Beyond De-Identification: A Structured Approach for Defining and Detecting Indirect Identifiers in Medical Texts

Ibrahim Baroud[1,2], **Lisa Raithel**[1,2,3], **Sebastian Möller**[1,2], **Roland Roller**[2]
[1]Quality & Usability Lab, Technische Universität Berlin
[2]German Research Center for Artificial Intelligence (DFKI), Berlin
[3]BIFOLD - Berlin Institute for the Foundations of Learning and Data
ibrahim.baroud@tu-berlin.de

Abstract

Sharing sensitive texts for scientific purposes requires appropriate techniques to protect the privacy of patients and healthcare personnel. Anonymizing textual data is particularly challenging due to the presence of diverse unstructured direct and indirect identifiers. To mitigate the risk of re-identification, this work introduces a schema of nine categories of indirect identifiers designed to account for different potential adversaries, including acquaintances, family members and medical staff. Using this schema, we annotate 100 MIMIC-III discharge summaries and propose baseline models for identifying indirect identifiers. We release the annotation guidelines, annotation spans (6,199 annotations in total) and the corresponding MIMIC-III document IDs to support further research in this area.[1]

1 Introduction

Access to data remains a major bottleneck in developing machine learning models for healthcare. Since data contains sensitive details about individuals, it cannot be shared readily outside hospitals. Interactions with legal departments and data security can be cumbersome, and regulations are somewhat unclear, particularly where text is concerned. However, the concept of de-identification is well-defined: according to HIPAA,[2] it requires the removal of a list of direct identifiers, known as protected health information (PHI),[3] including names and addresses.

Classical de-identification of text data has been explored for many years with various approaches (Sweeney, 1996; Gupta et al., 2004; He et al., 2015; Kocaman et al., 2023) and state-of-the-art de-identification systems achieve an F_1-score $\geq 95\%$ on academic benchmarks (Kocaman et al., 2023; Yogarajan et al., 2020). However, additional manual effort is needed to remove remaining PHIs, and more importantly, unstructured text often contains **additional information beyond PHIs** that can reveal an individual's identity (Feder et al., 2020), making the manual inspection process even more complex.

[...] Patient is a *33-year-old* male, admitted at *12:20* after a *motor vehicle accident*.
[...] He *works as a carpenter* and *lives with his 28-year-old girlfriend in assisted living*. No known *health insurance*, and he is *currently on disability assistance*. [...] He was noted to be *obese (BMI 32)* with a *height of 178 cm* and *weight of 110 kg*.
[...] He was evaluated by the *Emergency Department team* and consulted with *Orthopedics* for suspected fractures. [...] Patient reports *playing basketball once a week* [...].

Figure 1: A snippet of a fictitious discharge summary with annotations according to our IPI schema in red.

The concept of anonymization goes further: it is defined as an irreversible procedure that is applied to the data such that no information can be linked to any specific individual anymore (Meystre et al., 2010). While the terms de-identification and anonymization are often used interchangeably, they refer to distinct concepts (Chevrier et al., 2019). De-identification focuses solely on removing direct identifiers, whereas anonymization must also address indirect identifiers. Indirect identifiers are pieces of information that are potentially publicly known about an individual but do not lead to reidentification when considered alone. However, in combination with other background or external knowledge, they can be used to uniquely identify an individual (Pilán et al., 2022). Figure 1 shows a synthetic discharge summary with highlighted information (beyond direct identifiers) that may help reveal a person's identity.

[1] https://zenodo.org/records/15044596

[2] The U.S. Health Insurance Portability and Accountability Act of 1996.

