

Review

Clinical Text De-Identification Beyond PHI Detection: A Scoping Review

Murat Sariyar 

Institute for Optimization and Data Analysis (IODA), School of Engineering and Computer Science, Bern
University of Applied Sciences, 2502 Bienne, Switzerland; murat.sariyar@bfh.ch

Abstract

Background: Removing explicit protected health information (PHI) does not necessarily make a clinical narrative non-identifiable. Rare diagnoses, social circumstances, locations, temporal patterns, and treatment trajectories may still permit re-identification. **Methods:** This PRISMA-ScR-informed scoping review combined PubMed/MEDLINE, OpenAlex, and targeted NLP searches with OpenAlex title-and-abstract, Web of Science, and Scopus sensitivity checks. Eligible publications addressed clinical or biomedical text de-identification from 1 January 2018 to 31 March 2026. **Results:** The searches yielded 275 records and 258 unique records after deduplication. Eligibility assessment produced 110 core records, 11 background or review records, and 22 exclusions. All 143 eligibility-stage records underwent independent double screening and reconciliation; structured extraction covered all 110 core records. Abstract screening of 47 additional Scopus candidates found no task function or implementation paradigm outside the proposed classification. The resulting two-axis framework separates a method's role in the workflow from its technical implementation. **Conclusions:** Clinical text de-identification cannot be reduced to named-entity recognition. Local hybrid or transformer pipelines remain the most defensible baseline for routine PHI detection. LLMs are better suited to defined tasks in augmentation, transformation, and assurance, with local validation and explicit control of data exposure.

Keywords: clinical text de-identification; protected health information; anonymization; natural language processing; large language models; re-identification risk

1. Introduction

Clinical narratives are among the richest data sources in health care. Discharge summaries, nursing notes, radiology reports, pathology reports, psychiatric assessments, outpatient letters, and emergency department notes contain details often absent from structured electronic health record fields, including diagnostic reasoning, disease progression, treatment decisions, social circumstances, and functional status. Their value for research, quality improvement, decision support, epidemiology, and NLP development depends on privacy-preserving access to narrative text [1–4].

The same texts also contain sensitive and identifying information. Names, dates, addresses, phone numbers, medical record numbers, institutional names, and locations are explicit examples, but clinical identifiability often comes from context. Removing PHI spans is therefore not equivalent to anonymization. Rare diagnoses, uncommon life events, occupational details, family relationships, small geographic areas, or highly specific treatment trajectories can leave a patient recognizable even when conventional identifiers have been removed. Studies of hostile re-identification, downstream utility,



Academic Editor: Rüdiger Pryss

Received: 10 June 2026

Revised: 9 July 2026

Accepted: 14 July 2026

Published: 24 July 2026

Copyright: © 2026 by the author.

Licensee MDPI, Basel, Switzerland.

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

and synthetic or de-identified clinical notes show why residual context matters, especially in psychiatric notes, genetic reports, rare disease documentation, and longitudinal clinical narratives [5–8].

Text de-identification refers to the detection and removal, masking, replacement, generalization, or transformation of identifying information in free text. In the United States, many studies operationalize the task through HIPAA protected health information (PHI) categories and the Privacy Rule's de-identification pathways of Expert Determination or Safe Harbor [9]. In Europe, the legal framing differs. Under the General Data Protection Regulation (GDPR), pseudonymized data can remain personal data if re-identification is possible with additional information, while anonymous data require that the data subject is not or is no longer identifiable [10]. The EU Artificial Intelligence Act adds a further governance layer for AI systems, including documentation, risk management, human oversight, monitoring, and accountability expectations that become relevant when de-identification is embedded in automated clinical data processing workflows [11]. These distinctions matter for clinical free text, where a fixed list of entity types may miss identifying context [3,12].

Clinical text de-identification has moved from rules, regular expressions, dictionaries, and engineered features to CRFs, recurrent neural networks, BiLSTM-CRF architectures, and transformer encoders. Since 2019, BERT family models have generally framed de-identification as token classification or named-entity recognition, often with high F1 scores on established datasets such as i2b2/n2c2 [13–19].

Since 2023, LLMs have expanded the scope of the task. They can identify PHI, replace values with plausible surrogates, rewrite passages, generalize sensitive details, produce synthetic variants for training, or simulate adversarial re-identification. These uses introduce new risks: cloud processing may expose raw text before de-identification has succeeded, prompts and model versions may not be reproducible, rewriting can alter clinical meaning, and surface-level performance may conceal failure on rare or contextual identifiers [6,20–25].

The recent literature is no longer organized around one benchmark task. Kovacevic and colleagues reviewed English language clinical free-text de-identification up to February 2023 and found dominant machine learning and hybrid systems, high reported F1 scores, and persistent limits in corpus diversity and domain adaptation [12]. Since then, DeID-GPT, GPT-4 augmentation, temporal normalization, multi-specialty LLM benchmarks, information-loss analyses, and adversarial re-identification studies have expanded the task from PHI span detection to process design [6,20–23,26,27]. The relevant question is no longer only which model detects PHI best, but which processing route controls disclosure risk while preserving the clinical information needed for the intended use. LLMs are therefore treated as one component of risk-adapted de-identification, with method choice organized by workflow function rather than model family alone.

2. Objectives and Research Questions

The review covers clinical text de-identification published from 1 January 2018 to 31 March 2026. It examines implementation paradigms and task functions; the representation of datasets, domains, languages, document types, and PHI/PII categories; gaps in measures of disclosure risk, transformation quality, and clinical utility; and the conditions under which local encoders, hybrid systems, or LLM-assisted workflows are justified. A two-axis classification links detection, transformation, adaptation, assurance, and governance to technical implementation and deployment context.

3. Methods

The study was designed as a PRISMA-ScR-informed scoping review [28]. A scoping review clarifies concepts, identifies task functions, maps method families, and shows where evidence clusters or is missing in an emerging or heterogeneous field [29–31]. It is systematic in procedure but scoping in purpose. Unlike an effect-focused systematic review, it does not seek pooled effect estimation or a meta-analytic ranking of model families. This design fits a field in which studies address detection, transformation, utility preservation, residual-risk testing, and governance rather than one shared intervention or endpoint.

The workflow combined reproducible searches, retained logs, explicit eligibility rules, independent double screening of the 143 eligibility-stage records, reconciliation with agreement statistics, structured extraction for all 110 core records, endpoint-coverage coding, closed-database checks, and strength-of-evidence mapping.

The reproducible search component used PubMed/MEDLINE through the NCBI E-utilities API, OpenAlex, and targeted NLP-oriented searches. Deduplication, eligibility decisions, independent double screening, reconciliation, extraction, and synthesis are reported from search record to final category.

3.1. Eligibility Criteria

Eligibility criteria separated clinical free-text de-identification from broader privacy, security, and structured-data anonymization studies. Table 1 summarizes the operational rules. The criteria include technical systems and studies of residual privacy risk or clinical utility because recent work treats de-identification as more than span detection.

Table 1. Eligibility criteria.

Criterion	Inclusion	Exclusion
Time frame	Publications from 1 January 2018 to 31 March 2026; older papers may be cited only as background	Studies before 2018 unless essential for historical context; records indexed after the cut-off were not used as core evidence
Data type	Free text, clinical narratives, biomedical narratives, EHR notes, discharge summaries, nursing notes, radiology reports, psychiatric notes, pathology reports	Structured tabular anonymization without a free-text component; image-only de-identification
Task	De-identification, anonymization, pseudonymization, PHI detection, PII detection, masking, replacement, surrogate generation, re-identification audit	General privacy discussion without NLP/text methods
Domain	Clinical, medical, biomedical, health care, public health text	Non-health domains unless methodologically central and transferable
Methods	Rule-based, hybrid, CRF, BiLSTM-CRF, transformer, LLM, prompting, data augmentation, re-identification testing	Papers not describing or evaluating a relevant method, dataset, framework, or review
Evidence type	Peer-reviewed articles, conference papers, high-impact preprints, datasets, shared-task reports, systematic or scoping reviews	Editorials, opinion pieces, product pages, non-technical blog posts
Language	English and German; other languages included if abstract and core results are accessible	No accessible abstract or insufficient methodological detail

Structured k-anonymity, image de-identification, and general governance papers were retained only when they directly informed free-text de-identification, where privacy protection depends on contextual identifiability, annotation scope, and document genre rather than only explicit identifiers.

3.2. Information Sources and Open-Database Rationale

The search used an open-database scoping design to keep the review repeatable across an interdisciplinary field. PubMed/MEDLINE captured biomedical and clinical informatics work; ACL Anthology, arXiv, and targeted web searches captured NLP, engineering, and recent LLM papers; and OpenAlex broadened coverage beyond PubMed without

subscription-only exports. IEEE, ACM, AMIA, MEDINFO, medRxiv, and bioRxiv were not treated as separate counted database strata. Relevant records from these communities were sought through OpenAlex indexing, ACL/arXiv searches, PubMed overlap where applicable, and targeted title/source checks documented in the audit log.

