Supplementary Information

A systematic review of clinical translation of large language models using electronic health records

Taiyu Zhu[1,*], Mandy Miresmaeili[1], Christopher Tomlinson[1,2], Linglong Qian[1], James Teo[1,3,4], Zina Ibrahim[1], Richard Dobson[1,2]

[1]Department of Biostatistics & Health Informatics, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, United Kingdom

[2]Institute of Health Informatics, University College London, London, United Kingdom

[3]King's College Hospital NHS Foundation Trust, London, United Kingdom

[4]Guy's and St Thomas' NHS Foundation Trust, London, United Kingdom

[*]Correspondence: taiyu.zhu@kcl.ac.uk

This Supplementary Information accompanies the analysis with a publication cutoff of 1 June 2026. It contains Supplementary Notes 1–5, Supplementary Tables 1–11 and Supplementary Figs. 1–3. The completed PRISMA 2020 and PRISMA 2020 for Abstracts checklists are appended after the Supplementary References.

Supplementary Note 1. Executable search queries

The four database queries below are reproduced verbatim for reports published between 1 January 2022 and 1 June 2026. All four databases were last searched on 2 October 2026. Search concepts, field and date restrictions, language handling, and record-management procedures are described in the main Methods.

PubMed

```
("large language model*"[Title/Abstract] OR LLM[Title/Abstract] OR LLMs[Title/Abstract] OR "generative artificial
     intelligence"[Title/Abstract] OR "generative AI"[Title/Abstract] OR "foundation model*"[Title/Abstract] OR
     "instruction-tuned"[Title/Abstract] OR "instruction-following"[Title/Abstract] OR GPT[Title/Abstract] OR
     ChatGPT[Title/Abstract] OR OpenAI[Title/Abstract] OR Claude[Title/Abstract] OR Gemini[Title/Abstract] OR
     Gemma[Title/Abstract] OR Qwen[Title/Abstract] OR Llama[Title/Abstract] OR Mistral[Title/Abstract] OR
     Mixtral[Title/Abstract] OR DeepSeek[Title/Abstract] OR Vicuna[Title/Abstract] OR Alpaca[Title/Abstract] OR
     Falcon[Title/Abstract] OR Phi-3[Title/Abstract] OR "Med-PaLM"[Title/Abstract] OR MedGemma[Title/Abstract] OR
     BioGPT[Title/Abstract] OR Meditron[Title/Abstract] OR "Clinical-Camel"[Title/Abstract] OR ClinicalGPT[Title/Abstract] OR
     "retrieval augmented generation"[Title/Abstract])
AND ("electronic health record*"[Title/Abstract] OR EHR[Title/Abstract] OR EHRs[Title/Abstract] OR "electronic medical
     record*"[Title/Abstract] OR EMR[Title/Abstract] OR EMRs[Title/Abstract] OR "medical record*"[Title/Abstract] OR "patient
     record*"[Title/Abstract] OR "health record*"[Title/Abstract] OR "clinical note*"[Title/Abstract] OR "progress
     note*"[Title/Abstract] OR "admission note*"[Title/Abstract] OR "operative note*"[Title/Abstract] OR "consultation
     note*"[Title/Abstract] OR "discharge summar*"[Title/Abstract] OR "radiology report*"[Title/Abstract] OR "pathology
     report*"[Title/Abstract] OR "clinical report*"[Title/Abstract] OR "clinical text"[Title/Abstract] OR "clinical
     narrative*"[Title/Abstract] OR "free-text"[Title/Abstract] OR "unstructured clinical"[Title/Abstract] OR "patient
     portal*"[Title/Abstract] OR "clinical correspondence"[Title/Abstract] OR MIMIC[Title/Abstract] OR
     "MIMIC-III"[Title/Abstract] OR "MIMIC-IV"[Title/Abstract] OR eICU[Title/Abstract] OR i2b2[Title/Abstract] OR
     n2c2[Title/Abstract] OR PhysioNet[Title/Abstract] OR "Electronic Health Records"[MeSH])
AND (evaluat*[Title/Abstract] OR validat*[Title/Abstract] OR benchmark*[Title/Abstract] OR assess*[Title/Abstract] OR
     perform*[Title/Abstract] OR extract*[Title/Abstract] OR summar*[Title/Abstract] OR generat*[Title/Abstract] OR
     identif*[Title/Abstract] OR detect*[Title/Abstract] OR classif*[Title/Abstract] OR predict*[Title/Abstract] OR
     phenotyp*[Title/Abstract] OR annotat*[Title/Abstract] OR abstract*[Title/Abstract] OR document*[Title/Abstract] OR
     inform*[Title/Abstract] OR coding[Title/Abstract] OR ICD[Title/Abstract] OR triage[Title/Abstract] OR "decision
     support"[Title/Abstract] OR deploy*[Title/Abstract] OR implement*[Title/Abstract] OR workflow[Title/Abstract])
AND ("2022/01/01"[Date - Publication] : "2026/06/01"[Date - Publication])
AND English[Language]
NOT (review[Publication Type] OR meta-analysis[Publication Type] OR survey[Title])
```

Embase (Ovid)

```
1. ("large language model" OR "large language models" OR llm OR llms OR "generative artificial intelligence" OR "generative
     ai" OR "foundation model" OR "foundation models" OR gpt OR chatgpt OR openai OR claude OR gemini OR gemma OR qwen OR
     llama OR mistral OR mixtral OR deepseek OR "med-palm" OR medgemma OR "retrieval augmented generation").ti,ab.
2. ("electronic health record" OR "electronic health records" OR ehr OR ehrs OR "electronic medical record" OR "electronic
     medical records" OR emr OR emrs OR "medical record" OR "medical records" OR "patient record" OR "patient records" OR
     "health record" OR "health records" OR "clinical note" OR "clinical notes" OR "progress note" OR "progress notes" OR
     "admission note" OR "admission notes" OR "operative note" OR "operative notes" OR "consultation note" OR "consultation
     notes" OR "discharge summary" OR "discharge summaries" OR "radiology report" OR "radiology reports" OR "pathology
     report" OR "pathology reports" OR "clinical report" OR "clinical reports" OR "clinical text" OR "clinical narrative" OR
     "clinical narratives" OR "free text" OR "free-text" OR "unstructured clinical" OR "patient portal" OR "patient portals"
     OR "clinical correspondence" OR mimic OR "mimic-iii" OR "mimic-iv" OR eicu OR i2b2 OR n2c2 OR physionet).ti,ab.
3. (evaluate OR evaluates OR evaluated OR evaluation OR evaluating OR validate OR validates OR validated OR validation OR
     validating OR benchmark OR benchmarks OR benchmarked OR benchmarking OR assess OR assesses OR assessed OR assessment OR
     perform OR performs OR performed OR performance OR extract OR extracts OR extracted OR extraction OR summarize OR
     summarizes OR summarized OR summarization OR generate OR generates OR generated OR generation OR identify OR identifies
     OR identified OR identification OR detect OR detects OR detected OR detection OR classify OR classifies OR classified OR
     classification OR predict OR predicts OR predicted OR prediction OR phenotype OR phenotypes OR phenotyping OR annotate
     OR annotates OR annotated OR annotation OR abstract OR abstracts OR abstracted OR abstraction OR document OR documents
     OR documented OR documentation OR information OR coding OR icd OR triage OR "decision support" OR deploy OR deploys OR
     deployed OR deployment OR implement OR implements OR implemented OR implementation OR workflow).ti,ab.
4. 1 AND 2 AND 3
5. limit 4 to english language
6. limit 5 to yr="2022 - 2026"
```

Scopus

```
TITLE-ABS-KEY("large language model*" OR LLM OR LLMs OR "generative artificial intelligence" OR "generative AI" OR "foundation
     model*" OR GPT OR ChatGPT OR OpenAI OR Claude OR Gemini OR Gemma OR Qwen OR Llama OR Mistral OR Mixtral OR DeepSeek OR
     "Med-PaLM" OR MedGemma OR "retrieval augmented generation")
AND TITLE-ABS-KEY("electronic health record*" OR EHR OR EHRs OR "electronic medical record*" OR EMR OR EMRs OR "medical
     record*" OR "patient record*" OR "health record*" OR "clinical note*" OR "progress note*" OR "discharge summar*" OR
     "radiology report*" OR "pathology report*" OR "clinical report*" OR "clinical text" OR "clinical narrative*" OR MIMIC OR
     "MIMIC-III" OR "MIMIC-IV" OR eICU OR i2b2 OR n2c2 OR PhysioNet)
AND TITLE-ABS-KEY(evaluat* OR validat* OR benchmark* OR assess* OR extract* OR summar* OR predict* OR phenotyp* OR classif* OR
     coding OR ICD OR triage OR "decision support" OR deploy* OR implementation OR workflow)
AND PUBYEAR > 2021 AND PUBYEAR < 2027
AND LIMIT-TO(LANGUAGE, "English")
```

IEEE Xplore

```
("large language model" OR "large language models" OR LLM OR LLMs OR ChatGPT OR GPT OR OpenAI OR Claude OR Gemini OR Gemma OR
     Qwen OR Llama OR Mistral OR Mixtral OR DeepSeek OR "generative AI" OR "generative artificial intelligence" OR
     "foundation model" OR "retrieval augmented generation")
AND ("electronic health record" OR "electronic health records" OR EHR OR "electronic medical record" OR "electronic medical
     records" OR EMR OR "clinical note" OR "clinical notes" OR "medical record" OR "medical records" OR "patient record" OR
     "discharge summary" OR "discharge summaries" OR MIMIC)
AND (evaluation OR validation OR benchmark OR extraction OR summarization OR prediction OR classification OR coding OR triage
     OR "decision support" OR implementation OR workflow)
```

Supplementary Note 2. Human references and model comparisons

Supplementary Tables 1 and 2 report the ten-configuration comparisons of CALREP against the screening and profiling human consensus references. Reference construction, model identifiers, software versions and statistical procedures are described in the main Methods. These comparisons assess screening and the two primary profiling dimensions; they do not assess the descriptive fields.

Supplementary Table 1: Ten-configuration screening comparison against the 376-record human reference.

Configuration	Predicted eligible	Agreement	κ	Sensitivity	Specificity
GPT-5.6 Luna low	93	94.9%	0.864	90.2%	96.5%
GPT-5.6 Luna high	96	**95.7%**	**0.887**	93.5%	96.5%
GPT-5.6 Sol low	88	94.7%	0.854	87.0%	97.2%
GPT-5.6 Sol high	93	94.9%	0.864	90.2%	96.5%
GPT-5.6 Terra low	90	**95.7%**	0.884	90.2%	**97.5%**
GPT-5.6 Terra high	91	95.5%	0.877	90.2%	97.2%
Claude Sonnet 5 low	98	93.6%	0.831	90.2%	94.7%
Claude Sonnet 5 high	99	94.4%	0.853	92.4%	95.1%
Claude Opus 5 low	112	93.1%	0.826	**96.7%**	91.9%
Claude Opus 5 high	108	93.6%	0.837	95.7%	93.0%

Predicted eligible is the number of reference records classified for inclusion. The bold configuration name identifies the configuration used in the review workflow. Bold numeric values are the highest point estimates in each performance column; ties are retained.

Supplementary Table 2: Ten-configuration profile comparison on the same 96 reports.

Configuration	Dataset breadth: agreement; κ (95% CI)	Translation: agreement; κ (95% CI)
GPT-5.6 Luna, low	88.5%; 0.813 (0.701–0.910)	**97.9%**; 0.881 (0.653–1.000)
GPT-5.6 Luna, high	**90.6%**; **0.849** (0.744–0.934)	96.9%; 0.830 (0.586–1.000)
GPT-5.6 Sol, low	86.5%; 0.777 (0.657–0.882)	96.9%; 0.821 (0.585–1.000)
GPT-5.6 Sol, high	87.5%; 0.796 (0.679–0.898)	96.9%; 0.821 (0.585–1.000)
GPT-5.6 Terra, low	86.5%; 0.782 (0.664–0.884)	**97.9%**; 0.881 (0.653–1.000)
GPT-5.6 Terra, high	89.6%; 0.833 (0.724–0.921)	**97.9%**; **0.892** (0.699–1.000)
Claude Sonnet 5, low	80.2%; 0.668 (0.527–0.794)	92.7%; 0.584 (0.273–0.825)
Claude Sonnet 5, high	76.0%; 0.587 (0.444–0.724)	94.8%; 0.744 (0.476–0.937)
Claude Opus 5, low	83.3%; 0.725 (0.591–0.841)	95.8%; 0.774 (0.518–0.950)
Claude Opus 5, high	87.5%; 0.794 (0.676–0.896)	95.8%; 0.773 (0.518–0.950)

All configurations were evaluated on 96 reports (5.03% of initially eligible reports). Bold values mark the highest point estimates, including ties; the bold name identifies the selected configuration.

Supplementary Note 3. Full-text review and evidence sources

Clinical translation stages and dataset-validation categories are defined in the main Methods and Table 1.

All 90 Stage B, 94 Stage C and 37 unclear-translation reports in the title-and-abstract profile were targeted. Two Stage A records without abstracts were also reviewed. Full-text review showed that one was an invited commentary and the other a letter to the editor; both were excluded as non-empirical publication types. Of the 223 targeted reports, 220 had an available full text. Three reports remained under repository embargo and retained their original title-and-abstract labels in the adjusted map.

The full-text review excluded 39 reports and retained 181. Among retained reports, 70 retained both primary labels, 47 changed dataset breadth only, 28 changed translation stage only, and 36 changed both labels. The reviewed set contained 84 Stage C, 57 Stage B and 40 Stage A reports. The complete adjusted profile also includes the three unavailable reports and all other retained reports.

Only the two primary labels were substituted in the adjusted profile. Model names, access, use modes, clinical tasks, human-comparison categories, governance, safety and outcome fields remained the original title-and-abstract extractions. The separate trial categories and concise frontier characteristics came from the full-text review. Descriptive summaries therefore use the title-and-abstract fields consistently across all retained reports.

Supplementary Table 3: Evidence-profile dimensions before and after targeted full-text adjustment.

Category	Title and abstract (N = 1,909)	Adjusted map (N = 1,870)
Level 1: Single EHR dataset validation	1,077 (56.4%)	1,051 (56.2%)
Level 2: Multiple EHR datasets or sites	403 (21.1%)	450 (24.1%)
Level 3: Mixed public and private EHR validation	46 (2.4%)	53 (2.8%)
Unclear dataset breadth	383 (20.1%)	316 (16.9%)
Stage A: Retrospective/offline evaluation	1,688 (88.4%)	1,726 (92.3%)
Stage B: Prospective or standalone user-facing evaluation	90 (4.7%)	58 (3.1%)
Stage C: Integrated, interventional, or outcome-impact evaluation	94 (4.9%)	86 (4.6%)
Unclear translation stage	37 (1.9%)	0 (0.0%)

Values are n (%). The adjusted map substitutes primary labels for 181 eligible reviewed reports and removes 39 full-text-ineligible reports. It retains three targeted reports with unavailable full texts. Descriptive fields were not substituted.