[3] https://www.hhs.gov/hipaa/for-professionals/special-topics/de-identification/index.html

Proceedings of the Sixth Workshop on Privacy in Natural Language Processing, pages 75–85
April 4, 2025 ©2025 Association for Computational Linguistics

Despite the importance of anonymization, relatively few studies have systematically addressed text anonymization beyond traditional PHI detection. Gardner and Xiong (2008) developed a system for extracting and suppressing sensitive information other than PHIs, but it was limited to diagnoses. Kolditz et al. (2019) created a dataset with PHIs and added more categories, namely medical units, relatives and typists. Feder et al. (2020) annotated a set of demographic traits in clinical notes and proposed a framework for detecting sentences that include such traits. Pilán et al. (2022) presented a benchmark dataset comprising annotations of court cases and evaluation metrics to assess the performance of anonymization methods. The annotations cover categories such as names and quantities, and annotators mark each of the entities as a direct or indirect identifier. Moreover, Yang et al. (2024) proposed a framework for text anonymization based on large language models (LLMs). This framework measures anonymization success simply by checking whether an adversarial LLM can guess the name of the person to whom the text belongs.

Building on prior work, our study defines and identifies information beyond traditional personal health identifiers within a controlled framework. We introduce a schema of indirect personal identifiers (IPIs) optimized for a medical context and apply it to annotate relevant spans in discharge summaries from the Medical Information Mart for Intensive Care (MIMIC-III) dataset (Johnson et al., 2016). We define the problem of structurally identifying IPIs as a span classification problem, rather than a sentence classification problem as in Feder et al. (2020), to avoid removing whole sentences (which might include other medical information) and to reduce information loss during anonymization. Finally, we evaluate the performance of various models in detecting the annotated identifiers.

2 Indirect Personal Identifiers (IPI)

The type of information that may lead to re-identification in a given text is domain-dependent and requires unique analysis (Sweeney, 2000). In the following, we introduce a schema covering aspects of indirect personal identifiers (IPI) and use it to annotate spans in discharge summaries from MIMIC-III. To construct our dataset, we randomly sampled 100 summaries with lengths ranging from 500 to 2,500 words.[4]

2.1 IPI Schema

Our proposed schema builds on related work by Kolditz et al. (2019) and Feder et al. (2020), as well as our own manual analysis of discharge summaries. From prior work, we incorporate concepts like *medical unit* (Kolditz et al., 2019), expanding it to include medical services, teams and medical personnel. We adapt *family structure* from Feder et al. (2020), broadening it to include family decisions. We also integrate *living arrangements* into a new category, DETAILS, which covers indirect identifiers such as addresses (e.g., ‘lives in prison’), dates (‘he turned 18 right before COVID started’), and references to other PHIs like license numbers.

Additionally, we adapt the category *occupation* into SEC, which covers socio-economic and criminal history. Our LFSTL category includes habits, sports and diet alongside the *drug* category from Feder et al. (2020). We redefine the category *casually noticeable* in our category APPEARANCE to specifically cover body piercings, tattoos and scars. Based on our manual analysis, we introduce TIME to capture time-related expressions such as timestamps for taking lab values, admission days and time references around events such as surgeries. A brief overview of our final categories is provided below,[5] with further details available in Appendix A.

APPEARANCE Descriptions of appearance, e.g. *freshly healed scar behind right ear*, and mentions of weight, height or body modifications.

CIRCUMSTANCES Any mention of an event (e.g. an accident) that caused an injury or happened in a medical facility. This category also includes specific statements or behavior, e.g. *crashed his car into a dumpster* or *refused medication because she does not believe in it*.

SEC Mentions of information concerning socio-economic or criminal history, such as employment (e.g. *is a retired police officer*), health insurance (e.g. *has no health insurance*) or social/legal status (*does not have valid papers*).

FAMILY Any mention of family-related information, such as being adopted, as well as the

[4] More details on the dataset in Appendix B.

[5] The following examples were created by the authors to avoid presenting data from MIMIC-III directly.

family's medical history or involvement (e.g. *daughter serves as her health care proxy*).

FCLT_PERSONNEL Mentions of healthcare facilities (*ICU*) or medical personnel (*nursing team*).

TIME All mentions of age or time-related information (e.g., *postoperative day number 5*).