OpenAlex documentation identifies OpenAlex as a freely accessible bibliographic database with API-based access, while Scopus and Web of Science are provider-controlled commercial citation databases [32–34]. OpenAlex was used as the main cross-disciplinary open expansion layer, not as a one-to-one substitute for Scopus. Three checks assessed search stability: a Web of Science institutional search, an OpenAlex title-and-abstract search, and an institutional Scopus screen using the main TITLE-ABS-KEY query. Scopus was handled separately because its source selection, metadata fields, TITLE-ABS-KEY indexing, and coverage of conference proceedings, books, trade journals, and regional journals are provider controlled [32,33]. Comparative bibliometric and systematic review methodology studies show substantial overlap but non-identical coverage across Scopus, Web of Science, and open bibliographic sources [35–38]. Reference lists and citing articles were checked for central reviews and landmark systems, including i2b2/n2c2-derived work, Philter, Transformer-DeID, transformer-based de-identification, and LLM-based de-identification.

3.3. Search Strategy

The search strategy combined four concept blocks: de-identification terminology, clinical free-text terminology, NLP method terminology, and health-domain terminology. The de-identification block included spelling variants of de-identification, anonymization, pseudonymization, PHI, and PII. The text block included free text, clinical notes, electronic health records, discharge summaries, nursing notes, and radiology reports. The method block covered NLP, named-entity recognition, sequence labeling, CRF, LSTM, BiLSTM, BERT, RoBERTa, ClinicalBERT, transformers, large language models, GPT, and prompting. The domain block restricted retrieval to clinical, medical, health care, or biomedical contexts.

The full PubMed/MEDLINE string, the Scopus sensitivity query, Web of Science sensitivity query, targeted LLM strings, the OpenAlex query, and the OpenAlex title-and-abstract sensitivity query are reported in Supplementary File S2. The eligibility window was restricted to publications from 1 January 2018 to 31 March 2026, reflecting the period in which contextual embeddings, transformers, and LLM-based workflows became central to the field.

Targeted searches covered GPT-4, next-generation generative models, large language models, ClinicalBERT, German medical text, re-identification, and clinical information loss because emerging LLM work may appear in conference proceedings, preprints, or NLP venues before biomedical indexing. The original targeted search string included GPT-5 as an anticipatory term for later generative-model iterations; no mature GPT-5-specific clinical de-identification evidence was identified, and GPT-5-specific claims were therefore not used as a basis for the synthesis.

3.4. Selection Process

Records were deduplicated using normalized DOI where available and normalized title otherwise. This first deduplication step was applied before screening to remove bibliographically identical records retrieved from more than one open source. Title-level screening identified records likely to address clinical or biomedical text de-identification, PHI/PII detection, privacy leakage, anonymization, pseudonymization, or re-identification. Records were classified as title-level include candidates, direct title-level exclusions, or title/abstract assessment records. The title/abstract assessment records were assessed using

abstracts where available and title/source metadata where abstracts were not available in the retained export.

Title-level include candidates underwent eligibility assessment. Reports were included as core evidence when they addressed clinical or biomedical free-text de-identification, anonymization, pseudonymization, PHI/PII detection, surrogate generation, utility preservation, or re-identification risk. Reviews, terminology papers, governance papers, and adjacent PHI/privacy studies were retained as background evidence when they informed interpretation. Reports were excluded when they were duplicates, outside the date range, focused on structured or non-text de-identification, or lacked direct relevance. Duplicate or secondary-report exclusions at eligibility captured version records or overlapping reports identifiable only after closer inspection. Reasons are recorded in Supplementary Table S1, sheet Eligibility_143. The 35 title/abstract assessment records are documented in Supplementary Table S1, sheet Title_Abstract_Assessment, with assessment basis, abstract availability, decision, category, and explicit reason.

Screening was conducted as an independent double-screening workflow using predefined eligibility criteria, explicit exclusion categories, retained search logs, and tabular audit trails. In line with PRISMA-ScR reporting, no protocol was registered [28]. The author completed the initial title and eligibility screening, and a scientific staff member (MM) independently screened all 143 eligibility-stage records using the same predefined criteria before reconciliation. Agreement before reconciliation was 112/143 across harmonized eligibility decisions. Cohen's kappa was 0.552 across harmonized eligibility categories and 0.808 for the binary retain-versus-exclude decision. Most disagreements concerned the finer distinction between core evidence, background evidence, duplicate/secondary reporting, and uncertain records. Decision-level disagreements were reconciled against the predefined eligibility criteria and retained in a reconciliation log. Hard exclusion criteria took precedence over topical relevance. Records were retained as core evidence when they directly evaluated or developed clinical or biomedical free-text de-identification, PHI/PII detection, surrogate generation, utility preservation, or re-identification-risk assessment. Records were retained as background evidence when they informed terminology, governance, privacy risk, datasets, benchmarks, or review context without serving as primary method evidence. Bibliographic details, study type, dataset description, method family, evaluation target, and utility or re-identification fields were manually checked for the 110-record core extraction.

The 110 core records informed the qualitative synthesis and are included in Supplementary Table S1, sheets Core_Extraction_110 and Quantitative_Anchor_110. Each core record is retained with bibliographic fields, source, study type, document/domain signal, language signal, method/topic signal, evaluation focus, utility or re-identification signal, and extraction status. All 110 records were screened for endpoint information available from the documented extraction basis. Endpoint detail is used for concrete examples and endpoint display, not as the boundary of the evidence base. Supplementary Table S1, sheet Endpoint_Coverage, compares this layer with the full 110-record core corpus.

The PRISMA flow diagram reports identification, deduplication, title-level screening, eligibility assessment, exclusion reasons, and studies included in the qualitative synthesis. The reported PRISMA counts remain restricted to the open-database search and targeted supplementary searches; the Web of Science results are reported as a closed-database sensitivity check in Supplementary Table S1, sheet WoS_Sensitivity.

3.5. Data Charting

Data extraction used a structured charting form because studies differ in architecture, document type, language, PHI schema, deployment context, and evaluation logic. The form

captured bibliographic data, study type, clinical context, dataset characteristics, method family, model or prompt details, output type, evaluation metrics, privacy assessment, utility assessment, generalizability, and limitations. It separates detection performance from privacy and utility assessment, since high token-level F1 can coexist with contextual identifiers or clinically damaging overmasking. Extraction was completed for all core records in Supplementary Table S1, sheet Core_Extraction_110. That sheet contains the complete bibliographic record for each of the 110 included studies; the main reference list contains only works cited in the narrative. Record-level source fields and extraction-basis notes are documented in sheet Source_Extraction_Basis. Endpoint information was charted only to the depth supported by the documented source fields. All descriptive results and interpretations use the full core corpus.

3.6. Synthesis

Because studies differ in datasets, annotation schemes, model families, PHI categories, metrics, and deployment assumptions, a formal pooled performance meta-analysis was not appropriate. The quantitative synthesis therefore used study-level proportions with exact binomial 95% confidence intervals for corpus composition, endpoint availability, method signals, and evidence gaps. Endpoint values such as F1, precision, recall, leakage, utility loss, or throughput were extracted where source-level fields supported them, but were not pooled across incompatible corpora and label schemas.

Strength-of-evidence mapping connected the scoping objective with the review's deployment recommendations. Following evidence mapping logic, it identified where evidence was concentrated, where it was thin, and what type of inference each evidence cluster could support [39,40]. The mapping calibrated whether a record supported conceptual classification, endpoint display, workflow guidance, or deployment conclusions. Dimensions included publication and study type, dataset transparency, extraction-basis documentation, endpoint availability, direct comparative or benchmark evaluation, external validation, reproducibility of code/model/prompt details, deployment proximity, residual-PHI or re-identification analysis, utility assessment, privacy governance, and sensitivity-search behavior. Results were synthesized by method family, clinical domain, language, dataset type, evaluation approach, and generalizability. LLM-oriented reports were reviewed with explicit attention to peer-review status, external validation, and processing setting.

4. Results

4.1. Study Selection

The PubMed/MEDLINE search identified 145 records. Title-level screening identified 88 potentially relevant PubMed records containing de-identification, anonymization, pseudonymization, PHI, privacy leakage, or re-identification terms in the title. Targeted ACL Anthology and web searches identified seven additional methodologically relevant records that were not consistently captured in PubMed-oriented searching. An OpenAlex title search for de-identification records within the eligibility window identified 123 records and broadened coverage through an open bibliographic index.

Across PubMed, targeted ACL/web records, and OpenAlex, 275 raw records were identified. Deduplication by normalized DOI and normalized title removed 17 records, leaving 258 unique records for preliminary title-level screening. Of these, 143 were marked as title-level include candidates and assessed for eligibility. Eighty records were direct title-level exclusions. The remaining 35 records underwent title/abstract assessment with explicit record-level decisions. This assessment identified 15 background or adjacent records and 20 exclusions. The exclusions were mainly outside the eligibility window, medical-image or DICOM de-identification papers, structured-data or generic pri-

vacy/governance papers, non-clinical text de-identification papers, data-extraction studies in which anonymization was only a support step, or records lacking accessible abstract-level clinical free-text method detail. The supplementary audit tables preserve the screening and charting trail.