Supplementary Table 4: Design and evaluation characteristics of reports combining broader validation with Stage C evaluation (n = 58).

Year and report	Validation and EHR datasets or sites	Stage, evaluation and workflow	Study design and human role	Outcomes evaluated
2024 Constructing a Large Language Model to Generate Impressions from Findings in Radiology Reports.[1]	Level 2 Three named hospital sites: Shanghai Jiading Jiangqiao Hospital, Shanghai General Hospital Songjiang Branch, and Shanghai General Hospital *Data source:* Local/private only	Stage C Prospective user-facing Integrated into routine workflow	Six radiologists used the LLM during reporting; radiologists and an expert panel assessed outputs.	Similarity to final impressions and expert-rated terminology, coherence, diagnoses, recommendations, correctness, comprehensiveness, harmlessness, and bias.
2024 Evaluating a Natural Language Processing-Driven, AI-Assisted International Classification of Diseases, 10th Revision, Clinical Modification, Coding System for Diagnosis Related Groups in a Real Hospital Environment: Algorithm Development and Validation Study.[2]	Level 2 Two temporally distinct KMUCHH EHR cohorts/datasets; one hospital *Data source:* Local/private only	Stage C Mixed retrospective and prospective Operational deployment	Certified coding specialists reviewed AI recommendations, selected final codes, and provided reference coding.	ICD-10-CM coding F1, principal-diagnosis accuracy, MDC agreement, and coding error rate.
2024 The Impact of Large Language Model-Generated Radiology Report Summaries on Patient Comprehension: A Randomized Controlled Trial.[3]	Level 2 Two primary care clinics *Data source:* Local/private only	Stage C Prospective user-facing Interventional evaluation	Randomized clinical trial Patients were randomized to report formats and completed surveys assessing comprehension.	Patient comprehension of radiology reports measured through post-report surveys.
2024 Prompt engineering on leveraging large language models in generating response to InBasket messages.[4]	Level 2 5 sites within one health system; separate training, validation, and production/test cohorts *Data source:* Local/private only	Stage C Mixed retrospective and prospective Operational deployment	Clinicians rated drafts; patient advisors compared tone, quality, authorship, and whether messages addressed concerns.	Draft acceptance/helpfulness, hallucinations, patient tone and quality, clinician satisfaction, perceived cognitive load, and InBasket time.
2024 Prompt engineering with a large language model to assist providers in responding to patient inquiries: a real-time implementation in the electronic health record.[5]	Level 2 One health system (UW Health); three non-overlapping temporal EHR phases *Data source:* Local/private only	Stage C Prospective user-facing Operational deployment	Other interventional study Physician providers reviewed, edited, and sent or discarded GPT-4 drafts; 17 nurse clinicians joined the later phase.	Draft-message utilization, provider negative sentiment, editing distance, and prompt similarity; patient experience was not collected.
2024 Scaling Note Quality Assessment Across an Academic Medical Center with AI and GPT-4[6]	Level 2 Six inpatient facilities within one health system; exact GPT-4-specific dataset count unclear *Data source:* Local/private only	Stage C Mixed retrospective and prospective Operational deployment	Single-arm pilot implementation Physicians reviewed feedback before distribution; clinicians provided survey responses about usefulness and intended note changes.	GPT-4 F-scores; clinician feedback ratings; note quality, documentation burden, length of stay, and mortality.
2025 AI-generated draft replies to patient messages: exploring effects of implementation.[7]	Level 2 One hospital; two temporally distinct EHR message cohorts, with historical development-cohort size unclear and 919 prospective messages evaluated *Data source:* Local/private only	Stage C Mixed retrospective and prospective Operational deployment	Clinicians selected draft or blank replies, edited outputs, and approved final patient messages before sending.	Draft adoption, retained-text similarity, adjustment burden, response times, and variation across specialties.

Supplementary Table 4 continued

Year and report	Validation and EHR datasets or sites	Stage, evaluation and workflow	Study design and human role	Outcomes evaluated
2025 Application of unified health large language model evaluation framework to In-Basket message replies: bridging qualitative and quantitative assessments.[8]	Level 2 2 EHR-derived datasets; site count not reported *Data source:* Local/private only	Stage C Mixed retrospective and prospective Integrated into routine workflow	Clinicians annotated AI drafts, decided on use, and edited final replies.	Fluency, coherence, clarity, brevity, relevance, completeness, factuality, toxicity, and agreement with clinician use decisions.
2025 Artificial intelligence-generated draft replies to patient messages in pediatrics.[9]	Level 2 4 ambulatory specialty practices in one academic health system *Data source:* Local/private only	Stage C Prospective user-facing Integrated into routine workflow	Single-arm pilot implementation Physicians, nurses, medical assistants, and trainees used and surveyed the draft-response feature.	Draft utilization, task load, burnout, perceived time savings, likelihood to recommend, and net promoter score.
2025 A Bilingual On-Premises AI Agent for Clinical Drafting: Implementation Report of Seamless Electronic Health Records Integration in the Y-KNOT Project.[10]	Level 2 One hospital (Severance Hospital); two separately described EHR cohorts/stages *Data source:* Local/private only	Stage C Retrospective/offline Operational deployment	Physicians prepared training completions, graded drafts, and reviewed and approved generated documents.	Draft quality, safety, usability, satisfaction, decision-making impact, and discharge-summary completion rates.
2025 Comparative analysis of accuracy and completeness in standardized database generation for complex multilingual lung cancer pathological reports: large language model-based assisted diagnosis system vs. DeepSeek, GPT-3.5, and healthcare professionals with varied professional titles, with task load variation assessment among medical staff.[11]	**Level 3** Two reported datasets: Shanghai Chest Hospital and TCGA; exact case counts vary across sections *Data source:* Mixed public and local/private	Stage C Mixed retrospective and prospective Operational deployment	Ten clinicians generated databases; clinicians rated workload before and after LLM assistance.	Pathology-field accuracy/completeness versus clinicians and models, plus clinician NASA-TLX task load.
2025 Data augmented large language models for medical record generation[12]	Level 2 One hospital; two separately described task-specific EHR datasets (D2T and MTS) *Data source:* Local/private only	Stage C Mixed retrospective and prospective Integrated into routine workflow	Single-arm pilot implementation Physicians review generated records and directly use them after confirmation or minor adjustments.	Faithfulness, ROUGE scores, physician time saved, and workload reduction during medical-record writing.
2025 Development and evaluation of a clinical note summarization system using large language models.[13]	Level 2 Four hospitals *Data source:* Local/private only	Stage C Prospective user-facing Integrated into routine workflow	Physicians used, edited, validated, and rated summaries; primary-care physicians and patients provided survey/interview feedback.	Summary ratings, completeness, excessiveness, correctness, physician acceptance, and patient/physician attitudes toward AI-assisted discharge summaries.
2025 Development and prospective implementation of a large language model based system for early sepsis prediction.[14]	Level 2 Two UCSD Health hospitals/two emergency departments, with retrospective and separate prospective cohorts *Data source:* Local/private only	Stage C Mixed retrospective and prospective Operational deployment	Two clinicians reviewed 50 false-positive EHR charts; BPA was not shown to end-users.	Sensitivity, PPV, F1, false alarms, and clinician assessment of suspected bacterial infection.
2025 The Effect of Ambient Artificial Intelligence Scribes on Trainee Documentation Burden.[15]	Level 2 Two urban academic medical centers *Data source:* Local/private only	Stage C Prospective user-facing Integrated into routine workflow	Resident physicians used the ambient scribe during clinical practice and completed usability and burden surveys.	Documentation burden, cognitive load, usability, acceptability, and EHR note-activity time.

Supplementary Table 4 continued

Year and report	Validation and EHR datasets or sites	Stage, evaluation and workflow	Study design and human role	Outcomes evaluated
2025 Evaluation and practical application of prompt-driven ChatGPTs for EMR generation[16]	Level 2 3 medical centers *Data source:* Local/private only	Stage C Mixed retrospective and prospective Operational deployment	Clinicians quality-controlled and rated outputs and conducted consultations; patients answered guided questions and provided voice recordings.	EMR linguistic/clinical quality (BERTScore, ROUGE, 7DI) and clinician consultation time.
2025 Extracting Pulmonary Embolism Diagnoses From Radiology Impressions Using GPT-4o: Large Language Model Evaluation Study.[17]	Level 2 One health system; three temporally distinct cohorts/datasets (training, validation, post-deployment testing) *Data source:* Local/private only	Stage C Mixed retrospective and prospective Operational deployment	Healthcare professionals use the web app and confirm model results in a human-in-the-loop approach.	Sensitivity, specificity, F1 score, post-deployment performance, and demographic subgroup F1 scores.
2025 An eyecare foundation model for clinical assistance: a randomized controlled trial.[18]	Level 2 Multiple hospital-derived cohorts and clinical sites; exact validation-dataset total unclear *Data source:* Local/private EHR cohorts; public imaging datasets also used, but public EHR provenance not established	Stage C Mixed retrospective and prospective Integrated into routine workflow	Randomized clinical trial Ophthalmologists and junior eyecare providers used EyeFM for screening, diagnosis, referral, reporting and management suggestions.	Diagnostic/referral accuracy, report standardization, satisfaction, follow-up compliance, clinician usability, safety, and assessment time.
2025 Generative AI-Assisted Nursing Handover: Enhancing Clinical Data Integration and Work Efficiency[19]	Level 2 Three affiliated hospitals: Taipei Medical University Hospital, Wan Fang Hospital, and Shuang Ho Hospital *Data source:* Local/private only	Stage C Unclear Operational deployment	Nurses' needs informed the template, and user feedback was analyzed.	Handover documentation time, monthly hours saved, and AI-generated-content reliability.
2025 Home-based Dry Eye Assessment via Blink Kinematics Using mmWave and Clinical Knowledge Distillation[20]	Level 2 Two EHR-derived cohorts: 192 participants with matched EHRs and an additional 1,119 patients with clinic-test data *Data source:* Local/private only	Stage C Mixed retrospective and prospective Integrated into routine workflow	Ophthalmologists diagnosed participants; doctors performed TBUT, TMR, and MGD tests; participants underwent radar measurements.	TBUT mean absolute error and DED severity accuracy; robustness across distance, age, gender, duration, angle, outliers, and longitudinal consistency.
2025 Hybrid Ambient Clinical Documentation and Physician Performance: Work Outside of Work, Documentation Delay, and Financial Productivity.[21]	Level 2 14 adult primary care practices within one health system; multiple EHR data sources *Data source:* Local/private only	Stage C Retrospective/offline Integrated into routine workflow	Clinicians activated audio and finalized notes; virtual scribes reviewed, formatted, and transferred drafts.	After-hours EHR work, documentation delay, and clinician financial productivity measured by wRVUs.
2025 Impact of using an AI scribe on clinical documentation and clinician-patient interactions in allied health private practice: perspectives of clinicians and patients.[22]	Level 2 Multiple clinics within one allied health organisation; exact number unclear *Data source:* Local/private only	Stage C Prospective user-facing Integrated into routine workflow	AHPs used and reviewed AI-generated documentation; patients consented to scribe use and reported their experiences.	Documentation time and burden, productivity, therapeutic alliance, clinician-patient interaction, trust, and patient perceptions.

Supplementary Table 4 continued

Year and report	Validation and EHR datasets or sites	Stage, evaluation and workflow	Study design and human role	Outcomes evaluated
2025 Implementation of large language models in electronic health records.[23]	Level 2 Two EHR-derived sources; one named hospital plus regional eHealth network, with external site count unclear *Data source:* Local/private only	Stage C Prospective user-facing Operational deployment	Physicians, nurses, and students used the assistant for EHR retrieval, summarization, diagnosis support, and note drafting.	Adoption, conversation volume, task distribution, feedback, reported errors, and workflow incorporation; no clinical outcomes or time savings.
2025 IMQC: A Large Language Model Platform for Medical Quality Control[24]	**Level 3** One open-source EMR dataset plus de-identified institutional EMRs; exact site count unclear *Data source:* Mixed public and local/private	Stage C Unclear Operational deployment	Clinicians wrote and reviewed instruction samples; experts review indicator calculations.	Accuracy and F1 score, processing throughput, reported manpower savings, and quality-control indicator results.
2025 Information Extraction and Summarization for Neurovascular Consultations with GPT-4o: A Clinical Case Study.[25]	Level 2 Two temporal cohorts in one university hospital (5 prompt-development cases; 70 evaluation patients) *Data source:* Local/private only	Stage C Mixed retrospective and prospective Integrated into routine workflow	Clinicians reviewed summaries during routine consultation preparation and assessed correctness and completeness.	Extraction precision, recall, specificity, accuracy; summary usefulness, completeness, source accuracy, and clinician-rated correctness.
2025 An Integrated AI Cloud Sharing Framework for Predictive AI and Generative AI in Healthcare[26]	Level 2 Three hospitals/sites: Chi Mei Medical Center and two external hospitals *Data source:* Local/private only	Stage C Prospective user-facing Operational deployment	Healthcare staff use generative AI for medical documentation and ward-round assistance.	Framework operation, multisite deployment, and documentation assistance; workload and content-quality benefits are claimed, not quantitatively evaluated.
2025 KDA-TCM: Evidence-Grounded Prescription Generation From Symptoms in Chinese EMRs[27]	Level 2 One TCM department; 130,213 historical EMRs plus 500 new silent-trial cases *Data source:* Local/private only	Stage C Mixed retrospective and prospective Operational deployment	Three senior TCM physicians retrospectively rated 300 cases; consulting physicians were not exposed to silent-trial outputs.	Classification and prescription accuracy, fairness gaps, evidence consistency, safety violations, expert ratings, and silent-trial performance.
2025 Leave as Fast as You Can: Using Generative AI to Automate and Accelerate Hospital Discharge Reports[28]	Level 2 One hospital; two separately reported EHR-derived cohorts (60 prototype HDRs and 47 pilot HDRs) *Data source:* Local/private only	Stage C Mixed retrospective and prospective Integrated into routine workflow	Single-arm pilot implementation Physicians activated the tool, reviewed/edited drafts, and evaluated errors, omissions, style, and correctness.	ROUGE overlap, qualitative HDR discrepancies, and clinician assessments of factual errors, omissions, style, correctness, and editing concerns.
2025 MedCT: A Clinical Terminology Graph for Generative AI Applications in Healthcare[29]	**Level 3** At least two EHR datasets/sites: MIMIC-IV and one tertiary-care hospital in Zhejiang *Data source:* Mixed public and local/private	Stage C Mixed retrospective and prospective Operational deployment	Hospitalists used generated discharge-summary drafts; physicians annotated retrieval relevance and rated summaries.	NER/linking, retrieval precision-recall-F1, summary similarity, human ratings, and reported documentation-time reduction.
2025 MedStructGen: A Two-Stage Method for Medical Record Generation Using Large Language Models[30]	Level 2 Two proprietary real-world datasets from one partner hospital *Data source:* Local/private EHR datasets; public ACI-BENCH contains role-play dialogues	Stage C Mixed retrospective and prospective Operational deployment	Physicians' documentation time was measured; three medical professionals provided blind ratings.	BLEU-4, ROUGE, BERTScore, faithfulness, structural compliance, hallucination rate, and physician documentation time.