LFSTL Regular activities and habits, such as sports or diet (e.g. *reports sticking to low-sodium diet*), but also tobacco, alcohol or substance use.

DETAILS All mentions of PHIs that were not detected, or a description of a PHI (e.g. *lives in a halfway house*, which reveals information about the person's address).

OTHER All other kinds of non-medical but infrequent information that might be sensitive, e.g. languages, ethnicity or sexual orientation.

2.2 Data Annotation

Two annotators independently labelled the same set of 100 de-identified discharge summaries using the nine categories described above. The annotations were then consolidated, meaning that all annotations from both annotators were discussed and resolved into one final version of the corpus presented here. Inter-annotator agreement (IAA) was calculated using the average pairwise relaxed F_1-score between the annotators' marked entities.[6] We chose F_1-score for calculating agreement as it proved to be a more usable and interpretable measure for annotations such as span classification, where the number of negative examples is very large (or unknown) and the probability of chance agreement on positive examples (the desired spans) is close to zero (Hripcsak and Rothschild, 2005). The overall agreement resulted in an F_1-score of 0.87. Table 4 in Appendix B lists the scores for each category. The annotators achieved the highest agreement in the categories TIME ($F_1 = 0.89$), LFSTL ($F_1 = 0.88$) and FAMILY ($F_1 = 0.87$), and the lowest on DETAILS ($F_1 = 0.41$).

The finalized dataset consists of 6,199 annotations, the majority of them belonging to the categories TIME (64.62%) and FCLT_PERSONNEL (22.92%). This is expected, as most discharge summaries contain detailed temporal descriptions, department consultations and precise timestamps, such as when lab values were recorded. In contrast, information such as spoken languages or accident details appeared less frequently, as they were case-dependent and varied based on the typist's preference. Table 1 shows the number of annotations per category and their percentage in the overall annotations.

Category	#Annotations	Proportion
FAMILY	273	4.4%
APPEARANCE	132	2.13%
CIRCUMSTANCES	99	1.6%
SEC	59	0.95%
FCLT_PERSONNEL	1421	22.92%
TIME	4006	64.62%
LFSTL	144	2.32%
DETAILS	32	0.52%
OTHER	33	0.53%

Table 1: Number of annotations per category in 100 discharge summaries from MIMIC-III.

2.3 Data Characteristics

Overall, we focused on identifying indirect identifiers on the span level that may either be publicly known or describe a person's status, behaviour or appearance. Our final curated annotations reveal various such risks. For example, spans labeled as CIRCUMSTANCES contain descriptive information about accidents that could facilitate re-identification by witnesses. These details may enable an adversary to retrieve additional information about the patient, e.g. by searching online to find reports about the incident. Moreover, this category might encompass other sensitive or memorable descriptions, such as instances of patient aggression toward staff or refusal of medication.

The 59 annotations from the SEC category reveal information about a person's criminal history, which is public information in the U.S. (Jacobs and Larrauri, 2012) and therefore easy to look up even for a layperson. This category covers mentions of the patient being incarcerated, which may, in some cases, reveal the patient's exact address. Finally, the annotations include various information about patients' social status, such as being homeless or not having health insurance, or lifestyle, such as information about drinking, smoking or sports. Although these mentions are relatively infrequent in the dataset, they may pose a high re-identification risk. Unique or rare characteristics –

[6] Details about the annotators and IAA can be found in Appendix B.

especially those that distinguish an individual from the broader population – can drastically narrow down the pool of potential matches, making re-identification more feasible.

3 Experiments

To provide a first baseline for the automatic detection of the proposed set of indirect identifiers in medical texts, we experimented with BERT (Devlin et al., 2019) as well as open-source LLMs. We split the data into training (60%), development (15%) and test (25%) sets, and used the dev set for hyperparameter optimization. Table 2 shows statistics about the final data split.