All 143 title-level include candidates were then assessed for eligibility, producing 110 core clinical/free-text de-identification records, 11 background or review records, and 22 exclusions. The 22 eligibility-stage exclusions comprised 13 duplicate or secondary reports, five records outside the eligibility window, two structured-data-only de-identification papers, one record not focused on clinical free text, and one record not focused on text de-identification. Figure 1 reports the PRISMA-ScR flow for the full 143-record eligibility process within the reproducible open-database corpus and separately reports the completed assessment of the 35 title/abstract records. Records with clinical text relevance in this title/abstract assessment were retained as background or adjacent evidence because they concerned downstream utility, infrastructure or corpus reports, surrogate replacement risk, or benchmark/method variants already represented in the core evidence map.

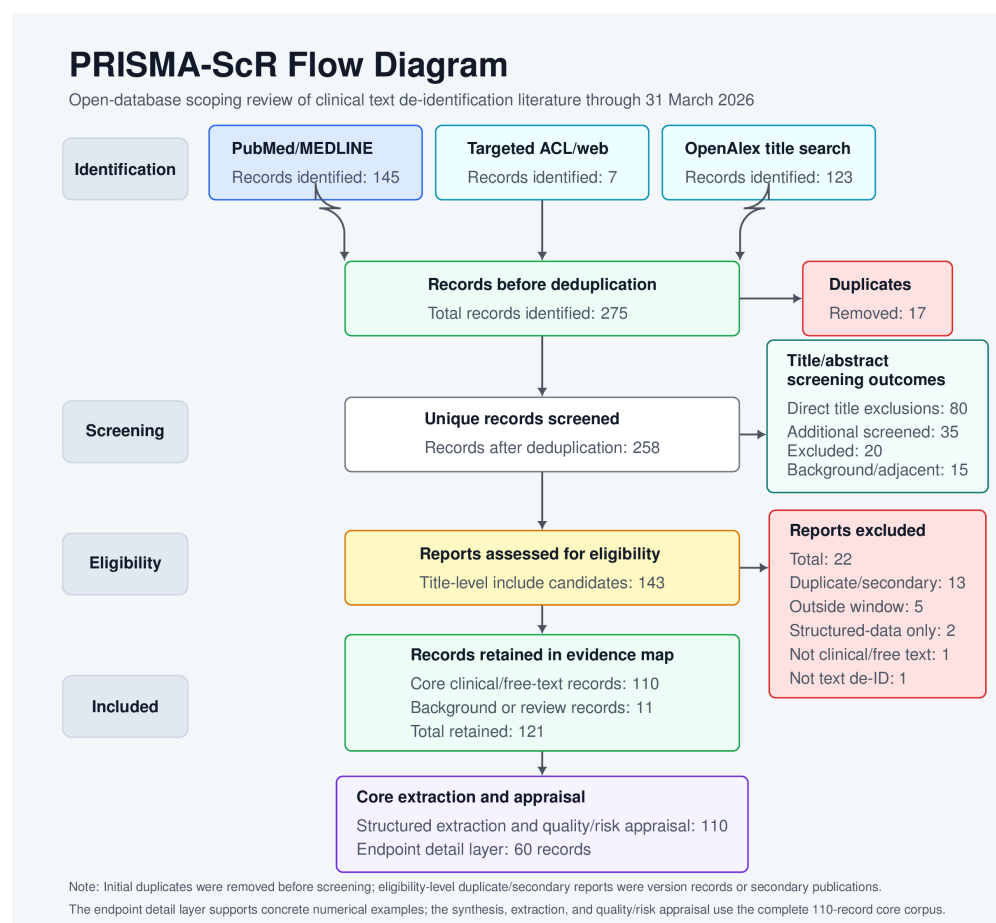


Figure 1. PRISMA-ScR flow diagram for the open-database scoping review. Arrows trace records through identification, screening, eligibility assessment, and inclusion; colors distinguish search sources and review stages for visual orientation only.

The additional OpenAlex title-and-abstract sensitivity search returned 65 records. Thirty-eight overlapped with the existing 143-record eligibility table by normalized title or DOI. Twenty-seven did not exactly match existing records; rapid title/source inspection showed that the apparently relevant non-overlapping records were image de-identification papers, secondary or preprint versions of already represented work, non-

clinical text de-identification, synthetic data or clinical NLP adjacency papers, or general privacy/governance papers.

4.2. Quantitative Evidence Summary

The principal descriptive counts and interpretive claims are based on the full 110-record core corpus. Supplementary Table S1 provides structured extraction and the quantitative anchor in sheets Core_Extraction_110 and Quantitative_Anchor_110.

The corpus shows a strong recent expansion of the field. Twenty-three records were published in 2018–2020, 35 in 2021–2023, and 52 in 2024–2026. Seventy-nine records appeared in journal or archival venues, 12 in conference or proceedings venues, 10 as preprints or repository records, and nine in unclear venue categories. Discovery sources were PubMed only for 52 records, OpenAlex only for 41, multiple PubMed/OpenAlex sources for 14, OpenAlex plus targeted search for two, and targeted search only for one. Language coding identified 38 English-only records, 33 non-English single-language records, five code-mixed or multilingual records, and 34 records where the language was not stated clearly enough in title/source metadata for confident coding. Title-based topic signals were LLM/generative for 17 records, corpus/benchmark/review for 16, hybrid/tool/pipeline for 13, NER/ML/deep learning for 12, transformer/BERT for nine, assurance/privacy/utility for eight, rule-based for three, and not inferable from title/source metadata for 32.

The corpus is dominated by method-development and detection-oriented work. In the core-extraction sheet, 57/110 records were coded as de-identification method, tool, or workflow evidence, 25/110 as PHI-detection-focused evidence, nine as comparative performance or benchmark evaluations, four as dataset or corpus creation, and the remaining records as portability, utility, privacy, re-identification, annotation, workflow, or mixed evaluation evidence. The quantitative anchor sheet classifies 60/110 records as full endpoint-charting records, 23/110 as core-level quantitative or evaluative signals requiring further endpoint verification before direct performance comparison, and 27/110 as bibliographic or evidence anchors without a source-level performance endpoint. Endpoint-level values display concrete numerical examples; they are not treated as pooled comparative performance evidence.

Table 2 reports the main quantitative synthesis with exact binomial confidence intervals. The estimates show that the evidence base is broader than the endpoint layer, and that LLM-oriented work is present but not dominant.

The institutional Web of Science sensitivity search returned 50 records. Twenty-one overlapped with the existing 143-record eligibility table by normalized title or DOI. Of the 29 non-overlapping records, five were classified as additional core sensitivity records, seven as background sensitivity records, and 17 as exclusions. The additional core records concerned hybrid discharge summary de-identification, LLM-based medical record de-identification, Norwegian synthetic de-identification benchmarks, synthetic free-text medical record generation with PHI masking, and de-identification/normalization of electronic medical record text notes.

The institutional Scopus sensitivity screen used the main TITLE-ABS-KEY query and returned 201 records. Eighty-seven records matched existing review or sensitivity records by normalized title, and five additional records were treated as already represented title variants or close conceptual duplicates pending DOI-level confirmation. Among the remaining records, title-level sensitivity screening classified 47 as additional core sensitivity candidates, 21 as background or adjacent sensitivity candidates, and 41 as exclusions. The 47 candidates broadened coverage of recent LLM-based de-identification, multilingual anonymization, pseudonymization tools, transformer/redaction systems, synthetic

or surrogate text work, and downstream-utility studies. As closed-database sensitivity evidence, these records are reported separately from the 110-record open-database core corpus. The full Scopus screening audit is provided in Supplementary Table S1, sheet Scopus_Screening.

Table 2. Quantitative synthesis of the complete 110-record core corpus.

Quantitative Anchor	Estimate	Interpretation for Main Claims
Records published in 2024–2026	52/110 (47.3%; 95% CI 37.7–57.0)	Nearly half of the core corpus is recent, including newer LLM and governance work.
Journal or archival venue	79/110 (71.8%; 95% CI 62.4–80.0)	Most records are traceable through formal venues, while venue status alone does not determine operational relevance.
Non-English or multilingual/ code-mixed signal	39/110 (35.5%; 95% CI 26.6–45.1)	Language transfer is a material deployment issue, not a marginal special case.
Broad LLM/generative signal	23/110 (20.9%; 95% CI 13.7–29.7)	LLMs are present across the corpus, but not the dominant evidence base.
Full endpoint-charting information	60/110 (54.5%; 95% CI 44.8–64.1)	Numerical endpoint display is informative but does not define the whole evidence base.
Direct comparative performance or benchmark evaluation	9/110 (8.2%; 95% CI 3.8–15.0)	Direct model ranking evidence is sparse relative to the size of the field.

All 47 additional Scopus core candidates were screened from publicly available abstracts; Supplementary Table S1 retains the records in Scopus_Additional_Core and the audit in Scopus_Abstract_Scan_47. The abstracts covered task-specific LLM use, local processing, privacy-preserving file workflows, context-aware surrogation, multilingual adaptation, pseudonymization, shared-task benchmarking, local tool validation, residual-risk assessment, synthetic data, and downstream utility. Method coding identified pseudonymization or anonymization tools (10/47), general de-identification methods (10/47), LLM or generative workflows (6/47), neural sequence models (6/47), synthetic or surrogate generation (5/47), transformer/BERT systems (5/47), hybrid or privacy-preserving machine learning approaches (3/47), and NER/PHI detection (2/47). Transformation and mixed transformation/adaptation roles were more common than detection alone. The records included Spanish, German, Chinese, Swedish, Italian, Thai, Norwegian, Indonesian, New Zealand, Australian, and Taiwanese or code-mixed settings, extending coverage of multilingual and conference literature.

