De-Identifying German Clinical Notes Under Small-Corpus Constraints

Transferring State-of-the-Art Approaches from English Benchmarks

Anna-Lena Artmann
Data Science and Artificial Intelligence
Center for Advanced Studies of the Baden-Württemberg Cooperative State University
Heilbronn, Germany
e-mail: cas417266@cas.dhbw.de

Abstract-**Most de-identification methods are trained on English corpora, limiting cross-lingual transfer to German where annotated data are scarce. We evaluate transferability of state-of-the-art approaches via a systematic review (Jan 2023 - Apr 2025; 55 publications) and a controlled experiment on Graz Synthetic Clinical text Corpus. We compare BiLSTM-CRF, gELECTRA, and an 8B LLaMA-3 variant (SauerkrautLM), reporting entity-level precision, recall, and F1 score. The review shows a continued shift to transformers and hybrids but limited cross-language comparability due to heterogeneous datasets and metrics. On GraSCCo, BiLSTM-CRF is a reliable baseline (entity-level Micro-F1 = 0.96, Macro-F1 = 0.95). gELECTRA performs well on structured identifiers but drops on rare or variable categories (micro-F1 0.66) due to data/label sparsity. One-shot decoding with Llama-3.1-SauerkrautLM was unreliable under on-premises constraints. For German small-corpus settings, compact encoders are the most pragmatic near-term solution, while larger corpora will be necessary to realize full potential of transformer encoders like gELECTRA.**

Keywords-De-identification; GSA region; gELECTRA; BiLSTM; GraSCCo.

I. INTRODUCTION

Clinical narratives are central to data-driven medical research, as they capture rich contextual information that structured health data cannot provide [1]. To enable secondary use, however, all Personally Identifiable Information (PII) must first be removed [2]. This process, known as de-identification, poses particular challenges for free-text records, which are less predictable than structured fields [3].

Over the past decade, de-identification research has evolved considerably. Rule-based and hybrid systems remain relevant for well-structured entities, yet recent surveys (e.g., [1]) document a clear methodological shift toward data-driven approaches. Especially since the introduction of transformer architectures, such as BERT (Bidirectional Encoder Representations from Transformers) [4], deep learning methods have consistently achieved state-of-the-art (SoTA) performance, with F1-scores surpassing 95% on English benchmarks like i2b2 (Informatics for Integrating Biology and the Bedside) [5] and MIMIC-III (Medical Information Mart for Intensive Care) [1][6]. These corpora and results have effectively set the reference standard for de-identification. While these advances are impressive, they reflect an English-centric research landscape.

The dominance of English corpora means that most methods are trained and validated in data-rich settings, while other languages remain underexplored [7]. For German, the limited availability of annotated corpora presents a critical bottleneck. To our knowledge, as of June 2025, the only publicly available resource suitable for training and evaluation is the Graz Synthetic Clinical text Corpus (GraSCCo), which comprises 64 synthetic examples [8]. This scarcity not only limits model development but also hinders meaningful benchmarking and comparison [9].

Moreover, linguistic and institutional differences reduce the straightforward transferability of English-trained methods [7][9]. High performance on English datasets does not necessarily generalize to German clinical contexts [10]. Consequently, developing effective methods for German de-identification requires both methodological adaptation and careful evaluation under data-scarce conditions.

This study addresses this gap by systematically analyzing SoTA approaches for clinical de-identification and evaluating their transferability to German data. To this end, promising methods are identified in international literature and applied to a small example dataset (GraSCCo). The study thereby explores the opportunities and limitations of modern approaches in a realistic data-limited scenario reflective of the current situation in clinical practice across the German-speaking GSA region (Germany, Austria, Switzerland). The guiding research question is to what extent state-of-the-art methods can be adapted to support de-identification of German clinical free text under small-corpus constraints.

This paper is structured as follows. Section 2 describes the methodology, combining a systematic literature review with a controlled laboratory experiment. Section 3 presents the results of both the literature review and the empirical evaluation on a small German clinical corpus. Section 4 discusses the findings, focusing on methodological reliability, comparability, and cross-lingual transferability. Finally, Section 5 concludes the paper and outlines implications for practice as well as directions for future research.

II. METHODS

This study combines a systematic literature review with a laboratory experiment to assess the transferability of SoTA de-identification methods from English benchmarks to German data-scarce settings.