We fine-tuned a BERT model for span classification using the HuggingFace library (Wolf et al., 2020). For the LLM experiments, we used Llama-3.1-8b-Instruct, Mistral-7B-Instruct-v0.3 and Qwen2.5-14B-Instruct in both zero-shot and three-shot settings leveraging Declarative Self-improving Python (DSPy) (Khattab et al., 2024) to automatically refine and optimize the prompt and Pydantic[7] to obtain structured and type-validated output from the LLMs. An example prompt is shown in Appendix E. We implemented an LLM agent for each category and provided DSPy with the description of each category as defined in the annotation guidelines. Model performance was assessed using relaxed precision, recall and F_1-score. Further details on data preprocessing, model fine-tuning and evaluation can be found in Appendix C.

	train	**dev**	**test**	**total**
#documents	60	15	25	100
#sections	592	162	253	1007
#annotations	3712	927	1560	6199

Table 2: Statistics for the train, development and test sets. '#sections' represents the number of sections the documents were split into for each set.

3.1 Results

Detailed evaluation results for the BERT model can be found in Table 3. Notably, recall is higher than precision in almost all cases. Phrases containing socio-economic or criminal information (SEC), medical facilities and personnel (FCLT_PERSONNEL) and time expressions (TIME) achieve higher scores than the other categories; i.e. less frequent categories tend to have a lower F_1-score, which was also true for the IAA scores. The lightweight LLMs, which are explored here for the first time for this specific task, performed poorly on the test set with F_1-score $\leq 51\%$ (micro) and recall $\leq 47\%$ (more details in Table 5). The 3-shot setting did not always improve performance. Interestingly, performance dropped in some cases when providing the models with examples. A similar phenomenon was also observed in Kwon et al. (2024) when using Llama3 for information extraction: the model achieved better results in some cases in the zero-shot setting in comparison to few-shot. This and the overall low performance of the LLMs in comparison to BERT highlights our doubts about the suitability and effectiveness of using LLMs for extracting our proposed categories of indirect identifiers. Moreover, our evaluation showed that the LLMs sometimes failed to follow the pre-defined output format and preserve the originality of the spans in the original texts. Moreover, they frequently hallucinated and extracted irrelevant or non-existent information.

Category	**P**	**R**	**F$_1$**	**Support**
DETAILS	0.13	0.50	0.21	4
FAMILY	0.67	0.96	0.79	73
APPEARANCE	0.52	0.59	0.55	29
CIRCUMSTANCES	0.18	0.23	0.20	30
SEC	0.59	0.71	0.65	14
FCLT_PERSONNEL	0.80	0.92	0.85	362
TIME	0.84	0.97	0.90	1006
LFSTL	0.57	0.86	0.68	35
OTHER	0.20	0.14	0.17	7
micro average	0.78	0.93	0.85	1560
macro average	0.50	0.65	0.55	1560

Table 3: Evaluation results on the test set for the **BERT-based** system in **P**recision, **R**ecall, and **F$_1$** score. Support shows the number of examples in the test set.

4 Discussion

As expected, the BERT-based model clearly outperformed the lightweight LLMs in both zero-shot and 3-shot settings, corroborating the results of Naguib et al. (2024) about BERT superiority against LLMs for span classification. This suggests that LLMs may be more powerful as supportive tools used to validate anonymization systems through inferring hidden information as proposed by Staab et al. (2024) rather than being used for span classification.

The BERT model shows a satisfactory micro

[7] https://pypi.org/project/pydantic/

F_1-score, with its comparably high recall being particularly advantageous for anonymization, as missing sensitive information can have serious consequences. However, the low macro F_1-score combined with the strong imbalance of the annotated categories indicates that the model struggles to detect less frequent, yet more critical, categories.

One reason for this may be the limited amount of training data, hampering the model's ability to learn robust representations for rare categories. Additionally, the inherent linguistic complexity within categories further complicates the task. In contrast to PHIs, such as names or addresses, which usually follow similar patterns across documents, IPIs exhibit greater lexical and semantic diversity. This not only makes them more challenging, but also highlights the urgency of accurately identifying them for effective anonymization. Given that annotating additional documents is both time- and resource-intensive, especially when rare events must be captured in sufficient numbers, it may be more realistic to investigate methods that perform well in low-resource scenarios.