4.3. Study Characteristics

Across the full 110-record corpus, included studies comprise system descriptions, comparative model evaluations, dataset papers, domain adaptation studies, multilingual studies, LLM evaluations, data augmentation work, utility preservation studies, and re-identification audits. The corpus confirms both the continuing dominance of English clinical text and the growth of Dutch, German, French, Chinese, Japanese, Korean, Spanish, Brazilian Portuguese, Swedish, and mixed-language evidence [41–48].

Studies from 2018 to 2021 emphasize contextual embeddings, recurrent sequence labeling, BERT-style models, rule-based pipelines, and hybrid systems. From 2022 onward, the literature adds transformer comparisons, low-resource settings, utility preservation, and cross-domain evaluation. From 2023 onward, GPT-4 and other LLMs appear first as zero-shot or few-shot de-identifiers and later as tools for augmentation, surrogate generation, information-loss analysis, and adversarial re-identification [15,6,17,18,20–22,27,49].

Clinical domains are unevenly represented across the corpus. General clinical notes, EHR free text, and medical records dominate, while nursing notes, psychiatric narratives, occupational health narratives, pathology reports, and some non-English clinical docu-

ments appear less consistently. Radiology and pathology reports may require different strategies because they contain fewer social details but still include dates, accession numbers, locations, clinician names, and rare diagnoses. German and Dutch studies show that English language models cannot simply be imported into European settings without local validation [41–43,46]. Japanese, Korean, Chinese, and code-mixed Chinese-English or Korean-English studies further show that whitespace tokenization assumptions do not transfer cleanly to character-based or mixed-script narratives, where name boundaries, honorifics, local institution names, transliterated terms, and administrative identifiers can be encoded differently from alphabetic English corpora [4,42,45,48].

4.4. Functional Classification of the Evidence

Studies were classified along two axes: the function addressed in de-identification and the implementation paradigm used to support it. Recent work does not map cleanly onto one-dimensional method classes. A transformer encoder may be used for PHI detection, a rule-based system may combine detection with date shifting, and an LLM may support augmentation, direct rewriting, temporal normalization, or adversarial re-identification. Supplementary Table S1, sheet Two_Axis_Classification, separates what a study contributes from how it is implemented, while Figure 2 visualizes the resulting workflow map.

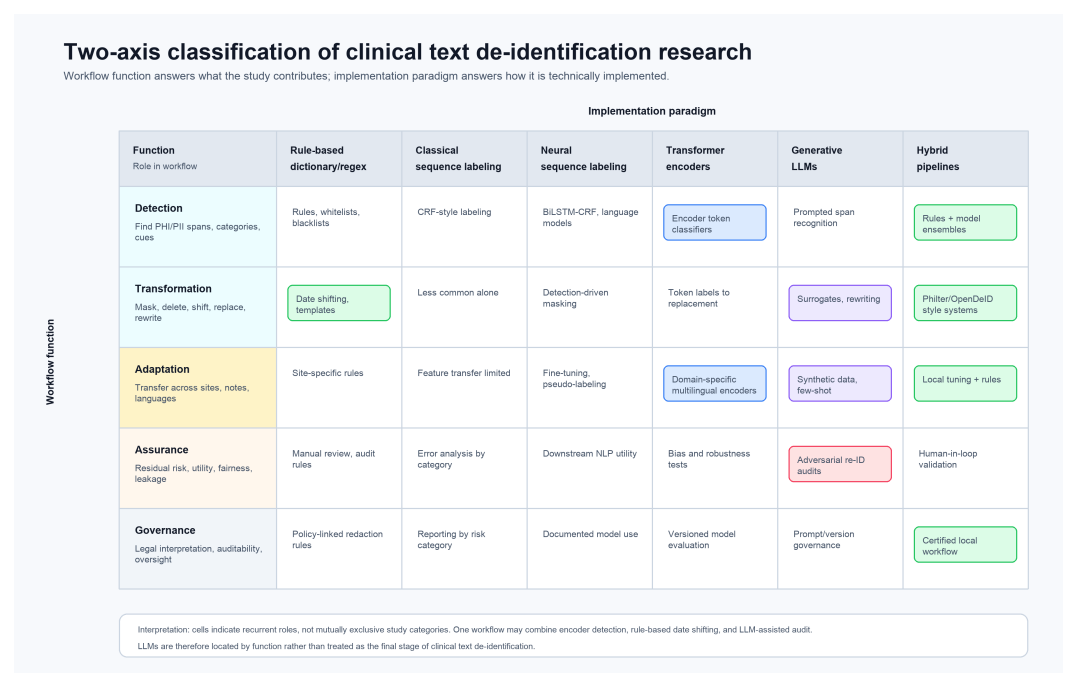


Figure 2. Workflow-function and implementation-paradigm map for clinical text de-identification research. Rows represent workflow functions and columns implementation paradigms. Colored boxes highlight recurrent combinations; the colors are visual guides and do not encode an evidence hierarchy.

The function axis captures five roles. Detection identifies PHI, PII, or sensitive contextual spans. Transformation modifies text by masking, deleting, generalizing, shifting, or replacing identifiers. Adaptation examines transfer across institutions, document types, languages, and low-resource settings. Assurance evaluates residual PHI, re-identification risk, clinical information loss, fairness, and downstream utility. Governance links technical outputs to auditability, legal interpretation, deployment setting, and human oversight.

The implementation axis captures rule-based and dictionary systems, statistical sequence labeling, neural sequence labeling, transformer encoders, generative LLM approaches, and hybrid pipelines. LLMs can appear in several functions, including augmentation, direct transformation, and assurance.

The classification reflects a shift from span detection toward risk-managed text transformation. Recent studies ask whether models adapt locally, replacements preserve temporal structure, downstream NLP performance remains stable, and residual context supports re-identification. LLMs are not simply stronger NER systems; they support augmentation, transformation, temporal normalization, or adversarial assurance, while task-specific transformers and hybrid systems remain stronger choices for routine PHI detection.

4.5. Representative Methods and Tools

BERT-family models and hybrid systems are the most established options for deployment, especially when rules handle structured identifiers and encoders handle contextual spans. Performance varies with language, document type, and institutional templates. LLM studies add prompting, augmentation, rewriting, and re-identification audits, together with reproducibility and governance problems that conventional benchmarks often miss. Representative methods appear in Table 3; Supplementary Table S1, sheet Representative_Methods, contains the full list.

Table 3. Representative methods and tools for clinical text de-identification.

Workflow Layer	Representative Procedures or Tools	Main Use	Operational Caution
Rule-based and hybrid systems	Philter; OpenDeID; CliniDeID; MIST; NLM Scrubber; pyDeid; dictionaries; regular expressions; date shifting	Structured identifiers, templated note sections, local deployment, and auditable preprocessing	Strong for predictable identifiers, but requires site-specific rules and residual-PHI audit [1,2,8,25,50,51].
Sequence-labeling baselines	CRF; BiLSTM-CRF; recurrent neural systems; NeuroNER-style pipelines	Historical baselines and PHI span detection where benchmark continuity matters	Useful comparators, but mostly surpassed by transformer encoders for new deployments [15,18,19].
Transformer encoders	BERT; RoBERTa; BioBERT; SciBERT; domain-tuned de-identification transformers	Contextual PHI detection under defined schemas and local fine-tuning	Strong default for known labels, but needs local validation for note type, language, and institution [13,14,16,17,26,42,52–54].
Weak supervision and adaptation	Regex bootstrapping; pseudo-labeling; local fine-tuning; multilingual or language-specific encoders	Low-label settings, new institutions, new languages, and domain transfer	Reduces annotation burden but can amplify weak-label bias without external validation [46,55,56].
LLM-assisted workflows	DeID-GPT; direct GPT prompting; GPT-4 augmentation; fine-tuned or LoRA-adapted LLMs; temporal-normalization prompts	Augmentation, surrogate generation, constrained rewriting, temporal normalization, and second-pass review	Most useful outside fixed-label detection; cloud use and model drift require explicit governance [20,21,23,27].
Assurance and attack testing	Manual residual-PHI audit; HIPS/parrot attacks; DIRI-style metrics; adversarial re-identification prompts; information-loss review	Residual-risk testing, utility assessment, and red-team evaluation	Assurance adds a check on residual risk but does not prove anonymity [5–8,22,49,57].