Supplementary Table 4 continued

Year and report	Validation and EHR datasets or sites	Stage, evaluation and workflow	Study design and human role	Outcomes evaluated
2025 A Novel Playbook for Pragmatic Trial Operations to Monitor and Evaluate Ambient Artificial Intelligence in Clinical Practice.[31]	Level 2 12 clinic locations within one health system *Data source:* Local/private only	Stage C Prospective user-facing Operational deployment	Randomized clinical trial Providers used the tool and were interviewed; certified professional coders reviewed diagnosis-documentation accuracy.	Ambient-AI utilization, note time, work outside work, coding accuracy, LLM–coder agreement, consent compliance, and provider interviews.
2025 PIRO: A web-based search platform for pathology reports, leveraging large language models to generate discrete searchable insights.[32]	Level 2 Two EHR/LIS datasets; one institution *Data source:* Local/private only	Stage C Unclear Operational deployment	Pathologists, trainees, and other personnel used PIRO; 55 staff and trainee respondents rated it.	Query volume, adoption, execution time, usability ratings, and LLM malignancy-risk accuracy, sensitivity, and specificity.
2025 A Pragmatic Randomized Controlled Trial of Ambient Artificial Intelligence to Improve Health Practitioner Well-Being.[33]	Level 2 Multiple ambulatory clinics across Wisconsin and Illinois; exact number unclear *Data source:* Local/private only	Stage C Prospective user-facing Operational deployment	Randomized clinical trial Health care practitioners used the ambient scribe during visits and reviewed AI-generated notes.	Practitioner well-being, documentation time, after-hours work, note quality, billing compliance, workflow efficiency, and software drift.
2025 A Prospective Comparison of Large Language Models for Early Prediction of Sepsis.[34]	Level 2 Two hospitals; two temporal cohorts (2,074 encounters total) *Data source:* Local/private only	Stage C Mixed retrospective and prospective Operational deployment	No end-user interaction; nurse-facing BPA was not shown.	Sensitivity, PPV, F1 score, and false alarms per patient hour in retrospective and prospective cohorts.
2025 Provision of Radiology Reports Simplified With Large Language Models to Patients With Cancer: Impact on Patient Satisfaction.[35]	Level 2 Two EHR-derived cohorts in one tertiary-care oncology center: 50 retrospective CT reports and 100 patient reports *Data source:* Local/private only	Stage C Mixed retrospective and prospective Standalone user-facing tool	Other interventional study Radiologist supervised output checking and patient interviews; technical staff administered questionnaires; oncologist and psychiatrist provided on-site consultation.	Readability; report accuracy/completeness; patient knowledge, confidence, understanding, perceived utility, reading time, distress, queries, and corrections.
2025 RedactOR: An LLM-Powered Framework for Automatic Clinical Data De-Identification[36]	**Level 3** At least two: public i2b2/UTHealth corpus and internal clinical audio dataset; production EHR pipeline also described *Data source:* Mixed public and local/private	Stage C Mixed retrospective and prospective Operational deployment	Humans review schemas before registry; authorized users analyze de-identified records.	PHI/PII precision, recall, F1, all-or-nothing recall, token usage, latency, and qualitative audio utility.
2025 TanaGPT: How Can I Help You? A Pilot Testing of a Genai Assistant in a Healthcare Workflow.[37]	Level 2 Two EHR-derived cohorts/datasets at one hospital: 163-record performance dataset and 54-patient pilot cohort *Data source:* Local/private only	Stage C Mixed retrospective and prospective Integrated into routine workflow	Five ED physicians used TanaGPT, verified answers, completed usability measures, and participated in interviews.	Usage, UMUX-Lite usability, NPS, perceived usefulness, response time, verification burden, and workflow efficiency.
2025 Utilization of Generative AI-drafted Responses for Managing Patient-Provider Communication.[38]	Level 2 Three clinical sites within one health system *Data source:* Local/private only	Stage C Prospective user-facing Integrated into routine workflow	Healthcare professionals reviewed, edited, and sent AI-generated drafts in the EHR InBasket.	Draft utilization, linguistic characteristics, turnaround time, InBasket actions, forwarding, and workflow efficiency.

Supplementary Table 4 continued

Year and report	Validation and EHR datasets or sites	Stage, evaluation and workflow	Study design and human role	Outcomes evaluated
2026 AI-Powered Documentation for Mental Health Providers: Retrospective Observational Mixed Methods Study.[39]	Level 2 One Talkspace platform with two temporally distinct deployment cohorts: pilot and post-launch full-network use *Data source:* Local/private only	Stage C Retrospective/offline Operational deployment	MHPs consented, reviewed and edited generated notes, and optionally rated quality and supplied feedback.	Use/uptake, generation errors, text similarity, MHP-rated note quality, feedback/satisfaction, working hours, completed sessions, and client caseload.
2026 Artificial intelligence enhances screening efficiency for inflammatory bowel disease clinical trials: a prospective study.[40]	Level 2 Two temporally distinct cohorts at one health system; exact site count not otherwise specified *Data source:* Local/private only	Stage C Mixed retrospective and prospective Operational deployment	Gastroenterology physician and clinical research coordinators manually reviewed AI-flagged candidates.	Eligibility identification, screening yield and efficiency, and comparison with physician referrals.
2026 Augmenting Clinical Decision-Making with an Interactive and Interpretable AI Copilot: A Real-World User Study with Clinicians in Nephrology and Obstetrics[41]	Level 2 3 EHR datasets/sites: XY-Nephrology, BS-Nephrology, and BC-Obstetrics *Data source:* Local/private only	Stage C Prospective user-facing Operational deployment	Clinicians interacted with AICare, assessed anonymized cases, rated workload, usability, trust, confidence, and gave interviews.	Task time, diagnostic accuracy, workload, usability, trust, confidence, interaction behavior, perceived usefulness, and interview themes.
2026 Automating the segmentation, date extraction, and classification of multi-report PDFs in outside medical records using optical character recognition and generative artificial intelligence.[42]	Level 2 116 distinct external institutions *Data source:* Local/private only	Stage C Prospective user-facing Integrated into routine workflow	Clinicians and staff used the tool and rated time savings, satisfaction, and usability.	Segmentation, classification, date-extraction accuracy, review time, workflow efficiency, and clinician satisfaction.
2026 Awaaz-e-Sehat: A Mobile Voice-based AI System for EMR Generation and Clinical Decision Support in Low-resource Maternal Healthcare[43]	Level 2 One hospital/site; at least two cohorts: an 8-patient development pilot and a 297-record deployment evaluation cohort *Data source:* Local/private only	Stage C Mixed retrospective and prospective Operational deployment	Single-arm pilot implementation Healthcare workers entered patient data; doctors reviewed EMRs and red flags, answered clarifications, and retained clinical decision control.	EMR field-level accuracy, ASR WER, usability, red-flag accuracy and relevance, and qualitative clinician workflow feedback.
2026 Brain compensation mechanisms of large language models in clinical decision-making in acupuncture: A fusion study using fNIRS and eye-tracking[44]	Level 2 Multiple sites; exact number unclear *Data source:* Unclear	Stage C Prospective user-facing Standalone user-facing tool	Other interventional study Licensed practitioners used the LLM or manually selected acupoints from medical records.	Acupoint compatibility performance, fNIRS cortical activation/connectivity, and eye-tracking fixation, pupil, and attentional measures.
2026 Design and implementation of an end-to-end AI-driven colonoscopy recall workflow at scale.[45]	Level 2 2 EHR-derived document collections; one academic medical center *Data source:* Local/private only	Stage C Retrospective/offline Operational deployment	Clinicians manually reviewed/annotated outputs; primary care physicians provided post-deployment feedback.	GPT-4/GPT-4 Turbo extraction accuracy, edge-case errors, quality-control failure rates, scalability/cost, projected cancer prevention, and physician feedback.

Supplementary Table 4 continued

Year and report	Validation and EHR datasets or sites	Stage, evaluation and workflow	Study design and human role	Outcomes evaluated
2026 Development and Validation of a Generative Artificial Intelligence-Based Pipeline for Automated Clinical Data Extraction From Electronic Health Records: Technical Implementation Study.[46]	Level 2 Three separately described MRI-report datasets/cohorts (1800 current; 424 and 228 prior validation); site count beyond current UCSF unclear *Data source:* Local/private only	Stage C Retrospective/offline Operational deployment	Researchers select cohorts, initiate extraction, monitor processing, and can review outputs against source reports.	Completion rate, processing time, token use and cost, throughput, batch consistency, extraction completeness, and schema validation.
2026 A domain-adapted large language model to support clinicians in psychiatric clinical practice[47]	**Level 3** PsychClinical from multiple psychiatric centres; PsychBench from three centres; one prospective hospital *Data source:* Mixed public and local/private	Stage C Mixed retrospective and prospective Operational deployment	Other interventional study Resident psychiatrists used PsychFound during interviews, diagnosis, medication planning, documentation, and monitoring; attendings reviewed decisions.	Clinical task accuracy, interview quality, medication recommendations, documentation time, patient symptoms, safety, guideline adherence, usability, and operational throughput.
2026 Ethical Assessment of Generative AI Tools for Clinical Summarization Tasks.[48]	Level 2 At least two clinical settings/data streams: inpatient hospital nursing EHR fields and one outpatient clinic; exact hospital count unclear *Data source:* Local/private only	Stage C Prospective user-facing Integrated into routine workflow	Patients, developers, clinicians, and prospective users participated in stakeholder interviews; physicians piloted the VS tool.	Ethical concerns, stakeholder perceptions, accuracy and monitoring barriers, documentation implications, and recommendations; actual impacts were not measured.
2026 Evaluating ambient artificial intelligence documentation: effects on work efficiency, documentation burden, and patient-centered care.[49]	Level 2 One health system; four named locations (Irvine, Newport Beach, Orange, and Fountain Valley) *Data source:* Local/private only	Stage C Prospective user-facing Integrated into routine workflow	Physicians used the tools, reviewed and edited generated notes, finalized EHR documentation, and completed surveys.	Documentation time, note length, encounter closure, perceived burden, efficiency, patient-centered care, safety, usability, and workflow.
2026 Integrating a Large Language Model to Streamline Nursing Handover Documentation Across Multiple Hospitals in Taiwan: Development and Implementation Study.[50]	Level 2 3 hospitals: Taipei Medical University Hospital, Wan Fang Hospital, and Shuang Ho Hospital *Data source:* Local/private only	Stage C Retrospective/offline Operational deployment	Nurses reviewed, revised, verified, and confirmed generated handover drafts before submission.	Documentation time, nurse revision rates, estimated nursing time savings, GenAI costs, and net labor value.
2026 A Language Model for Pediatric Occupational Therapy Documentation: Model Development and Pilot Study.[51]	Level 2 2 EHR-derived datasets; one institution (KidsAbility) *Data source:* Local/private only	Stage C Mixed retrospective and prospective Integrated into routine workflow	Ten occupational therapists used the tools, reviewed and edited generated drafts, and submitted final notes.	SOAP note quality, documentation time, editing similarity, and model output quality.
2026 Large language model-assisted radiology reporting in a single-radiologist implementation: a retrospective cohort study interpreted through a UTAUT lens.[52]	Level 2 One institution (Mayo Clinic Arizona); two non-overlapping temporal cohorts: 295 baseline and 314 post-implementation studies *Data source:* Local/private only	Stage C Retrospective/offline Integrated into routine workflow	One abdominal radiologist copied EHR notes, dictated findings, reviewed/edited LLM output, and finalized reports.	Inter-study interval time, perceived usefulness, training/cognitive effort, satisfaction, work-life balance, and willingness to continue use.

Supplementary Table 4 continued

Year and report	Validation and EHR datasets or sites	Stage, evaluation and workflow	Study design and human role	Outcomes evaluated
2026 A large language model for complex cardiology care.[53]	Level 2 Two non-overlapping cohorts from one SCICD center: 9 development cases and 107 RCT test cases *Data source:* Public/shareable only	Stage C Retrospective/offline Standalone user-facing tool	Randomized clinical trial General cardiologists used AMIE or not; blinded subspecialists rated the resulting assessments.	Subspecialist preferences, error and omission rates, reasoning, bias, usefulness, confidence, hallucinations, and time savings.
2026 Mixed-methods evaluation of clinician experiences and adoption patterns of an EHR-integrated generative AI-based clinical decision support uptake by clinicians in Kenya[54]	Level 2 16 Penda Health facilities/sites within one private outpatient provider *Data source:* Local/private only	Stage C Unclear Operational deployment	Clinicians used the AI Consult during consultations and provided interviews, observations, focus groups, and output feedback.	Uptake, LLM interaction patterns, clinician feedback, usability, trust, workflow integration, and patient NPS; no patient outcomes.
2026 Physician-Reported Safety Outcomes of AI-Generated Hospital Course Summaries.[55]	Level 2 1 Stanford Health Care inpatient unit; 2 temporally distinct EHR periods (pilot and prepilot baseline) *Data source:* Local/private only	Stage C Mixed retrospective and prospective Integrated into routine workflow	Single-arm pilot implementation Hospitalist physicians reviewed, rated, edited, used, or discarded summaries and completed safety and well-being surveys.	Physician-reported harm, errors, use, documentation time, EHR closure time, summary quality, cognitive burden, and burnout.
2026 Psychiatric Documentation and Management in Primary Care With Artificial Intelligence Scribe Use[56]	Level 2 2 large academic health systems; affiliated hospitals and outpatient practices, with exact site count unclear *Data source:* Local/private only	Stage C Retrospective/offline Operational deployment	Primary care clinicians used ambient AI scribes during routine annual visits.	RDoC symptom documentation, note length, depression diagnoses, antidepressant prescriptions, behavioral health referrals, and composite psychiatric intervention.
2026 Reporting efficiency in diagnostic imaging: Can plug-and-play general-purpose large language models outperform conventional speech recognition?[57]	Level 2 4 French academic centers/sites *Data source:* Local/private only	Stage C Prospective user-facing Standalone user-facing tool	Five radiologists dictated reports, used the GPT, and judged errors in their reports.	Generation times, editing times, qualitative errors, clinically meaningful errors, and Levenshtein similarity.
2026 Safety of a large language model-based clinical decision support system in African primary healthcare.[58]	Level 2 16 clinics *Data source:* Local/private only	Stage C Retrospective/offline Operational deployment	Clinical officers used and could accept or reject advice; Kenyan physicians retrospectively rated outputs and safety.	Documentation quality; diagnostic and management reasoning; hallucinations; harmful recommendations; clinician documentation changes; safety-risk mitigation; adverse outcomes.

All 58 reports are Stage C and have full-text-derived primary classifications and study characteristics. The five Level 3 reports are identified by bold validation-level labels. Data-source descriptions refer to EHR validation: public imaging data without established EHR provenance (EyeFM) and simulated dialogue data (ACI-BENCH in MedStructGen) do not establish Level 3. Superscript numbers identify the corresponding Supplementary References. Study-design labels are shown only when a specific design was reported. DOI links and study-level audit fields are also retained in the study-linked atlas. Categories describe the reported evaluation design and do not constitute a formal risk-of-bias assessment.