A. Literature Review

As a starting point, insights into the current research landscape were obtained by analyzing the two major de-identification challenges [5] and [11] and reviewing the recent comparative studies [1] and [2]. Building on these, a systematic literature review of international publications from January 2023 to April 2025 was conducted to capture the latest developments in clinical text de-identification.

The review was guided by three central questions: (1) which methodological approaches have achieved the most reliable performance, (2) what challenges limit comparability across studies, and (3) to what extent can findings from English benchmarks be transferred to the German clinical context?

The review followed Cooper's [12] taxonomy across six dimensions: focus on both methods and outcomes, critical integration of findings, conceptual and methodological organization, neutral perspective, selective yet representative coverage, and orientation toward a scientific audience in natural language processing. Search was conducted primarily in IEEE Xplore and PubMed, supplemented by Google Scholar. Eligible studies focused on the de-identification of clinical free text, reported transparent methods and results, and included performance metrics, such as F1-scores (or, for generative models, accuracy).

In total, 55 publications were reviewed, including more than 30 based on English data as well as the ten top-ranked methods summarized by [1]. For an in-depth analysis, the 25 best-performing studies (measured by F1-score) were compared in detail. Generative language models, which often report accuracy instead of F1-scores, were analyzed separately to account for methodological differences.

To capture the specific developments in the German research landscape, studies on German-language de-identification were treated separately. In this case, all available publications up to May 2025 were included, regardless of their publication year or reported performance. Cooper's taxonomy was therefore applied in a slightly adapted form, complemented by a chronological organization to trace the methodological development of German approaches alongside international trends.

The insights from this review provided the basis for selecting candidate methods for the subsequent laboratory experiment, which evaluates their applicability under data-scarce conditions.

B. Laboratory work

To complement the literature review and evaluate the most promising approaches under realistic data-poor conditions, we conducted a controlled experiment. Controlled experimentation allows for a systematic comparison of model architectures under constant conditions [13].

The independent variable was model architecture; the following systems were compared:

- BiLSTM-CRF (Bidirectional Long Short-term Memory with Conditional Random Field): A bidirectional recurrent neural network with a conditional random field output layer, widely used for sequence labeling in data-limited natural language processing tasks [14][15].
- gELECTRA (German Efficiently Learning an Encoder that Classifies Token Replacements): A German-adapted ELECTRA transformer, fine-tuned for token-level de-identification [16][17].
- LlaMA-3.1-SauerkrautLM-8B-Instruct (Large Language Model Meta AI): An instruction-tuned large language model based on Llama-3.1-8B-Instruct [18], fine-tuned on German-English data [19].

BiLSTM-CRF and gELECTRA were compared under identical conditions (dataset splits, runtime, and sentence-aware chunking into 256-token windows with a 64-token stride, matching gELECTRA's context). LLaMA-3.1 was evaluated separately in one-shot in-context mode on the same test set and metrics. Performance was measured by precision, recall, and F1-score (entity-level, micro- and macro-averaged) on GraSCCo (64 synthetic German clinical notes annotated in accordance with the HIPAA/Safe Harbor identifiers [20]), split 80/20 into training and test sets. BiLSTM-CRF (SpaCy-based [21]) and gELECTRA were trained in Google Colab (T4; 20 epochs; weight_decay 0.01), while LLaMA-3.1-SauerkrautLM was run locally via Ollama [22] to simulate data-sensitive deployment without cloud access. The inclusion of LLaMA-3.1-SauerkrautLM was exploratory, intended to highlight the limitations of large language models (LLM) under extreme data scarcity.

The working hypothesis was that with very small training sets (<100 texts), recurrent architectures, such as BiLSTM yield more stable de-identification performance than large pretrained transformer models or LLM-based one-shot approaches. The experiment thus examines whether compact recurrent or transformer models are more practical for de-identification in GSA-region hospitals, where data and infrastructure are limited.

III. RESULTS

In the following, the results of the literature review and the experimental evaluation on GraSCCo are presented.

A. Literature Review

1) Overall trend

Building on prior comparative reviews [1] and [2] and the shared tasks [5] and [11], our review of publications from January 2023 to April 2025 confirms the continued shift towards transformer-based approaches. Encoder-only transformer families (BERT, RoBERTa, DeBERTa, XLM-R, CamemBERT) dominate recent work (e.g., [37][42][46]). In many publications, the strongest systems appear to be hybrid pipelines that combine transformer encoders with rule-based modules, which reach top F1-scores above 0.98 (e.g.,

[37][42]). In contrast, purely rule-based systems perform substantially worse on modern benchmarks (e.g., MITDeID around F1 ≈ 0.64 [23]).