5 Conclusion

In this work, we introduced a dataset along with an annotation schema designed to capture a wide range of indirect identifiers in medical texts. The schema is inspired by medical records, but is adaptable to other domains and text genres with minimal modifications. We evaluated the performance of BERT and LLMs in detecting the proposed categories. The overall performance of the models highlights the inherent difficulty of this task, particularly in identifying less frequent and diverse indirect identifiers. However, our work provides a foundation for further exploration and adaptation, with an eye to improving privacy through structural information detection. In future work, we aim to develop a framework that (k-)anonymizes the proposed indirect identifiers and study the utility of the anonymized texts on downstream tasks.

Acknowledgements

We thank the anonymous reviewers for their constructive feedback on this paper. This research was supported by the German Federal Ministry of Education and Research (BMBF) through the project Medinym (16KISA006 and 16KISA007) and the grant BIFOLD25B.

Limitations

Our list of categories is diverse; however, indirect identifiers should not be limited to it, and further studies should explore more potential risks in unstructured data that do not fall under these categories. We plan to test the scalability of our schema to other datasets, languages and domains (such as legal or financial), but accessing similar relevant data is very limited due to privacy concerns, especially in languages other than English.

The LLM experiments are intended to provide a different baseline approach rather than to compare performance with the BERT model, as such a comparison would be unfair in a zero- or few-shot setting. The LLM approach could be improved, for example, by using bigger models or performing an instruction tuning using the training set instead of evaluating the models in a zero- or few-shot setting. We plan to use LLMs to augment the training set with synthetically generated examples to solve the problem of low numbers of examples for certain categories, which also did not suffice to train the BERT model.

BERT-based models have been shown to work well in NER tasks; however, they cannot be fully relied on for finding all instances of potentially sensitive information. Instead, these models can be used as a complement to help humans speed up the process of enhancing privacy. As for LLMs, we would not trust them to produce complete and reliable results since our experiments showed unfaithful output in terms of format (which hinders a structured evaluation) and "hallucinations."

We did not experiment with a hybrid approach (e.g., combining regular expressions and the approaches described) to improve the detection of categories with formulaic patterns for which we expect a better performance using regular expression, such as TIME.

Ethical Considerations

The data used in the above work is publicly available, de-identified data from the MIMIC-III database and therefore does not expose any patients or medical staff. It is only available after registration and training. We state that we only annotated potential indirect identifiers and did not attempt to re-identify any patients. All examples in this paper were created by the authors. They resemble texts from MIMIC-III, but are not copied from real discharge summaries. We only release the annotations and document IDs from MIMIC-III, but **not** the documents themselves.

Broader Impact Statement

This work contributes to protecting patient privacy by identifying and categorizing indirect personal identifiers in medical discharge summaries which are not considered in de-identification. Our annotated dataset offers a valuable resource for developing and evaluating privacy-enhancing machine learning models. Despite being optimized for medical discharge summaries, we encourage the further use and development of our schema in other domains, e.g., the legal and finance domain, to enhance data privacy and data sharing.

References

Raphaël Chevrier, Vasiliki Foufi, Christophe Gaudet-Blavignac, Arnaud Robert, and Christian Lovis. 2019. Use and understanding of anonymization and de-identification in the biomedical literature: scoping review. *Journal of medical Internet research*, 21(5):e13484.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *North American Chapter of the Association for Computational Linguistics*.

Amir Feder, Danny Vainstein, Roni Rosenfeld, Tzvika Hartman, Avinatan Hassidim, and Yossi Matias. 2020. Active deep learning to detect demographic traits in free-form clinical notes. *Journal of Biomedical Informatics*, 107:103436.