Supplementary Table 5: Publication year and evidence maturity.

Profile	Year	N	Level 2 or 3	Stage B or C	Frontier
Title and abstract	2022	7	0 (0.0%)	0 (0.0%)	0 (0.0%)
Title and abstract	2023	61	11 (18.0%)	4 (6.6%)	0 (0.0%)
Title and abstract	2024	434	70 (16.1%)	22 (5.1%)	3 (0.7%)
Title and abstract	2025	901	220 (24.4%)	95 (10.5%)	12 (1.3%)
Title and abstract	2026	506	148 (29.2%)	63 (12.5%)	10 (2.0%)
Adjusted	2022	7	0 (0.0%)	0 (0.0%)	0 (0.0%)
Adjusted	2023	60	13 (21.7%)	1 (1.7%)	0 (0.0%)
Adjusted	2024	426	76 (17.8%)	18 (4.2%)	6 (1.4%)
Adjusted	2025	882	250 (28.3%)	76 (8.6%)	32 (3.6%)
Adjusted	2026	495	164 (33.1%)	49 (9.9%)	20 (4.0%)

Percentages use the number of reports in that profile and year. The 2026 interval ends on 1 June. Frontier means Level 2 or 3 combined with Stage C.

Supplementary Table 6: Dataset provenance and named public resources among 1,870 retained reports.

Characteristic	n (%)
Public/shareable only	367 (19.6%)
Unclear	881 (47.1%)
Local/institutional only	569 (30.4%)
Mixed public and local/institutional	53 (2.8%)
Named public resource	
MIMIC-IV	124 (6.6%)
MIMIC-III	95 (5.1%)
MIMIC-CXR	79 (4.2%)
i2b2 / n2c2	25 (1.3%)
MIMIC (unspecified)	25 (1.3%)
eICU	10 (0.5%)
PhysioNet	6 (0.3%)

Provenance is a separate title-and-abstract extraction. Public-dataset categories are non-exclusive and use fixed string patterns applied to the named-dataset field.

Supplementary Table 7: Trial evidence extracted from the 181 eligible full texts.

Trial category	n (%)
None stated	148 (81.8%)
Other interventional study	8 (4.4%)
Single-arm pilot implementation	15 (8.3%)
Randomized clinical trial	8 (4.4%)
Unclear	2 (1.1%)

The full-text cohort was selected primarily for non-retrospective or unclear translation. These proportions do not estimate trial-design prevalence across all retained reports. Trial designation describes evaluation of the LLM system rather than the provenance of its source data.

Supplementary Table 8: Human involvement, governance, safety and outcomes in titles and abstracts.

Characteristic	n (%)
Human comparison or review	
Not reported	948 (50.7%)
Direct physician/clinician comparator	184 (9.8%)
Clinician adjudication or error review	297 (15.9%)
Clinician-created reference labels	292 (15.6%)
Human-in-the-loop workflow	80 (4.3%)
Patient-facing evaluation	23 (1.2%)
Oversight and governance	
Not reported	1,814 (97.0%)
Clinical safety or regulatory process described	12 (0.6%)
Mandatory clinician review or sign-off	11 (0.6%)
Escalation or failure pathway described	8 (0.4%)
Post-deployment monitoring or drift surveillance	3 (0.2%)
Safety evaluation	
Not reported	1,529 (81.8%)
Hallucination or fabrication assessed	122 (6.5%)
Error rate reported without severity grading	110 (5.9%)
Omission assessed	73 (3.9%)
Harm or severity taxonomy applied	27 (1.4%)
Clinical harm observed and reported	4 (0.2%)
Outcome measured	
Model accuracy or task performance	1,747 (93.4%)
Workflow or time outcome	163 (8.7%)
Safety or error outcome	284 (15.2%)
Clinician acceptance, edit, or override rate	21 (1.1%)
Not reported	31 (1.7%)
Clinician experience or burden	52 (2.8%)
Patient outcome or experience	46 (2.5%)

Denominator: 1,870 retained reports. Categories are non-exclusive. Model or task performance can be reported alongside clinical or operational outcomes. Non-reporting refers to title-and-abstract information; it does not establish absence from a study. Free-text Other responses remain in the study profiles and are not pooled into the controlled categories.

Supplementary Table 9: Model characteristics among 1,870 retained reports.

Characteristic	n (%)
Model access	
Unclear or other	928 (49.6%)
Proprietary only	485 (25.9%)
Both	152 (8.1%)
Open-weight only	305 (16.3%)
Named model categories	

Supplementary Table 9 continued

Characteristic	n (%)
Unclear	629 (33.6%)
GPT-4	355 (19.0%)
Llama	344 (18.4%)
GPT-4o	209 (11.2%)
GPT-3.5	138 (7.4%)
GPT/ChatGPT unspecified	108 (5.8%)
Gemini	100 (5.3%)
Mistral/Mixtral	98 (5.2%)
Qwen	84 (4.5%)
DeepSeek	78 (4.2%)
Claude	58 (3.1%)
Gemma/MedGemma	57 (3.0%)
FastChat-T5/Flan-T5	36 (1.9%)
Phi	26 (1.4%)
GPT-2/GPT-Neo	25 (1.3%)
GPT-OSS	16 (0.9%)
Use mode	
Prompting	940 (50.3%)
Embeddings/classification	313 (16.7%)
Information extraction	797 (42.6%)
Fine-tuning	382 (20.4%)
Unclear	16 (0.9%)
Summarization	263 (14.1%)
Multimodal language model	157 (8.4%)
Agentic workflow	70 (3.7%)
Retrieval-augmented generation	194 (10.4%)

Model names and use modes are non-exclusive. The table lists the most frequently named model categories and separately preserves GPT-2/GPT-Neo and GPT-OSS. All values derive from title-and-abstract profiles. Access categories reflect reported structured labels.

Supplementary Table 10: Broad EHR-task groups among 1,870 retained reports.

Task group	n (%)
Information extraction, abstraction, and structuring	727 (38.9%)
Clinical classification, coding, and staging	396 (21.2%)
Prediction, prognosis, and risk stratification	273 (14.6%)
Summarisation, documentation, and report generation	418 (22.4%)
Diagnostic reasoning, treatment planning, and decision support	439 (23.5%)
Patient communication and education	119 (6.4%)
EHR-grounded question answering, retrieval, and temporal analysis	109 (5.8%)
Cohort identification and trial matching	19 (1.0%)
Medical imaging interpretation and segmentation	285 (15.2%)
Data quality, safety, bias, and privacy	125 (6.7%)
Synthetic data generation and augmentation	33 (1.8%)
Workflow, administrative, and operational support	22 (1.2%)

Post-hoc groups are non-exclusive. The fixed keyword rules and original free-text task descriptions are retained in the source data and interactive atlas, respectively. Task grouping, validation scope and atlas construction are described in the main Methods.

Supplementary Note 4. Preprint sensitivity analysis

The preprint sensitivity procedures are described in the main Methods. Supplementary Table 11 presents the comparison with the additional 164 eligible database-indexed preprints. The title-and-abstract comparison uses 1,909 peer-reviewed reports and 2,073 combined reports; the adjusted-map comparison uses 1,870 peer-reviewed reports and 2,034 combined reports. Added preprints retain title-and-abstract classifications.

Supplementary Table 11: Preprint sensitivity comparisons.

Profile	N	Level 2 or 3	Stage B or C	Frontier
Peer-reviewed title/abstract	1909	449 (23.5%)	184 (9.6%)	25 (1.3%)
Additional preprints	164	61 (37.2%)	15 (9.1%)	4 (2.4%)
Combined title/abstract	2073	510 (24.6%)	199 (9.6%)	29 (1.4%)
Adjusted peer-reviewed map	1870	503 (26.9%)	144 (7.7%)	58 (3.1%)
Adjusted map plus preprints	2034	564 (27.7%)	159 (7.8%)	62 (3.0%)

Values are n (%) except N. The adjusted map contains mixed full-text and title-and-abstract primary classifications; added preprints retain title-and-abstract classifications.

The atlas will remain frozen during peer review, apart from documented corrections. After publication, we plan quarterly updates following human review, with dated releases and change logs; the published version will remain archived.

Supplementary Note 5. Review prompts

Primary screening prompt

```
You are screening titles and abstracts for a systematic review of generative large language models (LLMs) that use real-world
      electronic health record (EHR) data. Decide whether each record should advance to the review using only the supplied
      title, abstract, and publication metadata.

This is a title-and-abstract screen. Include when all four criteria are directly supported or reasonably indicated by the
      record. Exclude when the record provides a clear reason that a criterion fails. Do not exclude merely because an
      abstract omits details that are normally reported only in the full text. When genuine uncertainty remains, include for
      review, but do not infer an eligible generative LLM or clinical/EHR data from generic references to AI, healthcare, or
      foundation models alone.

## Eligibility criteria

### C1. Peer-reviewed original empirical study

Include peer-reviewed full journal articles, empirical letters, implementation or quality-improvement reports with original
      results, and archival full papers in peer-reviewed conference proceedings. Online-first articles are eligible when
      formally published by the review cutoff of 1 June 2026.

Exclude preprints, working papers, manuscripts under review, reviews, surveys, meta-analyses, editorials, commentaries,
      perspectives, protocols without results, errata, and abstract-only meeting, poster, oral, or supplement publications. A
      conference venue alone is not grounds for exclusion: full archival papers in ACL, EMNLP, NAACL, NeurIPS, ICML, ICLR,
      IEEE, ACM, CHIL, MLHC, ML4H, AMIA, and similar proceedings are eligible.

When publication metadata are incomplete, distinguish a database record with a missing abstract from an abstract-only
      publication. Exclude under C1 only when the metadata identify an ineligible publication type or provide no reasonable
      evidence of a peer-reviewed full paper.

### C2. Eligible generative LLM centrally evaluated

Include general-purpose, instruction-following, generative-text, multimodal generative, or medically adapted LLMs that are
      developed, prompted, fine-tuned, benchmarked, deployed, or validated as a main part of the study. A specific commercial
      model name is not required when the abstract clearly describes training, fine-tuning, prompting, generation, or
      evaluation of an LLM. Unnamed proprietary systems are eligible when a generative language function is explicit.

Exclude encoder-only models such as BERT, BioBERT, ClinicalBERT, PubMedBERT, or BlueBERT; EHR/code-sequence foundation models
      such as Med-BERT, BEHRT, CEHR-BERT, or ETHOS; classical machine learning; task-specific non-generative transformers; and
      image-only models. Also exclude studies in which an LLM is used only for preprocessing, embeddings, or as a trivial
      comparator. Such ineligible models may appear as baselines when a generative LLM is centrally evaluated.

A record using the term "LLM" without naming a model can satisfy C2 when its methods or task make central generative-model
      evaluation clear. A bare mention of an LLM, generative AI, or a foundation model does not satisfy C2 when no generative
      language function or substantive LLM evaluation is described.

### C3. Real-world EHR-derived data used

Include when real-world EHR-derived data are used in training, fine-tuning, retrieval or context construction, development,
      validation, or testing. Eligible data include structured EHR variables; clinical, progress, discharge, radiology, or
      pathology reports and notes; laboratory results; medications; diagnoses; vital signs; portal messages; clinical
      correspondence; patient timelines; routine-care documentation; and records from institutional data warehouses,
      hospitals, health systems, or EHR-derived registries.

Treat data described as EHR or EMR data, clinical notes, routine clinical text, deidentified clinical encounters, hospital
      records, or patient timelines as sufficient abstract-level evidence of EHR-derived data unless the record states that
      they are synthetic, educational, or otherwise not derived from patient care. Generic references to medical or clinical
      reports, an EHR system, patient information, or a "real dataset" are insufficient unless the record indicates actual
      patient-care data or a named clinical dataset. The abstract does not need to use the exact phrase "real-world" or name
      an institution. Recognised EHR datasets and tasks, including MIMIC, eICU, i2b2/n2c2, PhysioNet EHR datasets, EHRSQL, and
      EHRSHOT, satisfy C3.

Include studies combining synthetic and real EHR data. Exclude synthetic-only studies and studies using only exam questions,
      textbook vignettes, literature, guidelines, surveys, social media, standalone wearables, trial-only assessments not
      drawn from an EHR, or patient-chatbot transcripts without EHR context. Do not assume that model pretraining used EHR
      data unless stated.

Image data alone are insufficient. However, a clinical image-report dataset is eligible when patient images are paired with
      routine clinical reports or the evaluated task generates or reasons over those reports. Exclude image-only studies with
      no clinical text, report, or structured EHR component.

### C4. EHR-relevant task

Include diagnosis; risk, outcome, deterioration, or next-event prediction; prognosis; cohort identification; phenotyping;
      longitudinal EHR or patient-timeline analysis or visualization; summarization; report or note generation; information
      extraction; clinical coding; triage; decision support; documentation support; EHR-grounded question answering or patient
      communication; quality improvement; clinical-research support; chart review; and data abstraction.

Include question answering grounded in real EHR data. Exclude medical-knowledge question answering based only on exams,
      standalone vignettes, literature, guidelines, or general medical knowledge. Also exclude medical education or trainee
      assessment, teaching or exam-material generation, literature review, paper writing, citation management, and generic NLP
      benchmarks on non-clinical text.

## Decision rules

- Apply the criteria to the evidence reasonably available in a title and abstract.
```

- Include a dataset or benchmark paper when it uses real EHR-derived data and reports a substantive generative-LLM evaluation.
- Include clinically described institutional or multicentre data when notes, reports, encounters, records, timelines, or another routine-care source is identified, even if the EHR extraction process is not described.
- Do not treat ordinary omission of a model brand, hospital name, or the words "real-world" as evidence that C2 or C3 fails when the study methods otherwise support them.
- When multiple criteria clearly fail, use the first applicable exclusion reason in this order: C1, C2, C3, then C4.

Output

Return exactly one JSON object with no surrounding prose:

```json
{
  "decision": "include" | "exclude",
  "exclusion_reason": "none" | "C1: not an eligible peer-reviewed full empirical paper" | "C2: no eligible generative LLM centrally evaluated" | "C3: no real-world EHR-derived data used" | "C4: no eligible EHR-relevant task",
  "rationale": "one or two sentences, <=50 words, citing record evidence and linking the decision to C1-C4"
}
```

Use `exclusion_reason: "none"` only for included records. Every excluded record must use C1, C2, C3, or C4 according to the priority order. Base the rationale on evidence in the supplied record.