2) Model performance and state of the art

Specialized transformer variants (e.g., [37][42]) and strong BiLSTM-CRF baselines (e.g., [39][40]) both achieve competitive performance with F1-scores ≥ 0.99, indicating that recurrent architectures can still be viable. The top 25 systems by F1 are summarized in Table I, with hybrid and compact encoder models prevalent among the leaders.

TABLE I. TOP 25 STUDIES BY F1 SCORE (JAN 2023 – MAY 2025)

model	lang	data	cat	eval	F1
PubMedBERT [37]	EN	999 texts, i2b2 2014	21	T, B	0.995
Encoder with self-attention [38]	EN	MIMIC-III	18*	T, B	0.994
LSTM-CRF [39]	EN	MIMIC-III	18*	T, B	0.993
BiLSTM-CRF [40]	EN	600 texts, i2b2 2014	5	T, Mi	0.992
BiLSTM-CRF [41]	EN	MIMIC-III	21**	T, B	0.991
mDeBERTaV3 [42]	ES	MEDDOCAN [43]	8	E, Mi	0.990
BiLSTM-CRF [44]	EN	i2b2 2014	18*	T, B	0.990
Transformer-Ensemble [45]	EN	15'716 texts i2b2 2014	21	E, Mi	0.989 0.970
BERT-Large [46]	EN	i2b2 2014	21**	T, B	0.988
BiLSTM [47]	HU	15000 texts	10	T, B	0.987
Seq2Seq [48]	EN	i2b2 2014	18*	T, B	0.985
BERT-Ensemble [49]	EN	i2b2 2014	18	E, Mi	0.985
BiLSTM-CRF [50]	EN	i2b2 2014	5	T, B	0.983
GRU [51]	EN	i2b2 2014	21**	T, Mi	0.981
LSTM [52]	EN	i2b2 2014	7	T, B	0.980
CamelBERT [53]	AR	i2b2 2014	17	T, B	0.980
PI-RoBERTa [54]	EN	i2b2 2014 (translated)	8	T, Mi	0.980
XLM-RoBERTa large [55]	ES	MEDDOCAN	9	T, Mi	0.976
BERT(med)-BiLSTM-CRF [56]	ZH	33'107 texts	21**	E, Mi	0.976
KoBERT [57]	KO	11'281	6	T, B	0.971
BiLSTM-CRF [58]	FR	878'217	8	T, Mi	0.970
ClinicalBERT [59]	EN	i2b2 2014	18*	E, Mi	0.967
Mistral-7b [27]	FR	9'097 texts	6	T, B	0.967
BiLSTM-CRF [60]	EN	i2b2, CEGS N-GRID	21	E, Mi	0.965
Bert-base-german-cased [30]	DE	i2b2 2014 (translated)	21**	E, Ma	0.960

*18 HIPAA categories; ** i2b2 extended HIPAA categories; lang: language; cat: number of categories; eval: evaluation level (T: Token, E: Entity, B: Binary, Mi: Micro, Ma: Macro); EN: English; ES: Spanish; HU: Hungarian; AR: Arabic; ZH: Chinese; KO: Korean; FR: French; DE: German.

3) Generative LLMs

Decoder-based LLMs are a fast-moving research area, but remain limited in reliability, precision, and clinical usability. Zero-shot named entity recognition is especially challenging. LLaMA-3-8B, for instance, can reach very high recall (≈ 0.99) while yielding extremely low precision (entity-level F1 < 0.20), reflecting hallucinations and prompt sensitivity [24]. GPT-4 can occasionally match the performance of fine-tuned BERT-class models on small datasets, but still struggles with complex identifiers (e.g., hospital names), risking semantic information loss [25]. Prompting and fine-tuning help, yet privacy and compliance concerns constrain the real-world use of cloud-hosted GPT (Generative Pre-Trained Transformer) models [26]. Open-source LLMs like Mistral-7B show promise for local deployment when fine-tuned and quantized, though coverage across Protected Health Information (PHI) categories is often limited [27]. Overall, LLMs yield interesting experimental results, especially in few-shot settings, but remain less robust and transparent than compact encoder models, such as BERT.