DeID-GPT used GPT-4 (OpenAI, San Francisco, CA, USA) for zero-shot removal of identifiers and compared its output with BERT, RoBERTa, and ClinicalBERT baselines [27]. The publication did not identify a more specific model snapshot. Altalla et al. also evaluated GPT models as direct clinical-note de-identifiers [21]. Dai et al. tested in-context learning, fine-tuned LLMs, LoRA-style adaptation, and pattern-based temporal normalization for sensitive-health-information recognition and date normalization in EHR text [23]. Kim et al. used GPT-4 to augment training data rather than to produce the final de-identified text [20]. Morris et al. used adversarial prompts to test whether redacted notes still allowed patient re-identification, while Aghakasiri et al. assessed clinical information loss alongside PHI removal [6,22]. These studies cover direct redaction, surrogate generation, temporal normalization, data augmentation, and adversarial testing.

4.6. Evidence Patterns and Practical Selection

PHI categories are not fully standardized. HIPAA-derived categories are common, but some corpora use more granular labels and others merge labels for modeling [9,12]. For extraction, categories were harmonized into person names, dates and ages, locations, contact information, structured identifiers, institutions, professions and social roles, family and relationship details, rare events, and clinician signatures. Apparent model performance partly depends on annotation scope: a system evaluated against a narrow label set may appear stronger than one evaluated against richer contextual identifiers [7,42,58].

Second, benchmark concentration limits inference about deployment. i2b2/n2c2-style datasets, MIMIC-derived notes, and PhysioNet-hosted resources support comparison and reproducibility, but they do not remove the need for local validation. Nursing notes, psychiatric notes, emergency department records, radiology reports, pathology reports, occupational health narratives, and multilingual corpora differ in syntax, abbreviations, entity distributions, and temporal structure. Psychiatric narratives are especially challenging because identifying information is often embedded in life events, family relations, places, occupations, and rare social circumstances [17,26,42–46,48,54,55]. For example, a sentence such as “the patient is the only night-shift police dispatcher in a small mountain town and was brought in after a highly publicized family incident” may contain no conventional name, address, phone number, or medical record number, yet still provide a linkage path for someone with local knowledge. A standard BERT-style PHI detector trained mainly on benchmark labels may correctly leave most of this sentence untouched unless the annotation schema and local validation explicitly include contextual identifiers.

Third, evaluation remains narrower than deployment risk. Most studies report precision, recall, and F1 score at token or entity level, sometimes with strict and relaxed matching. These metrics are necessary, but two systems with similar F1 may differ substantially in high-risk missed identifiers, replacement consistency, preservation of temporal relations, clinical information loss, or reproducibility. Deployment-oriented evaluation can add PHI-level recall, manual residual-PHI audit, transformation-quality assessment, downstream utility testing, adversarial re-identification where relevant, and external validation on another institution, document type, language, or time period [5–8,22,49,57,59].

Methodological maturity also varies by task. Hybrid systems and fine-tuned transformers have the most mature deployment evidence when combined with local validation, rule-based safeguards, and audit procedures. Recurrent neural models now mainly serve as baselines. LLM prompting remains insufficiently stabilized for routine direct de-identification, whereas LLM-assisted augmentation and adversarial re-identification audits may strengthen local systems without making a cloud LLM the production de-identifier [1,2,6,8,20,21,23,27]. A use-case decision matrix is provided in Supplementary Table S1, sheet Practical_Framework.

The practical framework separates disclosure risk in the output from processing risk introduced by the tool. Local rule-based and encoder systems mainly create output-risk and maintenance-risk problems; cloud LLM workflows add upstream exposure because identifiable text may be transmitted, logged, retained, used for service improvement, or processed under a model version that cannot later be reconstructed. With external LLMs, privacy risk begins before the model has produced a de-identified text.

Operational risk assessment is better framed as linked questions than as a single score. What remains in the output? Missed PHI spans, rare contextual identifiers, unstable surrogates, or recognizable temporal patterns? Who could use those cues? An internal analyst, a linkage-capable institution, a hostile reader, or a public data user? What happened before the output existed? Local processing, cloud transmission, logging, retention, or vendor-side model use? What is the intended release context? Internal reuse, contractual sharing,

or public release? High entity-level F1 does not make a corpus low risk if contextual identifiers survive, if raw text left the institution before de-identification, or if the intended release context requires anonymity rather than pseudonymization.

5. Discussion

5.1. Clinical and Methodological Implications

Clinical text de-identification now includes controlled transformation of narrative evidence as well as named-entity recognition. When annotated data are available, fine-tuned transformers have the most consistent record for PHI detection; hybrid systems add deterministic handling of identifiers, dates, and templated signatures [1,2,8,17,20,21,23,27]. The 47 additional Scopus candidates added studies of transformation, pseudonymization, synthetic and surrogate text, multilingual adaptation, utility, and residual risk, but did not require another classification axis. Method choice still depends on the corpus, processing environment, transformation task, release context, and acceptable residual risk. The available studies justify defined LLM uses, not general production deployment.

The synthesis separates task function from implementation paradigm. A system may support detection, transformation, adaptation, assurance, or governance, and the same model family may appear in several roles. The two-axis classification and the practical selection framework in Supplementary Table S1, sheets Two_Axis_Classification and Practical_Framework, replace the broad question “rules, transformers, or LLMs?” with a more specific one: what combination of local processing, model-based detection, transformation, audit, and governance is justified for this corpus, sharing context, and risk threshold? [5,6,8,22,49,57]. Separating these axes prevents a high-performing detector from being treated as a complete privacy intervention. The framework is structured deployment guidance, not a definitive ranking of model families.

5.2. Why High F1 Is Not Enough

Benchmark F1 is a weak proxy for disclosure control. A system may achieve excellent micro-F1 while missing rare but high-risk identifiers, or remove so much context that downstream research value is impaired. Clinical de-identification requires assessment of both privacy risk and retained clinical value [5,7,8,22,49,59]. Generative models sharpen this problem because they may remove explicit PHI while preserving re-identifying context, or replace sensitive information with plausible but clinically misleading content [6,22–24]. A single F1 score can therefore make privacy protection look stronger than it is. Deployment-oriented studies are more informative when they report PHI-level recall, high-risk residual errors, strict and relaxed matching criteria, transformation quality, information loss, downstream task performance, and external validation across institution, document type, language, or time period [26,42,46,55].

Formal privacy metrics can help structure this assessment, but they do not transfer mechanically from tabular data to clinical narratives. In tabular release, k-anonymity, l-diversity, and related concepts operate over explicit quasi-identifier columns and equivalence classes. In free text, quasi-identifiers are dispersed across sentences, temporal sequences, occupations, locations, family relations, rare diseases, and narrative events. A de-identified note may satisfy a token-level PHI schema while still being unique in a local population. For clinical narratives, formal privacy reasoning therefore needs to be paired with residual-context review, linkage-scenario analysis, adversarial re-identification testing, and utility assessment; k-anonymity-style generalization may be useful for structured attributes or controlled replacements, but it cannot by itself certify narrative anonymity.

5.3. LLMs and Foundation Models

LLM performance depends on the task and processing environment. Under a stable PHI schema, fine-tuned encoders and hybrid systems are easier to validate, version, and keep under local control; the literature gives no general reason to replace them with prompted LLMs [1,2,8,17,26,42,56]. Generative models become relevant when the task involves rewriting, surrogate generation, date normalization, training-data augmentation, sparse annotation, or tests of residual identifiability [6,20,22,23,27]. Operational use still requires local validation and residual-risk assessment.

Direct prompting reduces the need for annotated training data, but it is not yet a dependable production architecture. DeID-GPT and related evaluations redact HIPAA-like identifiers without task-specific training [21,27]; recall and replacement behavior can still change with prompt wording, model version, context window, decoding settings, safety behavior, and vendor updates [22–24].

Evidence favors using LLMs alongside encoder models for specific subtasks. GPT-4-style augmentation can increase variation in training examples and may improve cross-dataset generalization while leaving a smaller local model as the deployable detector [20]. In-context learning, LoRA-style adaptation, and fine-tuned LLM workflows may help when the task includes temporal normalization or sensitive health information transformation rather than only PHI tagging [23]. Adversarial prompting can expose residual identifiability that token-level F1 misses [6–8]. Across the core corpus, 17 records had an LLM/generative title signal (15.5%; 95% CI 9.3–23.6), and broader record-level coding identified 23 LLM/generative records (20.9%; 95% CI 13.7–29.7). The Scopus screen surfaced further LLM-oriented candidates, including agentic anonymization, zero-shot commercial API evaluation, local LLM medical record de-identification, and LLM-supported synthetic or surrogate text generation. Direct comparative performance or benchmark evaluations accounted for only 9/110 records (8.2%; 95% CI 3.8–15.0), and deployment-relevant outcomes remained uncommon. The small comparative subset does not support broad model ranking.

For high-throughput detection of known PHI categories within a regulated institution, the evidence favors a local transformer-plus-rule pipeline with manual audit. LLMs have a stronger rationale where generative or inferential behavior is needed: low-resource adaptation, synthetic training data, semantic-preserving replacement, temporal normalization, or adversarial stress testing. If text leaves the institution for API-based processing, evaluation should include model version, prompt design, data pathway, privacy agreement, residual PHI audit, and clinical information-loss assessment.