Few-shot examples

Example 1: include, real and synthetic EHR data

Input: "A study of generative large language model for medical research and healthcare." This full npj Digital Medicine article reports training GatorTronGPT on de-identified UF Health clinical text and using it to generate and evaluate synthetic clinical text.

```json
{
  "decision": "include",
  "exclusion_reason": "none",
  "rationale": "The article trains GatorTronGPT on de-identified UF Health clinical text and evaluates synthetic clinical text, so C1-C4 hold: real EHR data are used alongside synthetic data."
}
```

Example 2: include, EHR-grounded question answering

Input: "EHRAgent: Code Empowers Large Language Models for Few-shot Complex Tabular Reasoning on Electronic Health Records." This full EMNLP paper evaluates a GPT-4-based agent on clinical questions over MIMIC-III, eICU, and TREQS EHR data.

```json
{
  "decision": "include",
  "exclusion_reason": "none",
  "rationale": "The full EMNLP paper evaluates GPT-4-based EHRAgent on clinical questions over MIMIC-III, eICU, and TREQS, satisfying C1-C4 as EHR-grounded QA."
}
```

Example 3: include, EHR-integrated proprietary system

Input: "AI-Generated Draft Replies Integrated Into Health Records and Physicians' Electronic Communication" (JAMA Network Open, 2024; doi:10.1001/jamanetworkopen.2024.6565). This full quality-improvement study evaluated an EHR-integrated generative AI system that drafted replies to patient portal messages for physicians at UC San Diego Health.

```json
{
  "decision": "include",
  "exclusion_reason": "none",
  "rationale": "The empirical quality-improvement study evaluates generative draft replies grounded in real EHR patient messages and integrated into physician workflows, so C1-C4 hold despite the underlying commercial model not being named."
}
```

Example 4: exclude, abstract-only publication

Input: "Context-Aware Concept Bottleneck Models: Integrating LLM Insights for Enhanced ARDS Identification." Source: European Respiratory Journal 2025;66(suppl 69):PA995; publication type: Conference Abstract; the publisher states that no full-text version is available.

```json
{
  "decision": "exclude",
  "exclusion_reason": "C1: not an eligible peer-reviewed full empirical paper",
  "rationale": "The publisher labels the record a Conference Abstract and states that no full-text version is available; therefore C1 fails."
}
```

Example 5: exclude, medical-knowledge question answering

Input: "How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment." This JMIR Medical Education article evaluates ChatGPT using AMBOSS and NBME examination questions.

```json
{
  "decision": "exclude",
  "exclusion_reason": "C3: no real-world EHR-derived data used",
  "rationale": "ChatGPT is evaluated on AMBOSS and NBME examination questions rather than real-world EHR-derived data, so C3 fails."
}
```

Evidence profiling prompt

```
You are classifying studies included in a systematic review of large language models (LLMs) using real-world electronic health record (EHR) data. Assign two primary evidence-profile labels and extract supporting study characteristics using only the supplied title, abstract, and metadata.

Classify what the study actually evaluated, not what the authors propose or describe as possible. Do not infer unreported datasets, sites, prospective evaluation, deployment, human involvement, oversight, safety assessment, outcomes, trial design, or synthetic-data use.

Absence of a statement is not evidence that an activity did not occur. Use ‘Not reported‘, ‘None stated‘, or ‘No use reported‘ to mean that the supplied record is silent, not that the study definitively lacked the feature.

## Primary dimensions

### D1. Dataset validation level

Choose exactly one.

- ‘Level 1: Single EHR dataset validation‘: one EHR-derived dataset, cohort, institution, or health system, including one public dataset such as MIMIC alone.
- ‘Level 2: Multiple EHR datasets or sites‘: more than one EHR-derived dataset, cohort, site, institution, or health system, including a clearly separate temporal validation cohort, without explicit use of both public and local/private EHR data.
- ‘Level 3: Mixed public and private EHR validation‘: explicit evaluation on both public/shareable EHR data and private/local/institutional EHR data.
- ‘Unclear dataset breadth‘: EHR-derived data are confirmed, but the record does not establish whether one or multiple datasets or sites were used.

Assign the highest level explicitly supported by the record: 3 over 2 over 1. Use ‘Unclear dataset breadth‘ only when no level can be determined. Deployment, clinician involvement, and study design do not increase the level. Level 3 requires explicit evaluation using both public/shareable and local/private EHR data.

### D2. Clinical translation stage

Choose exactly one.

- ‘Stage A: Retrospective/offline evaluation‘: evaluation on existing records or datasets without prospective monitoring, actual user-facing testing, routine workflow use, or intervention.
- ‘Stage B: Prospective or standalone user-facing evaluation‘: prospective, real-time, silent, or shadow evaluation; or actual testing of a standalone tool with clinicians, staff, or patients, without integration into routine clinical or operational workflow.
- ‘Stage C: Integrated, interventional, or outcome-impact evaluation‘: actual workflow integration or operational deployment; an RCT or other interventional evaluation; or direct measurement during clinical or operational use of effects on patient outcomes, clinician behaviour, safety, workflow, or care delivery.
- ‘Unclear translation stage‘: an eligible LLM evaluation is confirmed, but the record provides insufficient information to determine whether Stage A, B, or C applies.

Assign the most advanced stage explicitly supported by the record: C over B over A. Proposed deployment or integration does not qualify. Retrospective application to data from a prospective cohort or clinical trial remains Stage A. Clinician labels, adjudication, or ratings of retrospectively generated outputs alone do not raise Stage A to Stage B.

## Descriptive fields

These fields describe the evidence base but are not primary human-agreement outcomes.

### Model characteristics

- ‘model_names‘: extract specific centrally evaluated eligible generative LLM names separated by semicolons; exclude encoder-only models, classical machine-learning models, and other ineligible baselines; use ‘Unclear‘ if none is stated.
- ‘system_or_product_name‘: extract a named deployed system, platform, or vendor product separated by semicolons; do not repeat a model name; use ‘Unclear‘ if none is stated.
- ‘model_access_profile‘: choose one or more of ‘Open-weight, locally run‘, ‘Open-weight, API/hosted‘, ‘Open-weight, access unclear‘, ‘Closed/proprietary, locally deployed‘, ‘Closed/proprietary, API/hosted‘, ‘Closed/proprietary, access unclear‘, ‘Other: <brief description>‘, or ‘Unclear‘, separated by semicolons.
- ‘llm_use_mode‘: choose one or more of ‘Prompting‘, ‘Retrieval-augmented generation‘, ‘Fine-tuning‘, ‘Embeddings/classification‘, ‘Summarization‘, ‘Information extraction‘, ‘Agentic workflow‘, ‘Multimodal language model‘, ‘Other: <brief description>‘, or ‘Unclear‘, separated by semicolons.

### Dataset and clinical characteristics

- ‘named_dataset_or_setting‘: extract named datasets, hospitals, health systems, countries, or settings separated by semicolons; use ‘Unclear‘ if none is stated.
- ‘synthetic_data_use‘: choose ‘No use reported‘, ‘Training/development only‘, ‘Evaluation only‘, ‘Both training/development and evaluation‘, ‘Other: <brief description>‘, or ‘Unclear phase‘. Do not count ordinary LLM outputs unless they are reused as study data.
```

```
- ‘clinical_specialty‘: choose one or more of ‘Radiology‘, ‘Oncology‘, ‘Emergency medicine‘, ‘Critical care‘, ‘Internal
      medicine‘, ‘Cardiology‘, ‘Neurology‘, ‘Nephrology‘, ‘Surgery‘, ‘Primary care‘, ‘Paediatrics‘, ‘Obstetrics and
      gynaecology‘, ‘Psychiatry and mental health‘, ‘Pathology and laboratory medicine‘, ‘Pharmacy and medicines management‘,
      ‘Cross-specialty or general‘, ‘Other: <brief description>‘, or ‘Unclear‘, separated by semicolons.
- ‘clinical_task‘: state the main clinical, EHR, research, operational, or workflow task in a short phrase.

### Human and translational evidence

- ‘clinical_trial_evidence‘: choose ‘None stated‘, ‘Single-arm pilot implementation‘, ‘Randomized clinical trial‘, ‘Other
      interventional study: <brief design>‘, or ‘Unclear‘. Classify the LLM evaluation, not the source dataset.
- ‘human_comparison_or_review‘: choose one or more of ‘Not reported‘, ‘Clinician-created reference labels‘, ‘Clinician
      adjudication or error review‘, ‘Direct physician/clinician comparator‘, ‘Human-in-the-loop workflow‘, ‘Patient-facing
      evaluation‘, or ‘Other: <brief description>‘, separated by semicolons. Use ‘Not reported‘ only when no human role is
      stated.
- ‘oversight_and_governance‘: choose one or more of ‘Not reported‘, ‘Mandatory clinician review or sign-off‘, ‘Optional
      clinician review‘, ‘Escalation or failure pathway described‘, ‘Post-deployment monitoring or drift surveillance‘,
      ‘Clinical safety or regulatory process described‘, or ‘Other: <brief description>‘, separated by semicolons.
- ‘safety_evaluation‘: choose one or more of ‘Not reported‘, ‘Hallucination or fabrication assessed‘, ‘Omission assessed‘,
      ‘Error rate reported without severity grading‘, ‘Harm or severity taxonomy applied‘, ‘Clinical harm observed and
      reported‘, or ‘Other: <brief description>‘, separated by semicolons. Accuracy, F1, or agreement alone is not a safety
      evaluation.
- ‘outcome_measured‘: choose one or more of ‘Model accuracy or task performance only‘, ‘Clinician acceptance, edit, or
      override rate‘, ‘Workflow or time outcome‘, ‘Clinician experience or burden‘, ‘Patient outcome or experience‘, ‘Safety
      or error outcome‘, ‘Not reported‘, or ‘Other: <brief description>‘, separated by semicolons.

### Classification rationale

- ‘rationale‘: explain D1 and D2 in one or two sentences of no more than 60 words, citing the decisive record evidence rather
      than merely restating the labels.

Use ‘Other‘ only when no named category fits, and replace the placeholder with a description of no more than 10 words.

## Consistency rules

- Level 3 requires explicit evaluation using both public/shareable and local/private EHR data.
- Silent prospective evaluation or actual standalone user testing without routine integration is Stage B.
- Actual workflow integration, operational deployment, or evaluation of an LLM-supported intervention is Stage C.
- Clinician labels, adjudication, or retrospective comparison alone do not raise D2.
- Safety and outcome fields describe what was measured, not what was proposed.
- Do not infer oversight, sign-off, governance, model access, or deployment from the setting.
- If the abstract and metadata conflict, use the more conservative classification and explain the conflict.

## Output

Return exactly one JSON object with no surrounding prose:

‘‘‘json
{
  "dataset_validation_level": "Level 1: Single EHR dataset validation" | "Level 2: Multiple EHR datasets or sites" | "Level 3:
      Mixed public and private EHR validation" | "Unclear dataset breadth",
  "clinical_translation_stage": "Stage A: Retrospective/offline evaluation" | "Stage B: Prospective or standalone user-facing
      evaluation" | "Stage C: Integrated, interventional, or outcome-impact evaluation" | "Unclear translation stage",
  "descriptive_fields": {
    "model_names": "specific model names separated by semicolons, or Unclear",
    "system_or_product_name": "named systems or products separated by semicolons, or Unclear",
    "model_access_profile": "one or more allowed access profiles separated by semicolons",
    "llm_use_mode": "one or more allowed modes separated by semicolons",
    "named_dataset_or_setting": "named datasets or settings separated by semicolons, or Unclear",
    "synthetic_data_use": "No use reported" | "Training/development only" | "Evaluation only" | "Both training/development and
      evaluation" | "Other: brief description" | "Unclear phase",
    "clinical_specialty": "one or more allowed specialties separated by semicolons",
    "clinical_task": "short task phrase",
    "clinical_trial_evidence": "None stated" | "Single-arm pilot implementation" | "Randomized clinical trial" | "Other
      interventional study: brief design" | "Unclear",
    "human_comparison_or_review": "one or more allowed human roles separated by semicolons",
    "oversight_and_governance": "one or more allowed oversight categories separated by semicolons",
    "safety_evaluation": "one or more allowed safety categories separated by semicolons",
    "outcome_measured": "one or more allowed outcome categories separated by semicolons",
    "rationale": "one or two evidence-based sentences, <=60 words"
  }
}
‘‘‘

Use the named labels exactly. For an ‘Other‘ response, append the required brief description after the colon.

## Few-shot examples

The examples below are based on real peer-reviewed studies. Classify from the stated record evidence only.

### Example 1: single public dataset and retrospective evaluation

Paper: *DRG-LLaMA: tuning LLaMA model to predict diagnosis-related group for hospitalized patients* (npj Digital Medicine,
      2024; doi:10.1038/s41746-023-00989-3).

Record evidence: DRG-LLaMA-7B was fine-tuned with LoRA and evaluated on MIMIC-IV discharge summaries for DRG prediction. No
      prospective evaluation, workflow integration, trial, or human assessment was reported.

‘‘‘json
{
```

```
  "dataset_validation_level": "Level 1: Single EHR dataset validation",
  "clinical_translation_stage": "Stage A: Retrospective/offline evaluation",
  "descriptive_fields": {
    "model_names": "DRG-LLaMA-7B; LLaMA",
    "system_or_product_name": "Unclear",
    "model_access_profile": "Open-weight, access unclear",
    "llm_use_mode": "Fine-tuning; Embeddings/classification",
    "named_dataset_or_setting": "MIMIC-IV",
    "synthetic_data_use": "No use reported",
    "clinical_specialty": "Cross-specialty or general",
    "clinical_task": "DRG prediction from discharge summaries",
    "clinical_trial_evidence": "None stated",
    "human_comparison_or_review": "Not reported",
    "oversight_and_governance": "Not reported",
    "safety_evaluation": "Not reported",
    "outcome_measured": "Model accuracy or task performance only",
    "rationale": "The model was evaluated retrospectively on one public EHR dataset. No prospective monitoring, user-facing
     testing, workflow integration, or intervention was reported."
  }
}
```

Example 2: mixed public and local data with a clinician comparator

Paper: *Automatic structuring of radiology reports with on-premise open-source large language models* (European Radiology, 2025; doi:10.1007/s00330-024-11074-y).

Record evidence: A locally hosted Llama-2-70B-chat model retrospectively structured reports from MIMIC-CXR and a German university hospital. Its performance was compared directly with two human readers.

```json
{
  "dataset_validation_level": "Level 3: Mixed public and private EHR validation",
  "clinical_translation_stage": "Stage A: Retrospective/offline evaluation",
  "descriptive_fields": {
    "model_names": "Llama-2-70B-chat",
    "system_or_product_name": "Unclear",
    "model_access_profile": "Open-weight, locally run",
    "llm_use_mode": "Prompting; Information extraction",
    "named_dataset_or_setting": "MIMIC-CXR; German university hospital",
    "synthetic_data_use": "No use reported",
    "clinical_specialty": "Radiology",
    "clinical_task": "Structuring free-text chest radiograph reports",
    "clinical_trial_evidence": "None stated",
    "human_comparison_or_review": "Direct physician/clinician comparator",
    "oversight_and_governance": "Not reported",
    "safety_evaluation": "Not reported",
    "outcome_measured": "Model accuracy or task performance only",
    "rationale": "The retrospective evaluation used both public MIMIC-CXR reports and local hospital reports, establishing
     Level 3. Its clinician comparison did not constitute prospective or operational use, so it remains Stage A."
  }
}
```

Example 3: prospective standalone clinician comparison

Paper: *Comparative study of Claude 3.5-Sonnet and human physicians in generating discharge summaries for patients with renal insufficiency: assessment of efficiency, accuracy, and quality* (Frontiers in Digital Health, 2024; doi:10.3389/fdgth.2024.1456911).