4) Datasets, Evaluation and Transferability

Most studies rely on the i2b2-2014 benchmark, often combined with MIMIC-III or institutional hospital data. While this allows a certain degree of comparability, these benchmarks are repeatedly criticized for limited domain coverage and weak generalizability beyond the English-speaking context [28]. Evaluations further differ in whether they are token- or entity-based and in the use of micro- vs. macro-F1, hindering cross-study comparability [1][2]. This heterogeneity also weakens cross-language transfer claims and motivates German-specific validation.

Importantly for German, promising results were reported by Arzideh et al. [29] with gELECTRA and by Gunay et al. [30] using Bert-base-german-cased on machine-translated i2b2 2014 texts, both exceeding F1 > 0.95. These studies demonstrate that compact transformer models can achieve competitive performance even with limited resources, making them particularly relevant for the German-speaking clinical domain, where annotated corpora are small and access to real data remains highly restricted.

B. Research Status for German De-Identification

Research on German-language clinical text de-identification has evolved from early rule-based and hybrid approaches to modern transformer-based and LLM-driven methods. Early work included the Averbis system [31][32], which combined metadata, rule-based tagging, and machine learning and reported near-perfect performance in specific settings. Subsequent studies explored regex-based methods [33] and sequence models, such as CRF and BiLSTM, with BiLSTM reaching up to 96% F-scores [34].

More recent efforts integrated hybrid pipelines, such as Maskeeter [10], combining dictionaries, regex, and manual checks, while Gunay et al. [30] demonstrated strong performance (F1 ≈ 0.96) using German BERT model trained on synthetic and translated corpora.

Arzideh et al. [29] benchmarked multiple transformer models (mBERT, medBERTde, gBERT, gELECTRA, XLM-RoBERTa) on over 10'000 clinical documents, with ensemble strategies and gELECTRA achieving F1-scores up to 0.95, surpassing human annotators. At the same time, Sousa et al.

[35] tested zero- and one-shot prompting with LLMs (GPT-3.5, GPT-4, LLaMA), which showed promising recall but still lagged behind classical architectures in precision and robustness. Finally, Wiest et al. [36] proposed a locally deployable, privacy-preserving pipeline based on quantized LLMs, achieving over 99% sensitivity and 98% specificity while offering end-to-end de-identification workflows, but PHI coverage breadth and reliability remain open issues.

TABLE II. BILSTM-CRF AND GELECTRA PERFORMANCE ON TEST-SET-OBSERVED ENTITY TYPES

Entity	Precision		Recall		F1 Score		S*
Age	1.00	1.00	0.67	1.00	0.80	1.00	3
Date	0.97	0.98	0.95	0.98	0.96	0.98	316
ID	1.00	0.78	0.82	0.78	0.90	0.78	11
City	1.00	0.89	0.92	0.89	0.96	0.89	12
Hospital	1.00	0.57	1.00	0.67	1.00	0.62	19
Organization	1.00	0.00	1.00	0.00	1.00	0.00	2
Street name	0.80	1.00	1.00	1.00	0.89	1.00	12
Postal code	1.00	1.00	1.00	1.00	1.00	1.00	5
Clinician name	1.00	0.67	0.94	0.63	0.97	0.65	68
Patient name	1.00	0.89	0.87	0.94	0.93	0.91	55
Title	1.00	0.78	1.00	0.85	1.00	0.81	72
Profession	1.00	0.00	1.00	0.00	1.00	0.00	1
Micro avg	0.98	0.89	0.94	0.91	0.96	0.90	
Macro avg	0.98	0.66	0.93	0.67	0.95	0.66	

*S: Support

C. Laboratory results on small German corpus

The experimental evaluation on the GraSCCo corpus (see Table II) revealed clear performance differences under data-scarce conditions.