5.4. Generalizability and Multilingual Evidence

Generalizability remains limited by the available corpora. Systems trained on familiar corpora may not transfer to local templates, abbreviations, entity distributions, or documentation styles [12,26,42,44,46,55]. In de-identification, transfer failure is not only a performance problem; it can become a privacy failure. The Scopus screen reinforces language transfer and local adaptation as deployment problems. German language clinical text illustrates the point: compounds, variable capitalization, inflected names, local institutional terms, and mixed administrative–medical phrasing affect both PHI detection and clinical utility. Work on GRASCCO, GeMTEx, and German clinical document de-identification supports language-specific resources rather than direct importation of English language models [43,46].

Script and tokenization are not secondary implementation details; they shape the de-identification task itself. Chinese, Japanese, Korean, Thai, Arabic, and mixed-script clinical narratives challenge assumptions inherited from English benchmark corpora. Word boundaries may not be marked by spaces, subword tokenizers may split names or institutions inconsistently, and a single character sequence can carry information that would be

distributed across several tokens in English. Mixed-script notes can combine local scripts, Latin abbreviations, English drug names, transliterated names, hospital codes, and copy-and-pasted administrative identifiers in the same sentence. These patterns affect both rules and neural models: regular expressions may miss local identifier formats, dictionary matching may fail under transliteration or inflection, and transformer token classification can shift boundary errors into PHI recall errors. For character-based or mixed-script languages, evaluation should therefore report the segmentation or tokenizer used, whether PHI spans were annotated at character, subword, token, or entity level, how partial matches were scored, and whether errors were inspected separately for names, locations, institutions, dates, and administrative identifiers [4,42,45,48].

5.5. Governance and Deployment Architecture

Technical metrics do not settle the governance question. Under GDPR, pseudonymized data may remain personal data, and anonymization depends on whether re-identification is reasonably possible [10]. This is especially relevant for free text, where context can identify a person even without names or dates [5–8]. Many NLP papers use de-identification, pseudonymization, and anonymization interchangeably, but PHI removal for a specific research workflow does not establish anonymous data in the legal sense [3,9,10,12].

The EU Artificial Intelligence Act strengthens this governance perspective when AI systems are used in or around clinical workflows. Even when a de-identification system is not itself a diagnostic tool, its outputs can shape downstream research, model training, data sharing, and patient privacy. Deployment documentation should specify intended use, data-processing pathway, model version, human oversight, logging, post-deployment monitoring, and residual-risk review in addition to conventional NLP performance metrics [11]. De-identification is not a preprocessing step before governance begins; it is itself a governed data processing activity.

Deployment evidence points to layered systems rather than a single universal model. In clinical use, a typical pipeline combines high-recall rules for structured identifiers, a fine-tuned transformer for context-dependent spans, post-processing for replacement consistency and date shifting, and an audit layer for residual PHI, information loss, and failure modes [1–4,51].

LLMs fit limited and auditable roles in this architecture: synthetic training examples, surrogate proposals, second-pass review of contextual identifiers, or adversarial re-identification testing. They also shift the risk boundary. A local encoder can fail by missing identifiers in the output; a cloud LLM creates upstream exposure because identifiable text is transmitted before de-identification succeeds. External LLM services require formal data-processing agreements, privacy assessment, institutional approval, access logging, retention limits, no-training/no-secondary-use guarantees, and model/version documentation before raw clinical text is processed. Local or private-cloud LLMs reduce transfer risk, but still require validation, residual PHI audit, information-loss assessment, throughput planning, model update control, and prompt/output retention rules [6,20,22–24,27].

Risk thresholds are use-case specific, not properties of a model. A dataset released externally needs stricter controls than text used internally under ethics approval and a data-use agreement. A corpus for clinical NLP development may need temporal and semantic utility preserved, whereas a public teaching dataset may prioritize stronger masking. Authors should report which identifiers were targeted, which contextual cues remained out of scope, whether the output is de-identified or anonymous, which tool processed the raw text, and whether the corpus may be used internally, shared under contract, or released publicly.

5.6. Operational Comparison of LLMs and Encoder Pipelines

Complete processing routes should be compared by how well they control exposure, preserve useful text, and contain failure. For repeated processing of similar hospital note types, a rule-enhanced encoder is easier to validate, audit, and keep under local data control when the PHI schema is known [26,42,46,55,56]. LLMs may add functions that this baseline does not address, including the generation of training variants, review of uncertain spans, surrogate proposals, temporal normalization, and red-team testing. For externally shared corpora, their use requires an empirical improvement in residual-risk or utility assessment, not merely a newer model [6,20,22,23,26].

5.7. Research Gaps and Reporting Needs

The field still lacks corpora that reflect the settings in which de-identification is most consequential. English discharge summaries dominate, whereas European language narratives, psychiatric notes, rare disease documentation, social histories, and non-alphabetic or mixed-script languages remain underrepresented. The Scopus screen suggests that some of this evidence sits outside the most visible PubMed/OpenAlex routes, especially in conference proceedings and regional venues. That finding changes the research agenda more than the present classification: future reviews should deliberately combine reproducible open searches with closed-database checks when the goal is to map multilingual and workshop-level method development. Contextual identifiers such as rare events, occupations, family relations, place references, and longitudinal event chains need more systematic evaluation [12,42–46,48]. Future studies should treat external validation, utility preservation, and residual-risk assessment as core outcomes, and should report document type, language, writing system, PHI schema, matching criteria, train-test split, annotation guidelines, tokenizer or segmentation algorithm, span-boundary convention, tokenization error analysis, and evaluation setting.

LLM studies require additional reporting: model version, prompt structure, deployment setting, data processing pathway, whether sensitive text left the institution, whether prompts or logs were retained, whether vendor terms permitted secondary use, and whether outputs were manually audited for residual PHI and clinical information loss. Direct prompting should be compared with fine-tuned local transformer baselines rather than with generic or no-system baselines alone. Because templates, institutional names, communication practices, and model behavior change over time, deployment requires monitoring and recalibration [21–24,26,27].

5.8. Strengths and Limitations of This Review

The supplementary file and evidence workbook report the search strategy, deduplication, eligibility decisions, 110-record core extraction, sensitivity checks, independent second screening, and reconciliation decisions. The review remains limited by the reporting and availability of published studies. Many clinical de-identification systems are developed in institutional settings but not fully reported because of privacy restrictions, while public datasets may not represent real-world documentation diversity. Preprints and conference papers are prominent in the LLM literature but may not have undergone the same level of peer review as journal articles.

Screening and charting used predefined eligibility criteria, explicit exclusion categories, a retained screening log, and independent second screening. Agreement was stronger for the retain-versus-exclude boundary than for finer category distinctions, which required reconciliation. Consistent with PRISMA-ScR transparency principles [28], the supplementary materials provide eligibility decisions, exclusion reasons, the complete 110-record extraction, endpoint and quantitative anchors, verification fields, and reconciliation materials.

References, screening categories, extraction fields, manuscript claims, and supplementary files were manually checked by the author.

The open-database strategy improves reproducibility because PubMed/MEDLINE, OpenAlex, ACL Anthology, arXiv, and targeted search components can be repeated from openly accessible sources. Web of Science and Scopus added breadth; abstract screening of all 47 additional Scopus core sensitivity candidates placed them within mapped functions, paradigms, languages, document types, or utility/risk themes. These candidates are reported separately from the 110-record open-database core denominator. The review does not provide pooled effect estimates. Claims about the superiority of one model family should be tied to corpus, language, note type, PHI schema, matching criteria, deployment setting, and model version.

6. Conclusions

Clinical text de-identification depends on more than model performance. Transformation fidelity, local adaptation, residual risk, clinical utility, and governance determine whether a system is fit for use. Local, auditable pipelines remain the soundest baseline for repeated PHI detection in regulated settings, with rules for structured identifiers, dates, and templated signatures. An LLM warrants inclusion only when a defined function yields a measurable benefit under controlled data exposure and residual-risk testing.

LLMs can assist with sparse-label adaptation, synthetic training examples, meaning-preserving surrogate generation, temporal normalization, second-pass contextual review, and adversarial testing. None of these uses removes the need for validated de-identification infrastructure or a separate review before external release. Processing identifiable text through an external model makes transmission, retention, logging, secondary use, and model drift part of the privacy risk. A defensible workflow records what was removed, where the raw text was processed, how clinical meaning was preserved, and how residual identifiability was tested.

Supplementary Materials: The following supporting information can be downloaded at <https://www.mdpi.com/article/10.3390/j9030022/s1>: Supplementary File S1: PRISMA ScR Checklist; Supplementary File S2: Search Strategy and Audit Log; Supplementary Table S1: Evidence Workbook.

Funding: This research was supported by the Swiss National Science Foundation under grant number 211751.

Institutional Review Board Statement: Not applicable. This study is a scoping review of published literature and does not involve human participants or identifiable patient data.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new patient-level data were generated. The Supplementary Material includes the full search strategy and audit log, an evidence workbook with screening logs, eligibility decisions, complete 110-record extraction, quantitative anchors, source/extraction-basis fields, quality/risk appraisal and rubric, Web of Science and Scopus sensitivity checks, reconciliation sheets, endpoint coverage, practical framework, representative methods, two-axis classification, title/abstract assessment of 35 additional records, abstract-level scan of 47 additional Scopus core sensitivity candidates, and the PRISMA-ScR checklist.