Record evidence: In a prospective controlled study at Ningbo Hangzhou Bay Hospital, Claude 3.5-Sonnet and physicians independently generated discharge summaries for 100 patients. An expert physician panel compared accuracy, generation time, and quality; no workflow integration was reported.

```json
{
  "dataset_validation_level": "Level 1: Single EHR dataset validation",
  "clinical_translation_stage": "Stage B: Prospective or standalone user-facing evaluation",
  "descriptive_fields": {
    "model_names": "Claude 3.5-Sonnet",
    "system_or_product_name": "Unclear",
    "model_access_profile": "Closed/proprietary, access unclear",
    "llm_use_mode": "Prompting; Summarization",
    "named_dataset_or_setting": "Ningbo Hangzhou Bay Hospital; China",
    "synthetic_data_use": "No use reported",
    "clinical_specialty": "Nephrology",
    "clinical_task": "Generating discharge summaries for renal insufficiency",
    "clinical_trial_evidence": "None stated",
    "human_comparison_or_review": "Clinician adjudication or error review; Direct physician/clinician comparator",
    "oversight_and_governance": "Not reported",
    "safety_evaluation": "Not reported",
    "outcome_measured": "Model accuracy or task performance only; Workflow or time outcome",
    "rationale": "The LLM was prospectively compared with physicians on records from one hospital, supporting Level 1 and
     Stage B. The controlled evaluation did not report routine workflow integration or an intervention affecting care."
  }
}
```

Example 4: EHR-integrated quality-improvement study

````
Paper: *AI-Generated Draft Replies Integrated Into Health Records and Physicians' Electronic Communication* (JAMA Network
      Open, 2024; doi:10.1001/jamanetworkopen.2024.6565).

Record evidence: A randomized waiting-list quality-improvement study evaluated an unnamed commercial generative AI system
      integrated into the UC San Diego Health EHR. It drafted replies to patient messages for physician review and measured
      reading time, reply time, reply length, and clinician experience.

```json
{
  "dataset_validation_level": "Level 1: Single EHR dataset validation",
  "clinical_translation_stage": "Stage C: Integrated, interventional, or outcome-impact evaluation",
  "descriptive_fields": {
    "model_names": "Unclear",
    "system_or_product_name": "Epic-integrated GenAI draft-reply system",
    "model_access_profile": "Closed/proprietary, access unclear",
    "llm_use_mode": "Other: patient-message reply generation",
    "named_dataset_or_setting": "UC San Diego Health; United States",
    "synthetic_data_use": "No use reported",
    "clinical_specialty": "Primary care",
    "clinical_task": "Drafting replies to EHR patient messages",
    "clinical_trial_evidence": "Other interventional study: randomized waiting-list quality-improvement study",
    "human_comparison_or_review": "Human-in-the-loop workflow",
    "oversight_and_governance": "Mandatory clinician review or sign-off",
    "safety_evaluation": "Not reported",
    "outcome_measured": "Workflow or time outcome; Clinician experience or burden",
    "rationale": "The system used one health system's EHR messages and was evaluated after integration into physician
      workflows, supporting Level 1 and Stage C. The randomized waiting-list quality-improvement design measured operational
      and clinician outcomes."
  }
}
```
````

Preprint sensitivity screening prompt

```
You are screening titles and abstracts for a preprint sensitivity analysis of a systematic review of generative large language
      models (LLMs) that use real-world electronic health record (EHR) data. Decide whether each record should advance using
      only the supplied title, abstract, and publication metadata.

This is a title-and-abstract screen. Include when all four criteria are directly supported or reasonably indicated by the
      record. Exclude when the record provides a clear reason that a criterion fails. Do not exclude merely because an
      abstract omits details that are normally reported only in the full text. When genuine uncertainty remains, include for
      review, but do not infer an eligible generative LLM or clinical/EHR data from generic references to AI, healthcare, or
      foundation models alone.

## Eligibility criteria

### C1. Original full empirical study, including preprints

Include peer-reviewed full journal articles, empirical letters, implementation or quality-improvement reports with original
      results, archival full papers in peer-reviewed conference proceedings, and publicly available full empirical preprints.
      Online-first articles and preprints are eligible when available by the review cutoff of 1 June 2026.

Exclude working papers or manuscripts under review that are not publicly available as full preprints, reviews, surveys,
      meta-analyses, editorials, commentaries, perspectives, protocols without results, errata, and abstract-only meeting,
      poster, oral, or supplement publications. A conference venue alone is not grounds for exclusion: full archival papers in
      ACL, EMNLP, NAACL, NeurIPS, ICML, ICLR, IEEE, ACM, CHIL, MLHC, ML4H, AMIA, and similar proceedings are eligible.

When publication metadata are incomplete, distinguish a database record with a missing abstract from an abstract-only
      publication. Exclude under C1 only when the metadata identify an ineligible publication type or provide no reasonable
      evidence of a full empirical report.

### C2. Eligible generative LLM centrally evaluated

Include general-purpose, instruction-following, generative-text, multimodal generative, or medically adapted LLMs that are
      developed, prompted, fine-tuned, benchmarked, deployed, or validated as a main part of the study. A specific commercial
      model name is not required when the abstract clearly describes training, fine-tuning, prompting, generation, or
      evaluation of an LLM. Unnamed proprietary systems are eligible when a generative language function is explicit.

Exclude encoder-only models such as BERT, BioBERT, ClinicalBERT, PubMedBERT, or BlueBERT; EHR/code-sequence foundation models
      such as Med-BERT, BEHRT, CEHR-BERT, or ETHOS; classical machine learning; task-specific non-generative transformers; and
      image-only models. Also exclude studies in which an LLM is used only for preprocessing, embeddings, or as a trivial
      comparator. Such ineligible models may appear as baselines when a generative LLM is centrally evaluated.

A record using the term "LLM" without naming a model can satisfy C2 when its methods or task make central generative-model
      evaluation clear. A bare mention of an LLM, generative AI, or a foundation model does not satisfy C2 when no generative
      language function or substantive LLM evaluation is described.

### C3. Real-world EHR-derived data used

Include when real-world EHR-derived data are used in training, fine-tuning, retrieval or context construction, development,
      validation, or testing. Eligible data include structured EHR variables; clinical, progress, discharge, radiology, or
      pathology reports and notes; laboratory results; medications; diagnoses; vital signs; portal messages; clinical
      correspondence; patient timelines; routine-care documentation; and records from institutional data warehouses,
      hospitals, health systems, or EHR-derived registries.
```

````
Treat data described as EHR or EMR data, clinical notes, routine clinical text, deidentified clinical encounters, hospital
      records, or patient timelines as sufficient abstract-level evidence of EHR-derived data unless the record states that
      they are synthetic, educational, or otherwise not derived from patient care. Generic references to medical or clinical
      reports, an EHR system, patient information, or a "real dataset" are insufficient unless the record indicates actual
      patient-care data or a named clinical dataset. The abstract does not need to use the exact phrase "real-world" or name
      an institution. Recognised EHR datasets and tasks, including MIMIC, eICU, i2b2/n2c2, PhysioNet EHR datasets, EHRSQL, and
      EHRSHOT, satisfy C3.

Include studies combining synthetic and real EHR data. Exclude synthetic-only studies and studies using only exam questions,
      textbook vignettes, literature, guidelines, surveys, social media, standalone wearables, trial-only assessments not
      drawn from an EHR, or patient-chatbot transcripts without EHR context. Do not assume that model pretraining used EHR
      data unless stated.

Image data alone are insufficient. However, a clinical image-report dataset is eligible when patient images are paired with
      routine clinical reports or the evaluated task generates or reasons over those reports. Exclude image-only studies with
      no clinical text, report, or structured EHR component.

### C4. EHR-relevant task

Include diagnosis; risk, outcome, deterioration, or next-event prediction; prognosis; cohort identification; phenotyping;
      longitudinal EHR or patient-timeline analysis or visualization; summarization; report or note generation; information
      extraction; clinical coding; triage; decision support; documentation support; EHR-grounded question answering or patient
      communication; quality improvement; clinical-research support; chart review; and data abstraction.

Include question answering grounded in real EHR data. Exclude medical-knowledge question answering based only on exams,
      standalone vignettes, literature, guidelines, or general medical knowledge. Also exclude medical education or trainee
      assessment, teaching or exam-material generation, literature review, paper writing, citation management, and generic NLP
      benchmarks on non-clinical text.

## Decision rules

- Apply the criteria to the evidence reasonably available in a title and abstract.
- Include a dataset or benchmark paper when it uses real EHR-derived data and reports a substantive generative-LLM evaluation.
- Include clinically described institutional or multicentre data when notes, reports, encounters, records, timelines, or
      another routine-care source is identified, even if the EHR extraction process is not described.
- Do not treat ordinary omission of a model brand, hospital name, or the words "real-world" as evidence that C2 or C3 fails
      when the study methods otherwise support them.
- Do not exclude a record solely because it is a preprint.
- When multiple criteria clearly fail, use the first applicable exclusion reason in this order: C1, C2, C3, then C4.

## Output

Return exactly one JSON object with no surrounding prose:

‘‘‘json
{
  "decision": "include" | "exclude",
  "exclusion_reason": "none" | "C1: not an eligible full empirical paper" | "C2: no eligible generative LLM centrally
      evaluated" | "C3: no real-world EHR-derived data used" | "C4: no eligible EHR-relevant task",
  "rationale": "one or two sentences, <=50 words, citing record evidence and linking the decision to C1-C4"
}
‘‘‘

Use ‘exclusion_reason: "none"‘ only for included records. Every excluded record must use C1, C2, C3, or C4 according to the
      priority order. Base the rationale on evidence in the supplied record.

## Few-shot examples

### Example 1: include, real and synthetic EHR data

Input: "A study of generative large language model for medical research and healthcare." This full npj Digital Medicine
      article reports training GatorTronGPT on de-identified UF Health clinical text and using it to generate and evaluate
      synthetic clinical text.

‘‘‘json
{
  "decision": "include",
  "exclusion_reason": "none",
  "rationale": "The article trains GatorTronGPT on de-identified UF Health clinical text and evaluates synthetic clinical
      text, so C1-C4 hold: real EHR data are used alongside synthetic data."
}
‘‘‘

### Example 2: include, EHR-grounded question answering

Input: "EHRAgent: Code Empowers Large Language Models for Few-shot Complex Tabular Reasoning on Electronic Health Records."
      This full EMNLP paper evaluates a GPT-4-based agent on clinical questions over MIMIC-III, eICU, and TREQS EHR data.

‘‘‘json
{
  "decision": "include",
  "exclusion_reason": "none",
  "rationale": "The full EMNLP paper evaluates GPT-4-based EHRAgent on clinical questions over MIMIC-III, eICU, and TREQS,
      satisfying C1-C4 as EHR-grounded QA."
}
‘‘‘

### Example 3: include, EHR-integrated proprietary system
````

Input: "AI-Generated Draft Replies Integrated Into Health Records and Physicians' Electronic Communication" (JAMA Network Open, 2024; doi:10.1001/jamanetworkopen.2024.6565). This full quality-improvement study evaluated an EHR-integrated generative AI system that drafted replies to patient portal messages for physicians at UC San Diego Health.

```json
{
  "decision": "include",
  "exclusion_reason": "none",
  "rationale": "The empirical quality-improvement study evaluates generative draft replies grounded in real EHR patient messages and integrated into physician workflows, so C1-C4 hold despite the underlying commercial model not being named."
}
```

Example 4: exclude, abstract-only publication

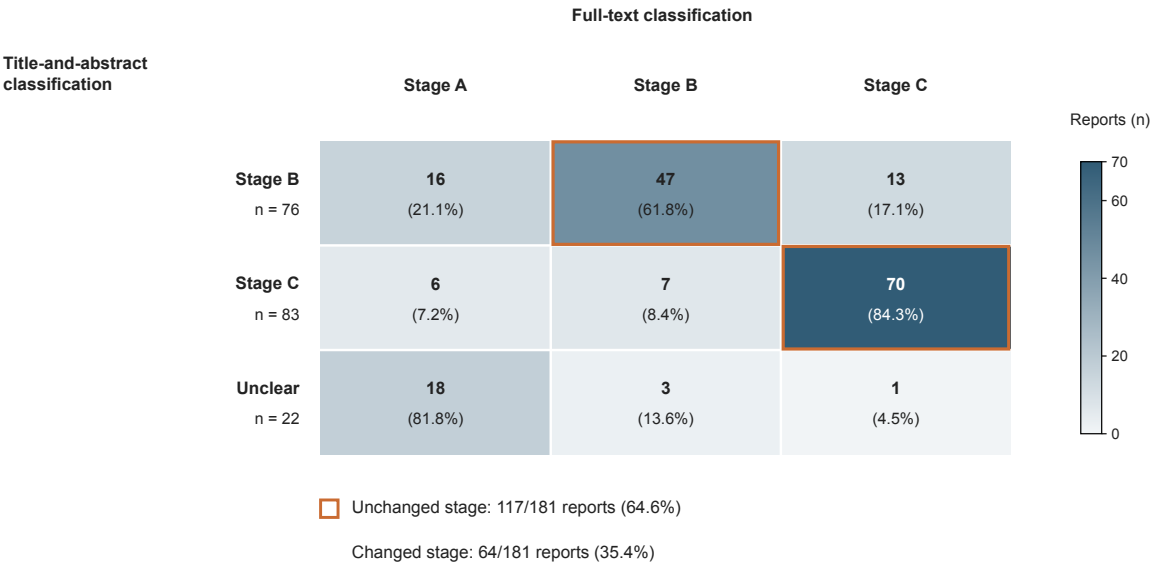
Input: "Context-Aware Concept Bottleneck Models: Integrating LLM Insights for Enhanced ARDS Identification." Source: European Respiratory Journal 2025;66(suppl 69):PA995; publication type: Conference Abstract; the publisher states that no full-text version is available.

```json
{
  "decision": "exclude",
  "exclusion_reason": "C1: not an eligible full empirical paper",
  "rationale": "The publisher labels the record a Conference Abstract and states that no full-text version is available; therefore C1 fails."
}
```