- BiLSTM-CRF: Achieved the best overall entity-level performance with Micro F1 = 0.96 (Micro P = 0.98, Micro R = 0.94; Macro F1 = 0.95). Except for the entities age (F1 = 0.80) and street names (F1 = 0.89), all categories reached Micro F1 scores between 0.90 and 1.00, indicating strong robustness, including on rare entities.
- gELECTRA: Delivered Micro F1 = 0.90 (Micro P = 0.89, Micro R = 0.91), but a substantially lower Macro F1 = 0.66, reflecting weaknesses on infrequent classes. Performance was near-perfect on structured entities (e.g., date: 0.98; postal code: 1.00), yet lower on semantically variable or sparse categories (e.g., clinician name: 0.65, hospital: 0.62) and missed organization (support = 2) and profession (support = 1).
- LlaMA-3.1-SauerkrautLM: The decoder model proved impractical in a one-shot setting. Although the output format was consistent, label assignment was unreliable, with frequent hallucinations, misclassifications, and omissions.

IV. DISCUSSION

On the small synthetic GraSCCo corpus (n = 64) [8][21], BiLSTM-CRF emerges as a reliable baseline (Micro-F1 = 0.96; Macro-F1 = 0.95), consistent with prior evidence for recurrent models in data-scarce sequence tagging [15][16]. gELECTRA performs strongly on structured identifiers, while its Macro-F1 drop (Micro-F1: 0.90; Macro-F1: 0.66) for rare or semantically variable categories might be influenced by data scarcity and label sparsity and should not be taken as definitive evidence of an inherent encoder limitation [1][29][61]. Based on the results in Table I, we expect encoder-based transformers, such as gELECTRA to achieve substantially higher performance and greater stability on a larger, more realistic dataset. Given compliance and infrastructure constraints, BiLSTM-CRF provides a prudent baseline; hybrid systems that combine encoders with deterministic rules merit further exploration [1][2][10].

Recent work exhibits heterogeneous protocols (ranging from token- to entity-level metrics, and micro- vs. macro-averaging) limiting comparability and possibly exaggerating perceived performance gains [1][2]. Our own experiment, though standardized, faces similar constraints due to its small synthetic corpus (GraSCCo, n = 64). In addition, implementation details, such as preprocessing and hyperparameters are often not mentioned in the publications. Together, these factors underline the need for transparent, harmonized benchmarking frameworks to ensure reproducibility and meaningful cross-study comparison.

English shared Tasks and benchmarks catalyzed progress [5][11], yet linguistic/institutional differences and scarce German corpora limit direct transfer [3][7][8][9]. Consequently, German-specific validation is indispensable. Future resources will only be impactful if paired with consistent, openly documented evaluation protocols [1][2] [12][13].

V. CONCLUSION AND FUTURE WORK

Under small-corpus constraints, partial adaptation of English SoTA is feasible. When data are severely lacking, BiLSTM-CRF is currently the most reliable option. gELECTRA remains promising, but our present evaluation is insufficient to judge its full potential.

Implications. For institutions in the GSA region with stringent privacy rules and limited annotation capacity, the most reliable near-term path appears to be a compact encoder (BiLSTM-CRF now and gELECTRA as data grow) complemented by deterministic components for highly structured PHI, deployed fully on-premises [1][2][10][16]. Results are preliminary given the synthetic, small-scale evaluation, and additional data plus careful tuning are likely

to narrow the gap for transformer encoders [1][8][16][17][20].

Looking ahead, further evaluation of gELECTRA on a real annotated corpus, potentially in combination with resources like the currently emerging German Medical Text Corpus (GeMTeX) [62], could provide a solid foundation for developing a privacy-compliant de-identification tool tailored to hospitals in the GSA region.

REFERENCES

[1] A. Kovacevic, B. Basaragin, N. Milosevic, and G. Nenadic "De-identification of clinical free text using natural language processing: a systematic review of current approaches," Artificial Intelligence in Medicine, 2024, ISSN: 0933-3657.

[2] B. Negash et al., "De-identification of free text data containing personal health information: a scoping review of reviews," International Journal of Population Data Science, vol. 8, 2023.

[3] C. Moore, J. Ranisau, W. Nelson, J. Petch, and A. Johnson, "PyCLIPSE: a library for de-identification of free-text clinical notes," arXiv, 2023.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1, 2019, pp. 4171-4186, doi: https://doi.org/10.18653/v1/N19-1423.

[5] A. Stubbs, C. Kotfila, and Ö. Uzuner, "Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task track 1," Journal of Biomedical Informatics, vol. 58, pp. S11–S19, 2015.

[6] A. E. W. Johnson et al., "MIMIC-III, a freely accessible critical care database," Scientific Data, vol. 3, 2016.