Acknowledgments: ChatGPT 5.5 (OpenAI, San Francisco, CA, USA) supported language editing during manuscript preparation. The review design, interpretation, reference verification, and final text remain the author's responsibility.

Conflicts of Interest: The author declares no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial intelligence
BERT	Bidirectional Encoder Representations from Transformers
CRF	Conditional random field
EHR	Electronic health record
GDPR	General Data Protection Regulation
HIPAA	Health Insurance Portability and Accountability Act
LLM	Large language model
NER	Named-entity recognition
NLP	Natural language processing
PHI	Protected health information
PII	Personally identifiable information
PRISMA-ScR	PRISMA extension for Scoping Reviews

References

1. Norgeot, B.; Muenzen, K.; Peterson, T.A.; Fan, X.; Glucksberg, B.S.; Schenk, G.; Rutenberg, E.; Oskotsky, B.; Sirota, M.; Yazdany, J.; et al. Protected Health Information filter (Philter): Accurately and securely de-identifying free-text clinical notes. *npj Digit. Med.* **2020**, *3*, 57. [CrossRef] [PubMed]
2. Liu, J.; Gupta, S.; Chen, A.; Wang, C.K.; Mishra, P.; Dai, H.J.; Wong, Z.S.Y.; Jonnagaddala, J. OpenDeID Pipeline for Unstructured Electronic Health Record Text Notes Based on Rules and Transformers: Deidentification Algorithm Development and Validation Study. *J. Med. Internet Res.* **2023**, *25*, e48145. [CrossRef] [PubMed]
3. Tannier, X.; Wajsburt, P.; Calliger, A.; Dura, B.; Mouchet, A.; Hilka, M.; Bey, R. Development and Validation of a Natural Language Processing Algorithm to Pseudonymize Documents in the Context of a Clinical Data Warehouse. *Methods Inf. Med.* **2024**, *63*, 21–34. [CrossRef] [PubMed]
4. Wang, C.K.; Wang, F.D.; Lee, Y.Q.; Chen, P.T.; Wang, B.H.; Su, C.H.; Kuo, J.C.C.; Wu, C.S.; Chien, Y.L.; Dai, H.J.; et al. Principle-Based Approach for the De-Identification of Code-Mixed Electronic Health Records. *IEEE Access* **2022**, *10*, 22875–22885. [CrossRef]
5. Ben Cheikh Larbi, I.; Burchardt, A.; Roller, R. Clinical Text Anonymization, its Influence on Downstream NLP Tasks and the Risk of Re-Identification. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 105–111. [CrossRef]
6. Morris, J.X.; Champion, T.R.; Nutheti, S.L.; Peng, Y.; Raj, A.; Zabih, R.; Cole, C.L. DIRI: Adversarial Patient Reidentification with Large Language Models for Evaluating Clinical Text Anonymization. In *Proceedings of the AMIA Joint Summits on Translational Science Proceedings*; American Medical Informatics Association: Washington, DC, USA, 2025; pp. 355–364.
7. Carrell, D.S.; Malin, B.A.; Cronkite, D.J.; Aberdeen, J.S.; Clark, C.; Li, M.R.; Bastakoty, D.; Nyemba, S.; Hirschman, L. Resilience of clinical text de-identified with “hiding in plain sight” to hostile reidentification attacks by human readers. *J. Am. Med. Inform. Assoc.* **2020**, *27*, 1374–1382. [CrossRef] [PubMed]
8. Radhakrishnan, L.; Schenk, G.; Muenzen, K.; Oskotsky, B.; Ashouri Choshali, H.; Plunkett, T.; Israni, S.; Butte, A.J. A certified de-identification system for all clinical text documents for information extraction at scale. *JAMIA Open* **2023**, *6*, ooad045. [CrossRef] [PubMed]
9. Guidance Regarding Methods for De-Identification of Protected Health Information in Accordance with the HIPAA Privacy Rule, 2026. U.S. Department of Health and Human Services, Office for Civil Rights. Available online: <https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html> (accessed on 4 May 2026).
10. Regulation (EU) 2016/679, Recital 26: Not Applicable to Anonymous Data, 2026. European Parliament and Council of the European Union. Available online: <https://eur-lex.europa.eu/eli/reg/2016/679/oj> (accessed on 4 May 2026).
11. Regulation (EU) 2024/1689 of the European Parliament and of the Council: Artificial Intelligence Act, 2024. European Parliament and Council of the European Union. Available online: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj> (accessed on 5 June 2026).
12. Kovacevic, A.; Basaragin, B.; Milosevic, N.; Nenadic, G. De-identification of clinical free text using natural language processing: A systematic review of current approaches. *Artif. Intell. Med.* **2024**, *151*, 102845. [CrossRef] [PubMed]
13. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the NAACL-HLT*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 4171–4186. [CrossRef]
14. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692. [CrossRef]

15. Tang, B.; Jiang, D.; Chen, Q.; Wang, X.; Yan, J.; Shen, Y. De-identification of Clinical Text via Bi-LSTM-CRF with Neural Language Models. In *Proceedings of the AMIA Annual Symposium Proceedings*; American Medical Informatics Association: Washington, DC, USA, 2019; pp. 857–863.
16. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2019**, *36*, 1234–1240. [[CrossRef](#)] [[PubMed](#)]
17. Johnson, A.E.W.; Bulgarelli, L.; Pollard, T.J. Deidentification of free-text medical records using pre-trained bidirectional transformers. In *Proceedings of the ACM Conference on Health, Inference, and Learning*; ACM CHIL '20; ACM: New York, NY, USA, 2020; pp. 214–221. [[CrossRef](#)] [[PubMed](#)]
18. Meaney, C.; Hakimpour, W.; Kalia, S.; Moineddin, R. A Comparative Evaluation of Transformer Models for De-Identification of Clinical Text Data. *arXiv* **2022**, arXiv:2204.07056. [[CrossRef](#)]
19. Zhao, Y.S.; Zhang, K.L.; Ma, H.C.; Li, K. Leveraging text skeleton for de-identification of electronic medical records. *BMC Med. Inform. Decis. Mak.* **2018**, *18*, 18. [[CrossRef](#)] [[PubMed](#)]
20. Kim, W.; Hahm, S.; Lee, J. Generalizing Clinical De-identification Models by Privacy-safe Data Augmentation using GPT-4. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 21204–21218. [[CrossRef](#)]
21. Altalla, B.; Abdalla, S.; Altamimi, A.; Bitar, L.; Al Omari, A.; Kardan, R.; Sultan, I. Evaluating GPT Models for Clinical Note De-identification. *Sci. Rep.* **2025**, *15*, 3852. [[CrossRef](#)] [[PubMed](#)]
22. Aghakasiri, K.; Zambare, N.; Thai, J.; Ye, C.; Mehta, M.; Mitchell, J.R.; Abdalla, M. Not What the Doctor Ordered: Surveying LLM-based De-identification and Quantifying Clinical Information Loss. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 32175–32191. [[CrossRef](#)]
23. Dai, H.J.; Mir, T.H.; Chen, C.T.; Chen, C.C.; Yang, H.P.; Lee, C.H.; Chou, Y.Y.; Teng, Y.C.; Gupta, S.; Panchal, O.; et al. Leveraging large language models for the deidentification and temporal normalization of sensitive health information in electronic health records. *npj Digit. Med.* **2025**, *8*, 517. [[CrossRef](#)] [[PubMed](#)]
24. Chim, J.; Pearson, J.; Schofield, D.; Liakata, M. Evaluating privacy leakages in LLM-driven ambient clinical documentation. *Front. Digit. Health* **2026**, *8*, 1761624. [[CrossRef](#)] [[PubMed](#)]
25. Jonnagaddala, J.; Chen, A.; Batongbacal, S.; Nekkanti, C. The OpenDeID corpus for patient de-identification. *Sci. Rep.* **2021**, *11*, 19973. [[CrossRef](#)] [[PubMed](#)]
26. Kuo, R.; Soltan, A.A.; O'Hanlon, C.; Hasanica, A.; Clifton, D.A.; Collins, G.; Furniss, D.; Eyre, D.W. Benchmarking transformer-based models for medical record de-identification in a single center multi-specialty evaluation. *iScience* **2025**, *28*, 113732. [[CrossRef](#)] [[PubMed](#)]
27. Liu, Z.; Huang, Y.; Yu, X.; Zhang, L.; Wu, Z.; Cao, C.; Dai, H.; Zhao, L.; Li, Y.; Shu, P.; et al. DeID-GPT: Zero-shot Medical Text De-Identification by GPT-4. *arXiv* **2023**, arXiv:2303.11032. [[CrossRef](#)]
28. Tricco, A.C.; Lillie, E.; Zarin, W.; O'Brien, K.K.; Colquhoun, H.; Levac, D.; Moher, D.; Peters, M.D.; Horsley, T.; Weeks, L.; et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Ann. Intern. Med.* **2018**, *169*, 467–473. [[CrossRef](#)] [[PubMed](#)]
29. Arksey, H.; O'Malley, L. Scoping studies: Towards a methodological framework. *Int. J. Soc. Res. Methodol.* **2005**, *8*, 19–32. [[CrossRef](#)]
30. Levac, D.; Colquhoun, H.; O'Brien, K.K. Scoping studies: Advancing the methodology. *Implement. Sci.* **2010**, *5*, 69. [[CrossRef](#)] [[PubMed](#)]
31. Munn, Z.; Peters, M.D.J.; Stern, C.; Tufanaru, C.; McArthur, A.; Aromataris, E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Med. Res. Methodol.* **2018**, *18*, 143. [[CrossRef](#)] [[PubMed](#)]
32. *Scopus Content Coverage Guide*; Elsevier: Amsterdam, The Netherlands, 2023. Available online: https://supportcontent.elsevier.com/Support%20Hub/Scopus/Scopus_ContentCoverage_Guide_WEB.pdf (accessed on 5 June 2026).
33. *Web of Science Core Collection*; Clarivate: London, UK, 2026. Available online: <https://clarivate.com/academia-government/scientific-and-academic-research/research-discovery-and-referencing/web-of-science/web-of-science-core-collection/> (accessed on 5 June 2026).
34. *How Does OpenAlex Work?*; OpenAlex Help Center: Vancouver, BC, Canada, 2025. Available online: <https://help.openalex.org/hc/en-us/articles/28932712154391-How-does-OpenAlex-work> (accessed on 5 June 2026).
35. Mongeon, P.; Paul-Hus, A. The journal coverage of Web of Science and Scopus: A comparative analysis. *Scientometrics* **2016**, *106*, 213–228. [[CrossRef](#)]
36. Bramer, W.M.; Rethlefsen, M.L.; Kleijnen, J.; Franco, O.H. Optimal database combinations for literature searches in systematic reviews: A prospective exploratory study. *Syst. Rev.* **2017**, *6*, 245. [[CrossRef](#)] [[PubMed](#)]