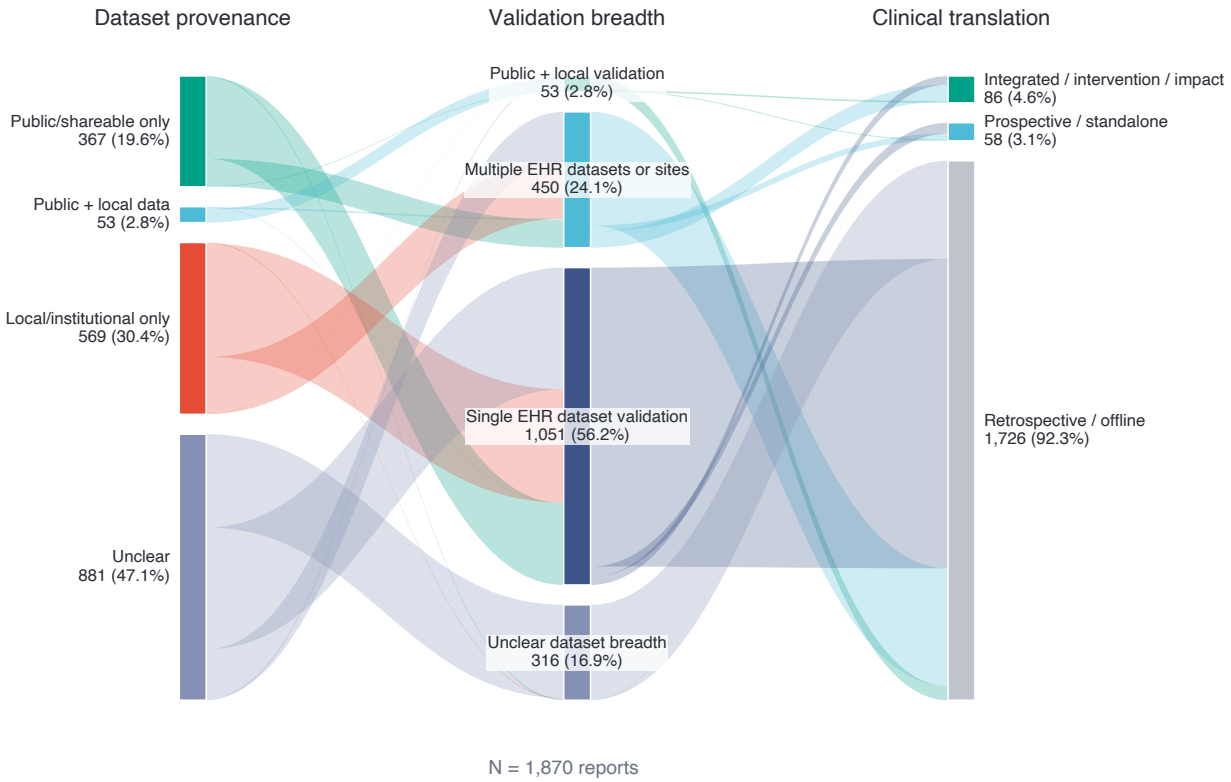
Example 5: exclude, medical-knowledge question answering

Input: "How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment." This JMIR Medical Education article evaluates ChatGPT using AMBOSS and NBME examination questions.

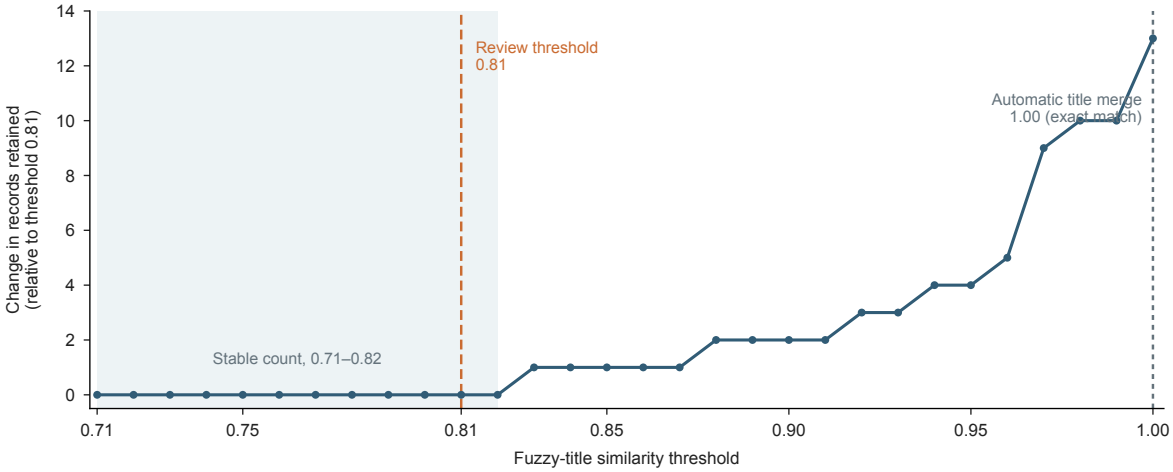
```json
{
  "decision": "exclude",
  "exclusion_reason": "C3: no real-world EHR-derived data used",
  "rationale": "ChatGPT is evaluated on AMBOSS and NBME examination questions rather than real-world EHR-derived data, so C3 fails."
}
```



Supplementary Fig. 1: Translation-stage changes among eligible reviewed reports. Rows show initial title-and-abstract labels and columns show full-text labels for the same 181 reports. Cell labels give counts and within-row percentages; colour intensity represents counts on a linear scale. Orange outlines identify unchanged stages. Translation stage changed for 64 reports (35.4%). The two title-only Stage A records reviewed because they had no abstracts were excluded as non-empirical publication types and therefore contribute no cells to this eligible-report comparison.



Supplementary Fig. 2: Evidence-atlas pathways from reported EHR dataset provenance to validation breadth and clinical translation in the adjusted map of 1,870 reports. Ribbon width is proportional to the number of reports following each pathway. Node labels show counts and percentages using all retained reports as the denominator. Categories were assigned from titles, abstracts and metadata, with the two primary dimensions replaced by full-text classifications for the 181 eligible reviewed reports. Unclear categories indicate insufficient information in the available evidence. The interactive atlas links these aggregate pathways to study-level profiles.



Supplementary Fig. 3: Sensitivity of record count to the fuzzy-title threshold. The curve shows the change in records that would be retained under automatic fuzzy-title merging relative to a threshold of 0.81. The count was unchanged from 0.71 to 0.82 and increased as the threshold became more restrictive. A similarity of at least 0.81 was used only to flag candidate pairs for human review; automatic title merging required an exact normalised-title match (1.00). Human review of fuzzy-title and metadata-conflict groups confirmed 23 of 26 groups as duplicates and retained three separately.

Supplementary References

[1] L Zhang et al. Constructing a Large Language Model to Generate Impressions from Findings in Radiology Reports. *Radiology*, 2024. https://doi.org/10.1148/radiol.240885.

[2] HJ Dai et al. Evaluating a Natural Language Processing-Driven, AI-Assisted International Classification of Diseases, 10th Revision, Clinical Modification, Coding System for Diagnosis Related Groups in a Real Hospital Environment: Algorithm Development and Validation Study. *Journal of medical Internet research*, 2024. https://doi.org/10.2196/58278.

[3] K Berigan et al. The Impact of Large Language Model-Generated Radiology Report Summaries on Patient Comprehension: A Randomized Controlled Trial. *Journal of the American College of Radiology : JACR*, 2024. https://doi.org/10.1016/j.jacr.2024.06.018.

[4] S Yan et al. Prompt engineering on leveraging large language models in generating response to InBasket messages. *Journal of the American Medical Informatics Association : JAMIA*, 2024. https://doi.org/10.1093/jamia/ocae172.

[5] M Afshar et al. Prompt engineering with a large language model to assist providers in responding to patient inquiries: a real-time implementation in the electronic health record. *JAMIA open*, 2024. https://doi.org/10.1093/jamiaopen/ooae080.

[6] J Feldman et al. Scaling Note Quality Assessment Across an Academic Medical Center with AI and GPT-4. *NEJM Catalyst Innovations in Care Delivery*, 2024. https://doi.org/10.1056/cat.23.0283.

[7] CMHHT Bootsma-Robroeks et al. AI-generated draft replies to patient messages: exploring effects of implementation. *Frontiers in digital health*, 2025. https://doi.org/10.3389/fdgth.2025.1588143.

[8] C Hong et al. Application of unified health large language model evaluation framework to In-Basket message replies: bridging qualitative and quantitative assessments. *Journal of the American Medical Informatics Association : JAMIA*, 2025. https://doi.org/10.1093/jamia/ocaf023.

[9] AS Liang et al. Artificial intelligence-generated draft replies to patient messages in pediatrics. *JAMIA open*, 2025. https://doi.org/10.1093/jamiaopen/ooaf159.

[10] H Kim et al. A Bilingual On-Premises AI Agent for Clinical Drafting: Implementation Report of Seamless Electronic Health Records Integration in the Y-KNOT Project. *JMIR medical informatics*, 2025. https://doi.org/10.2196/76848.

[11] H Hang et al. Comparative analysis of accuracy and completeness in standardized database generation for complex multilingual lung cancer pathological reports: large language model-based assisted diagnosis system vs. DeepSeek, GPT-3.5, and healthcare professionals with varied professional titles, with task load variation assessment among medical staff. *Frontiers in medicine*, 2025. https://doi.org/10.3389/fmed.2025.1618858.

[12] X Zhang et al. Data augmented large language models for medical record generation. *Applied Intelligence*, 2025. https://doi.org/10.1007/s10489-024-05934-9.

[13] JD Oliveira et al. Development and evaluation of a clinical note summarization system using large language models. *Communications medicine*, 2025. https://doi.org/10.1038/s43856-025-01091-3.

[14] SP Shashikumar et al. Development and prospective implementation of a large language model based system for early sepsis prediction. *NPJ digital medicine*, 2025. https://doi.org/10.1038/s41746-025-01689-w.

[15] DS Wright et al. The Effect of Ambient Artificial Intelligence Scribes on Trainee Documentation Burden. *Applied clinical informatics*, 2025. https://doi.org/10.1055/a-2647-1142.

[16] H Ding et al. Evaluation and practical application of prompt-driven ChatGPTs for EMR generation. *npj Digital Medicine*, 2025. https://doi.org/10.1038/s41746-025-01472-x.

[17] M. Mahyoub, K. Dougherty, and A. Shukla. Extracting Pulmonary Embolism Diagnoses From Radiology Impressions Using GPT-4o: Large Language Model Evaluation Study. *JMIR medical informatics*, 2025. https://doi.org/10.2196/67706.

[18] Y Wu et al. An eyecare foundation model for clinical assistance: a randomized controlled trial. *Nature medicine*, 2025. https://doi.org/10.1038/s41591-025-03900-7.

[19] Y.-H. Tu, T.-H. Chang, and Y.-S. Lo. Generative AI-Assisted Nursing Handover: Enhancing Clinical Data Integration and Work Efficiency. *Studies in Health Technology and Informatics*, 2025. https://doi.org/10.3233/shti251283.

[20] M Xue et al. Home-based Dry Eye Assessment via Blink Kinematics Using mmWave and Clinical Knowledge Distillation. *ACM MobiCom 2025 - Proceedings of the 2025 the 31st Annual International Conference on Mobile Computing and Networking*, 2025. https://doi.org/10.1145/3680207.3765268.

[21] LM Moura et al. Hybrid Ambient Clinical Documentation and Physician Performance: Work Outside of Work, Documentation Delay, and Financial Productivity. *Journal of general internal medicine*, 2025. https://doi.org/10.1007/s11606-025-09979-5.

[22] K Evans et al. Impact of using an AI scribe on clinical documentation and clinician-patient interactions in allied health private practice: perspectives of clinicians and patients. *Musculoskeletal science & practice*, 2025. https://doi.org/10.1016/j.msksp.2025.103333.

[23] M. Griot, J. Vanderdonckt, and D. Yuksel. Implementation of large language models in electronic health records. *PLOS digital health*, 2025. https://doi.org/10.1371/journal.pdig.0001141.

[24] Q Ye et al. IMQC: A Large Language Model Platform for Medical Quality Control. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025. https://doi.org/10.1609/aaai.v39i28.35145.

[25] AK Indrakanti et al. Information Extraction and Summarization for Neurovascular Consultations with GPT-4o: A Clinical Case Study. *Clinical neuroradiology*, 2025. https://doi.org/10.1007/s00062-025-01538-z.

[26] C.-J. Chen, Y.-S. Fang, and C.-F. Liu. An Integrated AI Cloud Sharing Framework for Predictive AI and Generative AI in Healthcare. *Studies in Health Technology and Informatics*, 2025. https://doi.org/10.3233/shti250314.

[27] H. Chen, N. Li, H. Huang, J. Lai, and C. Wang. KDA-TCM: Evidence-Grounded Prescription Generation From Symptoms in Chinese EMRs. *Proceedings - 2025 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2025*, 2025. https://doi.org/10.1109/bibm66473.2025.11356770.

[28] A Trejo Omeñaca et al. Leave as Fast as You Can: Using Generative AI to Automate and Accelerate Hospital Discharge Reports. *Computers*, 2025. https://doi.org/10.3390/computers14060210.

[29] Y Chen et al. MedCT: A Clinical Terminology Graph for Generative AI Applications in Healthcare. *Lecture Notes in Computer Science*, 2025. https://doi.org/10.1007/978-3-031-97557-8_3.

[30] Z. Ma, R. Liu, and Y. Wang. MedStructGen: A Two-Stage Method for Medical Record Generation Using Large Language Models. *Proceedings of the International Conference on Parallel and Distributed Systems - ICPADS*, 2025. https://doi.org/10.1109/icpads67057.2025.11323208.

[31] M Afshar et al. A Novel Playbook for Pragmatic Trial Operations to Monitor and Evaluate Ambient Artificial Intelligence in Clinical Practice. *NEJM AI*, 2025. https://doi.org/10.1056/aidbp2401267.

[32] S. Robertson, V. Koppireddy, J. Cumbo, H. Rashidi, and S. Albahra. PIRO: A web-based search platform for pathology reports, leveraging large language models to generate discrete searchable insights. *Journal of pathology informatics*, 2025. https://doi.org/10.1016/j.jpi.2025.100436.

[33] M Afshar et al. A Pragmatic Randomized Controlled Trial of Ambient Artificial Intelligence to Improve Health Practitioner Well-Being. *NEJM AI*, 2025. https://doi.org/10.1056/aioa2500945.

[34] S. P. Shashikumar and S. Nemati. A Prospective Comparison of Large Language Models for Early Prediction of Sepsis. *Pacific Symposium on Biocomputing*, 2025. https://doi.org/10.1142/9789819807024_0009.

[35] A Gupta et al. Provision of Radiology Reports Simplified With Large Language Models to Patients With Cancer: Impact on Patient Satisfaction. *JCO clinical cancer informatics*, 2025. https://doi.org/10.1200/cci-24-00166.

[36] P Singh et al. RedactOR: An LLM-Powered Framework for Automatic Clinical Data De-Identification. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2025. https://doi.org/10.18653/v1/2025.acl-industry.36.

[37] A Segalini et al. TanaGPT: How Can I Help You? A Pilot Testing of a Genai Assistant in a Healthcare Workflow. *Studies in health technology and informatics*, 2025. https://doi.org/10.3233/shti250813.

[38] S Mandal et al. Utilization of Generative AI-drafted Responses for Managing Patient-Provider Communication. *NPJ digital medicine*, 2025. https://doi.org/10.1038/s41746-025-01972-w.