[7] J. L. Leevy, T. M. Khoshgoftaar, and F. Villanustre, "Survey on RNN and CRF models for de-identification of medical free text," Journal of Big Data, vol. 7, no. 73, 2020.

[8] L. Modersohn, S. Schulz, C. Lohr, and U. Hahn, "GRASCCO-The first publicly shareable, multiply-alienated German clinical text corpus," Studies in Health Technology and Informatics, vol. 296, pp. 66–72, 2022.

[9] T. Kolditz et al., "Annotating German Clinical Documents for De-Identification" Studies in Health Technology and Informatics, vol. 264, pp. 203–207, 2019.

[10] M. Baumgartner et al., "Masketeer: An Ensemble-Based Pseudonymization Tool with Entity Recognition for German Unstructured Medical Free Text," Future Internet, vol. 16, 2024.

[11] A. Stubbs, M. Filannino, and Ö. Uzuner, "De-identification of psychiatric intake records: Overview of 2016 CEGS N-GRID shared tasks track 1," Journal of Biomedical Informatics, vol. 75S, pp. S4–S18, 2017.

[12] H. M. Cooper, "Organizing knowledge syntheses: A taxonomy of literature reviews," Knowledge in Society, vol. 1, pp. 104–126, 1988.

[13] C. Wohlin et al., "Experimentation in Software Engineering," Berlin, Germany: Springer, pp. 9–20, 2012.

[14] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," arXiv, 2015.

[15] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," Proc. NAACL-HLT, San Diego, CA, USA, pp. 260–270, 2016.

[16] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: Pre-training text encoders as discriminators rather than generators," ICLR, Addis Ababa, Ethiopia, 2020.

[17] Deepset. *German ELECTRA (gelectra-base).* [Online]. Available from: https://huggingface.co/deepset/gelectra-base [retrieved: January 2026]

[18] Meta AI, *Llama 3.1.* [Online]. Available from: https://ai.meta.com/blog/llama-3-1 [retrieved: January 2026]

[19] VAGOsolutions. *Llama-3.1-8B-Instruct-German-SauerkrautLM*. [Online]. Available from: https://huggingface.co/VAGOsolutions/Llama-3.1-8B-Instruct-German-SauerkrautLM [retrieved: January 2026]

[20] C. Lohr et al., "GraSCCo_PHI - Graz Synthetic Clinical text Corpus with Protected Health Information Annotations," Zenodo, doi: 10.5281/zenodo.11502329.

[21] Explosion AI. *Training Pipelines & Models*. [Online]. Available from: https://spacy.io/usage/training#ner [retrieved: January 2026]

[22] Ollama. *Ollama.* [Online]. Available from: https://www.ollama.com [retrieved: January 2026]

[23] I. Neamatullah et al., "Automated de-identification of free-text medical records," BMC Medical Informatics and Decision Making, vol. 8, p. 32, 2008, doi: 10.1186/1472-6947-8-32.

[24] R. Kuo et al., "Comparative evaluation of large-language models and purpose-built software for medical record de-identification," Research Square, 2024, doi: 10.21203/rs.3.rs-4870585/v1. (*preprint*)

[25] F. J. Moreno-Barea et al, "Named entity recognition for de-identifying Spanish electronic health records," Comput. Biol. Med., vol. 185, p. 109576, 2025, doi: 10.1016/j.compbiomed.2024.109576.

[26] Z. Liu et al., "DeID-GPT: Zero-shot Medical Text De-Identification by GPT-4," arXiv, 2023, doi: 10.48550/arXiv.2303.11032.

[27] O. Dorémus et al., "Harnessing Moderate-Sized Language Models for Reliable Patient Data De-identification in Emergency Department Records: Algorithm Development, Validation, and Implementation Study," JMIR AI, 2025, doi: 10.2196/57828.

[28] J. L. Leevy and T. M. Khoshgoftaar, "A Short Survey of LSTM Models for De-identification of Medical Free Text," Proc. IEEE CIC 2020, Atlanta, GA, USA, pp. 117–124, 2020, doi: 10.1109/CIC50333.2020.00023.

[29] K. Arzideh, et al., "A Transformer-Based Pipeline for German Clinical Document De-Identification," Appl. Clin. Inform., vol. 16, no. 1, pp. 31–43, 2025, doi: 10.1055/a-2424-1989.