37. Martín-Martín, A.; Orduna-Malea, E.; Thelwall, M.; Delgado López-Cózar, E. Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *J. Inf.* **2018**, *12*, 1160–1177. [\[CrossRef\]](#)
38. Visser, M.; van Eck, N.J.; Waltman, L. Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. *Quant. Sci. Stud.* **2021**, *2*, 20–41. [\[CrossRef\]](#)
39. Miake-Lye, I.M.; Hempel, S.; Shanman, R.; Shekelle, P.G. What is an evidence map? A systematic review of published evidence maps and their definitions, methods, and products. *Syst. Rev.* **2016**, *5*, 28. [\[CrossRef\]](#) [\[PubMed\]](#)
40. Snilstveit, B.; Vojtkova, M.; Bhavsar, A.; Stevenson, J.; Gaarder, M. Evidence and gap maps: A tool for promoting evidence informed policy and strategic research agendas. *J. Clin. Epidemiol.* **2016**, *79*, 120–129. [\[CrossRef\]](#) [\[PubMed\]](#)
41. Trienes, J.; Trieschnigg, D.; Seifert, C.; Hiemstra, D. Comparing Rule-based, Feature-based and Deep Neural Methods for De-identification of Dutch Medical Records. *arXiv* **2020**, arXiv:2001.05714. [\[CrossRef\]](#)
42. Chen, F.; Bokhari, S.M.A.; Cato, K.; Gursoy, G.; Rossetti, S. Examining the Generalizability of Pretrained De-identification Transformer Models on Narrative Nursing Notes. *Appl. Clin. Inform.* **2024**, *15*, 357–367. [\[CrossRef\]](#) [\[PubMed\]](#)
43. Lohr, C.; Matthies, F.; Faller, J.; Modersohn, L.; Riedel, A.; Hahn, U.; Kiser, R.; Boeker, M.; Meineke, F. De-Identifying GRASCCO: A Pilot Study for the De-Identification of the German Medical Text Project (GeMTEx) Corpus. In *German Medical Data Sciences 2024; Studies in Health Technology and Informatics*; IOS Press: Amsterdam, The Netherlands, 2024; Volume 317, pp. 171–179. [\[CrossRef\]](#) [\[PubMed\]](#)
44. Azzouzi, M.E.; Coatrieux, G.; Bellafqira, R.; Delamarre, D.; Riou, C.; Oubenali, N.; Cabon, S.; Cuggia, M.; Bouzille, G. Automatic de-identification of French electronic health records: A cost-effective approach exploiting distant supervision and deep learning models. *BMC Med. Inform. Decis. Mak.* **2024**, *24*, 54. [\[CrossRef\]](#) [\[PubMed\]](#)
45. Lee, Y.Q.; Chen, C.T.; Chen, C.C.; Lee, C.H.; Chen, P.; Wu, C.S.; Dai, H.J. Unlocking the Secrets Behind Advanced Artificial Intelligence Language Models in Deidentifying Chinese-English Mixed Clinical Text: Development and Validation Study. *J. Med. Internet Res.* **2024**, *26*, e48443. [\[CrossRef\]](#) [\[PubMed\]](#)
46. Arzideh, K.; Baldini, G.; Winnekens, P.; Friedrich, C.M.; Nensa, F.; Idrissi-Yaghir, A.; Hosch, R. A Transformer-Based Pipeline for German Clinical Document De-Identification. *Appl. Clin. Inform.* **2025**, *16*, 31–43. [\[CrossRef\]](#)
47. Schiezar, M.; Rosa, G.; Campos, B.A.G.; Pedrini, H. Guardians of the data: NER and LLMs for effective medical record anonymization in Brazilian Portuguese. *Front. Public Health* **2026**, *13*, 1717303. [\[CrossRef\]](#) [\[PubMed\]](#)
48. Chang, J.; Park, J.; Kim, C.; Park, R.W. A De-Identification Model for Korean Clinical Notes: Using Deep Learning Models. In *Proceedings of the MEDINFO 2023—The Future Is Accessible*; IOS Press: Amsterdam, The Netherlands, 2024. [\[CrossRef\]](#) [\[PubMed\]](#)
49. Vakili, T.; Dalianis, H. Utility Preservation of Clinical Text After De-Identification. In *Proceedings of the 21st Workshop on Biomedical Language Processing*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 383–388. [\[CrossRef\]](#)
50. Meystre, S.; Heider, P. High Accuracy Open-Source Clinical Data De-Identification: The CliniDeID Solution. In *MEDINFO 2023—The Future Is Accessible*; Studies in Health Technology and Informatics; IOS Press: Amsterdam, The Netherlands, 2024; Volume 310, pp. 1375–1376. [\[CrossRef\]](#) [\[PubMed\]](#)
51. Sundrelingam, V.; Parimoo, S.; Pogacar, F.; Koppula, R.; Shin, S.; Pou-Prom, C.; Roberts, S.B.; Verma, A.A.; Razak, F. pyDeid: An improved, fast, flexible, and generalizable rule-based approach for deidentification of free-text medical records. *JAMIA Open* **2024**, *8*, ooae152. [\[CrossRef\]](#) [\[PubMed\]](#)
52. Beltagy, I.; Lo, K.; Cohan, A. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 3613–3618. [\[CrossRef\]](#)
53. Alsentzer, E.; Murphy, J.R.; Boag, W.; Weng, W.H.; Jin, D.; Naumann, T.; McDermott, M. Publicly Available Clinical BERT Embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 72–78. [\[CrossRef\]](#)
54. Moore, C.; Bulgarelli, L.; Pollard, T.; Johnson, A. Transformer-DeID: Deidentification of free-text clinical notes with transformers. *PhysioNet* **2023**, RRID:SCR_007345. Available online: <https://physionet.org/content/transformer-deid/1.0.0/> (accessed on 5 June 2026).
55. Yang, X.; Lyu, T.; Li, Q.; Lee, C.Y.; Bian, J.; Hogan, W.R.; Wu, Y. A study of deep learning methods for de-identification of clinical notes in cross-institute settings. *BMC Med. Inform. Decis. Mak.* **2019**, *19*, 232. [\[CrossRef\]](#) [\[PubMed\]](#)
56. An, J.; Kim, J.; Sunwoo, L.; Baek, H.; Yoo, S.; Lee, S. De-identification of clinical notes with pseudo-labeling using regular expression rules and pre-trained BERT. *BMC Med. Inform. Decis. Mak.* **2025**, *25*, 82. [\[CrossRef\]](#) [\[PubMed\]](#)
57. Sarkar, A.R.; Chuang, Y.S.; Mohammed, N.; Jiang, X. De-identification is not enough: A comparison between de-identified and synthetic clinical notes. *Sci. Rep.* **2024**, *14*, 29669. [\[CrossRef\]](#) [\[PubMed\]](#)

58. Leevy, J.L.; Khoshgoftaar, T.M.; Villanustre, F. Survey on RNN and CRF models for de-identification of medical free text. *J. Big Data* **2020**, *7*, 73. [[CrossRef](#)]
59. Xiao, Y.; Lim, S.; Pollard, T.J.; Ghassemi, M. In the Name of Fairness: Assessing the Bias in Clinical Record De-identification. In *Proceedings of the 2023 ACM Conference on Fairness Accountability and Transparency; FAccT '23*; ACM: New York, NY, USA, 2023; pp. 123–137. [[CrossRef](#)]

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