[39] K. E. McCrudden, M. S. Swirbul, E. E. Peake, M. J. Rodio, and A. Padmanabhan. AI-Powered Documentation for Mental Health Providers: Retrospective Observational Mixed Methods Study. *JMIR formative research*, 2026. https://doi.org/10.2196/84628.

[40] A Gupta et al. Artificial intelligence enhances screening efficiency for inflammatory bowel disease clinical trials: a prospective study. *Inflammatory bowel diseases*, 2026. https://doi.org/10.1093/ibd/izag021.

[41] Y Zhu et al. Augmenting Clinical Decision-Making with an Interactive and Interpretable AI Copilot: A Real-World User Study with Clinicians in Nephrology and Obstetrics. *Conference on Human Factors in Computing Systems - Proceedings*, 2026. https://doi.org/10.1145/3772318.3791458.

[42] S Damani et al. Automating the segmentation, date extraction, and classification of multi-report PDFs in outside medical records using optical character recognition and generative artificial intelligence. *JAMIA open*, 2026. https://doi.org/10.1093/jamiaopen/ooag027.

[43] M Mustafa et al. Awaaz-e-Sehat: A Mobile Voice-based AI System for EMR Generation and Clinical Decision Support in Low-resource Maternal Healthcare. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2026. https://doi.org/10.1145/3790115.

[44] W Lin et al. Brain compensation mechanisms of large language models in clinical decision-making in acupuncture: A fusion study using fNIRS and eye-tracking. *Expert Systems with Applications*, 2026. https://doi.org/10.1016/j.eswa.2025.130870.

[45] A Mohapatra et al. Design and implementation of an end-to-end AI-driven colonoscopy recall workflow at scale. *JAMIA open*, 2026. https://doi.org/10.1093/jamiaopen/ooag070.

[46] MN Carlisle et al. Development and Validation of a Generative Artificial Intelligence-Based Pipeline for Automated Clinical Data Extraction From Electronic Health Records: Technical Implementation Study. *JMIR bioinformatics and biotechnology*, 2026. https://doi.org/10.2196/70708.

[47] R Wang et al. A domain-adapted large language model to support clinicians in psychiatric clinical practice. *Nature Machine Intelligence*, 2026. https://doi.org/10.1038/s42256-026-01224-w.

[48] D. Char, N. L. Downing, A. Youssef, and M. M. Mello. Ethical Assessment of Generative AI Tools for Clinical Summarization Tasks. *The American journal of bioethics : AJOB*, 2026. https://doi.org/10.1080/15265161.2026.2657875.

[49] Y Guo et al. Evaluating ambient artificial intelligence documentation: effects on work efficiency, documentation burden, and patient-centered care. *Journal of the American Medical Informatics Association : JAMIA*, 2026. https://doi.org/10.1093/jamia/ocaf180.

[50] RJ Chen et al. Integrating a Large Language Model to Streamline Nursing Handover Documentation Across Multiple Hospitals in Taiwan: Development and Implementation Study. *Journal of medical Internet research*, 2026. https://doi.org/10.2196/81604.

[51] R DiMaio et al. A Language Model for Pediatric Occupational Therapy Documentation: Model Development and Pilot Study. *JMIR AI*, 2026. https://doi.org/10.2196/73274.

[52] N. Tan. Large language model-assisted radiology reporting in a single-radiologist implementation: a retrospective cohort study interpreted through a UTAUT lens. *Abdominal radiology (New York)*, 2026. https://doi.org/10.1007/s00261-026-05524-y.

[53] JW O'Sullivan et al. A large language model for complex cardiology care. *Nature medicine*, 2026. https://doi.org/10.1038/s41591-025-04190-9.

[54] C Obong'o et al. Mixed-methods evaluation of clinician experiences and adoption patterns of an EHR-integrated generative AI-based clinical decision support uptake by clinicians in Kenya. *BMJ Digital Health & AI*, 2026. https://doi.org/10.1136/bmjdhai-2025-000207.

[55] F Grolleau et al. Physician-Reported Safety Outcomes of AI-Generated Hospital Course Summaries. *JAMA network open*, 2026. https://doi.org/10.1001/jamanetworkopen.2026.16556.

[56] V. M. Castro, T. H. McCoy, P. Verhaak, A. Ramachandiran, and R. H. Perlis. Psychiatric Documentation and Management in Primary Care With Artificial Intelligence Scribe Use. *JAMA Psychiatry*, 2026. https://doi.org/10.1001/jamapsychiatry.2025.4303.

[57] C de Margerie-Mellon et al. Reporting efficiency in diagnostic imaging: Can plug-and-play general-purpose large language models outperform conventional speech recognition? *European radiology*, 2026. https://doi.org/10.1007/s00330-026-12524-5.

[58] A Agweyu et al. Safety of a large language model-based clinical decision support system in African primary healthcare. *Nature health*, 2026. https://doi.org/10.1038/s44360-026-00082-5.

PRISMA 2020 Checklist

Section and Topic	Item #	Checklist item	Location where item is reported
TITLE			
Title	1	Identify the report as a systematic review.	Title (p. 1): A systematic review of clinical translation of large language models using electronic health records.
ABSTRACT			
Abstract	2	See the PRISMA 2020 for Abstracts checklist.	Abstract (p. 1) and accompanying abstract checklist.
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of existing knowledge.	Introduction, paragraphs 2–4 (pp. 1–2).
Objectives	4	Provide an explicit statement of the objective(s) or question(s) the review addresses.	Introduction, paragraph 5 (p. 2).
METHODS			
Eligibility criteria	5	Specify the inclusion and exclusion criteria for the review and how studies were grouped for the syntheses.	Methods: Search strategy and eligibility criteria (pp. 9–10); Table 1 (p. 16).
Information sources	6	Specify all databases, registers, websites, organisations, reference lists and other sources searched or consulted to identify studies. Specify the date when each source was last searched or consulted.	Methods: Search strategy and eligibility criteria (pp. 9–10); Supplementary Note 1 (Supplementary pp. 1–2).
Search strategy	7	Present the full search strategies for all databases, registers and websites, including any filters and limits used.	Supplementary Note 1: Executable search queries (Supplementary pp. 1–2).
Selection process	8	Specify the methods used to decide whether a study met the inclusion criteria of the review, including how many reviewers screened each record and each report retrieved, whether they worked independently, and if applicable, details of automation tools used in the process.	Methods: Agent-assisted screening and profiling with human validation (pp. 10–11); Targeted full-text review and adjusted evidence map (pp. 11–12).
Data collection process	9	Specify the methods used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and if applicable, details of automation tools used in the process.	Methods: Agent-assisted screening and profiling with human validation (pp. 10–11); Targeted full-text review and adjusted evidence map (pp. 11–12).
Data items	10a	List and define all outcomes for which data were sought. Specify whether all results that were compatible with each outcome domain in each study were sought (e.g. for all measures, time points, analyses), and if not, the methods used to decide which results to collect.	Table 1 (p. 16); Supplementary Tables 7–8 (Supplementary p. 14); Supplementary Note 5: Evidence profiling prompt (Supplementary pp. 19–22).
	10b	List and define all other variables for which data were sought (e.g. participant and intervention characteristics, funding sources). Describe any assumptions made about any missing or unclear information.	Table 1 (p. 16); Supplementary Note 5: Evidence profiling prompt (Supplementary pp. 19–22).
Study risk of bias assessment	11	Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked independently, and if applicable, details of automation tools used in the process.	Discussion, limitations paragraph (pp. 8–9).
Effect measures	12	Specify for each outcome the effect measure(s) (e.g. risk ratio, mean difference) used in the synthesis or presentation of results.	Methods: Data synthesis and atlas construction (pp. 12–13).
Synthesis methods	13a	Describe the processes used to decide which studies were eligible for each synthesis (e.g. tabulating the study intervention characteristics and comparing against the planned groups for each synthesis (item #5)).	Methods: Targeted full-text review and adjusted evidence map (pp. 11–12); Data synthesis and atlas construction (pp. 12–13).
	13b	Describe any methods required to prepare the data for presentation or synthesis, such as handling of missing summary statistics, or data conversions.	Methods: Record management and publication reconciliation (p. 10); Data synthesis and atlas construction (pp. 12–13).

Section and Topic	Item #	Checklist item	Location where item is reported
	13c	Describe any methods used to tabulate or visually display results of individual studies and syntheses.	Methods: Data synthesis and atlas construction (pp. 12–13); Figs. 1–5 (pp. 17–21); Supplementary Tables 1–11 (Supplementary pp. 3–16).
	13d	Describe any methods used to synthesize results and provide a rationale for the choice(s). If meta-analysis was performed, describe the model(s), method(s) to identify the presence and extent of statistical heterogeneity, and software package(s) used.	Methods: Data synthesis and atlas construction (pp. 12–13).
	13e	Describe any methods used to explore possible causes of heterogeneity among study results (e.g. subgroup analysis, meta-regression).	Methods: Data synthesis and atlas construction (pp. 12–13).
	13f	Describe any sensitivity analyses conducted to assess robustness of the synthesized results.	Methods: Preprint sensitivity analysis (p. 12); Supplementary Note 4 (Supplementary p. 16); Supplementary Fig. 3 (Supplementary p. 27).
Reporting bias assessment	14	Describe any methods used to assess risk of bias due to missing results in a synthesis (arising from reporting biases).	No formal reporting-bias assessment reported.
Certainty assessment	15	Describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome.	No formal certainty-of-evidence assessment reported.
RESULTS			
Study selection	16a	Describe the results of the search and selection process, from the number of records identified in the search to the number of studies included in the review, ideally using a flow diagram.	Results: Search, screening and human validation (pp. 2–3); Fig. 1 (p. 17).
	16b	Cite studies that might appear to meet the inclusion criteria, but which were excluded, and explain why they were excluded.	Results: Search, screening and human validation (pp. 2–3); Fig. 1 (p. 17).
Study characteristics	17	Cite each included study and present its characteristics.	Supplementary Table 4 (Supplementary pp. 5–12) and Supplementary References (Supplementary pp. 28–30); Data availability: study-linked atlas (p. 12).
Risk of bias in studies	18	Present assessments of risk of bias for each included study.	Discussion, limitations paragraph (pp. 8–9).
Results of individual studies	19	For all outcomes, present, for each study: (a) summary statistics for each group (where appropriate) and (b) an effect estimate and its precision (e.g. confidence/credible interval), ideally using structured tables or plots.	Supplementary Table 4 (Supplementary pp. 5–12); Data availability: study-linked atlas (descriptive study profiles; p. 12).
Results of syntheses	20a	For each synthesis, briefly summarise the characteristics and risk of bias among contributing studies.	Results (pp. 2–6); Supplementary Tables 3–11 (Supplementary pp. 4–16). Risk of bias not assessed.
	20b	Present results of all statistical syntheses conducted. If meta-analysis was done, present for each the summary estimate and its precision (e.g. confidence/credible interval) and measures of statistical heterogeneity. If comparing groups, describe the direction of the effect.	Results (pp. 2–6); Figs. 2–5 (pp. 18–21); Supplementary Tables 1–11 (Supplementary pp. 3–16).
	20c	Present results of all investigations of possible causes of heterogeneity among study results.	Results: Publication growth and evidence maturity (p. 3); Supplementary Table 5 (Supplementary p. 13).
	20d	Present results of all sensitivity analyses conducted to assess the robustness of the synthesized results.	Results: Preprint sensitivity and interactive atlas (pp. 5–6); Supplementary Note 4 and Table 11 (Supplementary p. 16); Supplementary Fig. 3 (Supplementary p. 27).
Reporting biases	21	Present assessments of risk of bias due to missing results (arising from reporting biases) for each synthesis assessed.	No formal reporting-bias assessment reported.
Certainty of evidence	22	Present assessments of certainty (or confidence) in the body of evidence for each outcome assessed.	No formal certainty-of-evidence assessment reported.

PRISMA 2020 Checklist

Section and Topic	Item #	Checklist item	Location where item is reported
DISCUSSION			
Discussion	23a	Provide a general interpretation of the results in the context of other evidence.	Discussion (pp. 6–9).
	23b	Discuss any limitations of the evidence included in the review.	Discussion, paragraphs 3–8 and limitations paragraph (pp. 6–9).
	23c	Discuss any limitations of the review processes used.	Discussion, limitations paragraph (pp. 8–9).
	23d	Discuss implications of the results for practice, policy, and future research.	Discussion, paragraphs addressing research, practice and policy, and concluding paragraph (pp. 8–9).
OTHER INFORMATION			
Registration and protocol	24a	Provide registration information for the review, including register name and registration number, or state that the review was not registered.	Methods: Search strategy and eligibility criteria (pp. 9–10).
	24b	Indicate where the review protocol can be accessed, or state that a protocol was not prepared.	Methods: Search strategy and eligibility criteria (pp. 9–10).
	24c	Describe and explain any amendments to information provided at registration or in the protocol.	Methods: Search strategy and eligibility criteria (pp. 9–10).
Support	25	Describe sources of financial or non-financial support for the review, and the role of the funders or sponsors in the review.	Acknowledgments (p. 13).
Competing interests	26	Declare any competing interests of review authors.	Competing interests (p. 13).
Availability of data, code and other materials	27	Report which of the following are publicly available and where they can be found: template data collection forms; data extracted from included studies; data used for all analyses; analytic code; any other materials used in the review.	Data availability and Code availability (pp. 12–13); Supplementary Note 5 (Supplementary pp. 17–24).

From: Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. doi: 10.1136/bmj.n71. This work is licensed under CC BY 4.0. To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/

PRISMA 2020 for Abstracts checklist

Locations refer to the manuscript abstract (p. 1) unless otherwise stated.

Item	Abstract requirement	Location or explanation
1	Systematic review in title	Title (p. 1).
2	Review objective	Abstract, opening sentence (p. 1).
3	Eligibility criteria	Abstract (p. 1): peer-reviewed empirical reports using EHR data, published between 1 January 2022 and 1 June 2026.
4	Information sources and dates	Abstract (p. 1): four databases.
5	Risk-of-bias methods	Not reported in the abstract. No formal study-level risk-of-bias assessment was undertaken (Discussion, limitations paragraph, pp. 8–9).
6	Synthesis methods	Abstract (p. 1): descriptive evidence mapping using agent-assisted screening and profiling, human validation and targeted full-text review.
7	Included studies and participants	Abstract (p. 1): 1,870 reports, including 181 with full-text primary classifications. Participant totals were not aggregated for this evidence map.
8	Principal results	Abstract (p. 1): principal percentages for validation breadth and clinical translation.
9	Limitations of evidence	Abstract (p. 1): predominantly retrospective/offline evaluations and limited evidence connecting broader validation with clinical use.
10	Interpretation	Abstract, final two sentences (p. 1).
11	Funding	Not reported in the abstract. Funding is reported in Acknowledgments (p. 13).
12	Registration	Not reported in the abstract. Registration is reported in Methods: Search strategy and eligibility criteria (pp. 9–10; OSF: TCM8Q).

Source: PRISMA 2020 for Abstracts, https://www.prisma-statement.org/abstracts.