[30] M. Gunay, B. Keles, and R. Hizlan, "LLMs-in-the-Loop Part 2: Expert Small AI Models for Anonymization and De-identification of PHI Across Multiple Languages," arXiv, 2024, doi: 10.48550/arXiv.2412.10918.

[31] K. Tomanek, D. Wermter, and U. Hahn, "An Interactive De-Identification-System," Proceedings of the 5th International Symposium on Semantic Mining in Biomedicine (SMBM 2012), Zurich, Switzerland, 2012, doi: 0.5167/UZH-64476.

[32] H. Seuss et al., "Semi-automated De-identification of German Content Sensitive Reports for Big Data Analytics," Rofo, vol. 189, no. 7, pp. 661–671, 2017, doi: 10.1055/s-0043-102939.

[33] P. Richter-Pechanski, S. Riezler, and C. Dieterich, "De-Identification of German Medical Admission Notes," German Medical Data Sciences: A Learning Healthcare System, vol. 253, pp. 165–169, 2018, doi: 10.3233/978-1-61499-896-9-165.

[34] P. Richter-Pechanski, A. Amr, H. A Katus, and C. Dieterich, "Deep Learning Approaches Outperform Conventional Strategies in De-Identification of German Medical Reports," Studies in health technology and informatics, pp. 101-109, 2019, doi: 10.3233/SHTI190813.

[35] S. Sousa, A. M. Jantscher, M. Kröll, and R. Kern, "Large Language Models for Electronic Health Record De-Identification in English and German," Information, vol. 16, no. 2, p. 112, 2025, doi: 10.3390/info16020112.

[36] I. C. Wiest et al., "Deidentifying Medical Documents with Local, Privacy-Preserving Large Language Models: The LLM-Anonymizer," NEJM AI, vol. 2, no. 4, 2025, doi: 10.1056/aidbp2400537.

[37] P. J. Chambon, S. Bluethgen, J. Dreyfuss, A. Lungren, and D. Rubin, "Automated deidentification of radiology reports combining transformer and 'hide in plain sight' rule-based methods," JAMIA, vol. 30, no. 2, pp. 318–328, 2023, doi: 10.1093/jamia/ocac219.

[38] T. Ahmed, M. M. A. Aziz, and N. Mohammed, "De-identification of electronic health record using neural network," Scientific reports, vol. 10, no. 1, p. 18600, 2020, doi: 10.1038/s41598-020-75544-1.

[39] J. Y. Lee, F. Dernoncourt, O. Uzuner, and P. Szolovits, "Feature-Augmented Neural Networks for Patient Note De-identification," ClinicalNLP (COLING 2016), Osaka, Japan, 2016, pp. 17–22. doi: 10.48550/arXiv.1610.09704.

[40] L. Liu, O. Perez-Concha, A. Nguyen, V. Bennett, and L. Jorm, "De-identifying Australian hospital discharge summaries: An end-to-end framework using ensemble of deep learning models," Journal of biomedical informatics, vol. 135, 2022, doi: 10.1016/j.jbi.2022.104215.

[41] F. Dernoncourt, J. Y. Lee, O. Uzuner, and P. Szolovits, "De-identification of Patient Notes with Recurrent Neural Networks," arXiv, 2016, doi: 10.48550/arXiv.1606.03475.

[42] C. Aracena et al., "A Privacy-Preserving Corpus for Occupational Health in Spanish: Evaluation for NER and Classification Tasks," Proceedings of the 6th Clinical Natural Language Processing Workshop, Mexico City, Mexico: ACL, pp. 111–121, 2024, doi: 10.18653/v1/2024.clinicalnlp-1.11.

[43] M. Montserrat et al, "Automatic De-identification of Medical Texts in Spanish: the MEDDOCAN Track, Corpus, Guidelines, Methods and Evaluation of Results," Proceedings of the Iberian Languages Evaluation Forum, 2019.

[44] A. Aloqaily et al., "Deep Learning Framework for Advanced De-Identification of Protected Health Information," Future Internet, vol. 17, no. 1, p. 47, 2025, doi: 10.3390/fi17010047.

[45] K. Murugadoss et al., "Scaling text de-identification using locally augmented ensembles," medRxiv, 2024. (*preprint*)

[46] A. E. W. Johnson, L. Bulgarelli, and T. J. Pollard, "Deidentification of free-text medical records using pre-trained bidirectional transformers," ACM CHIL '20, Toronto, ON, Canada, pp. 214–221, 2020, doi: 10.1145/3368555.3384455.

[47] A. Berzi et al., "NLP-based removal of personally identifiable information from Hungarian electronic health records," Front. Artif. Intell., vol. 8, 2025, doi: 10.3389/frai.2025.1585260.

[48] M. M. Anjum, N. Mohammed, and X. Jiang, "De-identification of Unstructured Clinical Texts from Sequence to Sequence Perspective," Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security, pp. 2438–2440, 2021, doi: 10.1145/3460120.3485354.

[49] K. Murugadoss et al., "Building a best-in-class automated de-identification tool for electronic health records through ensemble learning," Patterns, vol. 2, no. 6, p. 100255, 2021, doi: 10.1016/j.patter.2021.100255.

[50] L. Liu et al., "Web-Based Application Based on Human-in-the-Loop Deep Learning for Deidentifying Free-Text Data in Electronic Medical Records: Development and Usability Study" Interactive journal of medical research, vol. 12, 2023, doi: 10.2196/46322

[51] Y.-S. Zhao, K.-L. Zhang, H.-C. Ma, and K. Li, "Leveraging text skeleton for de-identification of electronic medical records," BMC Med. Inform Decis Mak, vol. 18, suppl. 1, p. 18, 2018, doi: 10.1186/s12911-018-0598-6.

[52] K. Li, Y. Chai, H. Zhao, X. Nan, and Y. Zhao, "Learning to Recognize Protected Health Information in Electronic Health Records with Recurrent Neural Network," Natural Language Understanding and Intelligent Applications, Cham: Springer, pp. 575–582, 2016, doi: 10.1007/978-3-319-50496-4_51.

[53] V. Kocaman, Y. Mellah, H. Haq, and D. Talby, "Automated De-Identification of Arabic Medical Records," Proceedings of ArabicNLP 2023, pp. 33–40, Singapore, 2023, doi: 10.18653/v1/2023.arabicnlp-1.4.

[54] S. Singh et al., "Generation and De-Identification of Indian Clinical Discharge Summaries using LLMs," Proceedings of the 23rd Workshop on Biomedical Natural Language Processing, Bangkok, Thailand, pp. 342–362, 2024, doi: 10.18653/v1/2024.bionlp-1.26.

[55] G. López-García et al., "Named Entity Recognition for De-identifying Real-World Health Records in Spanish," Computational Science – ICCS 2023, Prague, Czech Republic, pp. 228–242, 2023, doi: 10.1007/978-3-031-36024-4_17.

[56] K. Xu, Y. Song, and J. Ma, "Identifying protected health information by transformers-based deep learning approach in Chinese medical text," Health Informatics Journal, vol. 31, no. 1, 2025, doi: 10.1177/14604582251315594.

[57] J. An et al., "De-identification of clinical notes with pseudo-labeling using regular expression rules and pre-trained BERT," BMC medical informatics and decision making, vol. 25, no. 1, 2025, doi: 10.1186/s12911-025-02913-z.

[58] M. E. Azzouzi et al., "Automatic de-identification of French electronic health records: a cost-effective approach exploiting distant supervision and deep learning models," BMC medical informatics and decision making, vol. 24, no. 1, p. 54, 2024, doi: 10.1186/s12911-024-02422-5.

[59] A. Paul, D. Shaji. L. Han, W. Del-Pinto, and G. Nenadic, "DeIDClinic: A Multi-Layered Framework for De-identification of Clinical Free-text Data," arXiv, 2024. doi: 10.48550/arXiv.2410.01648.

[60] S. Meystre and P. Heider, "High Accuracy Open-Source Clinical Data De-Identification: The CliniDeID Solution," Studies in Health Technology and Informatics, vol. 310, pp.1370-1371, 2024, doi: 10.3233/SHTI231199.

[61] A. Ezen-Can, "A Comparison of LSTM and BERT for Small Corpus," arXiv, 2020. doi: 10.48550/arXiv.2009.05451.

[62] Medizininformatik-Initiative. *GeMTeX – Medizinische Texte für die Forschung automatisiert erschließen [GeMTeX – Automated extraction of medical texts for research]*. [Online]. Available from: https://www.medizininformatik-initiative.de/de/gemtex-medizinische-texte-fuer-die-forschung-automatisiert-erschliessen [retrieved: January 2026]