James Gardner and Li Xiong. 2008. HIDE: an integrated system for health information DE-identification. In *2008 21st IEEE international symposium on computer-based medical systems*, pages 254–259. IEEE.

Dilip Gupta, Melissa Saul, and John Gilbertson. 2004. Evaluation of a deidentification (De-Id) software engine to share pathology reports and clinical documents for research. *American journal of clinical pathology*, 121(2):176–186.

Bin He, Yi Guan, Jianyi Cheng, Keting Cen, and Wenlan Hua. 2015. CRFs based de-identification of medical records. *Journal of biomedical informatics*, 58:S39–S46.

George Hripcsak and Adam S Rothschild. 2005. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American medical informatics association*, 12(3):296–298.

James B Jacobs and Elena Larrauri. 2012. Are criminal convictions a public matter? The USA and Spain. *Punishment & Society*, 14(1):3–28.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2024. DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines.

Veysel Kocaman, D Talby, and H Ul Hak. 2023. RWD143 Beyond Accuracy: Automated De-Identification of Large Real-World Clinical Text Datasets. *Value in Health*, 26(12):S532.

Tobias Kolditz, Christina Lohr, Johannes Hellrich, Luise Modersohn, Boris Betz, Michael Kiehntopf, and Udo Hahn. 2019. Annotating German clinical documents for de-identification. In *MEDINFO 2019: Health and Wellbeing e-Networks for All*, pages 203–207. IOS Press.

Yeonsu Kwon, Jiho Kim, Gyubok Lee, Seongsu Bae, Daeun Kyung, Wonchul Cha, Tom Pollard, Alistair Johnson, and Edward Choi. 2024. EHRCon: Dataset for Checking Consistency between Unstructured Notes and Structured Tables in Electronic Health Records. *ArXiv*, abs/2406.16341.

Stephane M Meystre, F Jeffrey Friedlin, Brett R South, Shuying Shen, and Matthew H Samore. 2010. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC medical research methodology*, 10:1–16.

Ines Montani and Matthew Honnibal. Prodigy: A modern and scriptable annotation tool for creating training data for machine learning models.

Marco Naguib, Xavier Tannier, and Aurélie Névéol. 2024. Few shot clinical entity recognition in three languages: Masked language models outperform LLM prompting. *arXiv preprint arXiv:2402.12801*.

Ildikó Pilán, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet. 2022. The Text Anonymization Benchmark (TAB): A Dedicated Corpus and Evaluation Framework for Text Anonymization. *Computational Linguistics*, 48(4):1053–1101.

Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. 2013. SemEval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 341–350, Atlanta, Georgia, USA. Association for Computational Linguistics.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. 2024. Beyond Memorization: Violating Privacy via Inference with Large Language Models. In *The Twelfth International Conference on Learning Representations*.

Latanya Sweeney. 1996. Replacing personally-identifying information in medical records, the Scrub system. In *Proceedings of the AMIA annual fall symposium*, page 333. American Medical Informatics Association.

Latanya Sweeney. 2000. Simple Demographics Often Identify People Uniquely.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Tianyu Yang, Xiaodan Zhu, and Iryna Gurevych. 2024. Robust Utility-Preserving Text Anonymization Based on Large Language Models. *arXiv preprint arXiv:2407.11770*.

Vithya Yogarajan, Bernhard Pfahringer, and Michael Mayo. 2020. A review of automatic end-to-end de-identification: Is high accuracy the only metric? *Applied Artificial Intelligence*, 34(3):251–269.

A Detailed Descriptions of the IPI Categories

APPEARANCE Mention of a person's (also infant's) weight, height or a description of a person's body or body modifications, e.g., a scar under the eye, very tall, very short, gained/lost weight over a specific period of time, tattoos, piercings, etc.

CIRCUMSTANCES Any mention or description of an event (accident, storm, wildfire, etc.) that caused, e.g., a person's injury or happened in the clinical center such as patient being aggressive, rejecting help or medicine, leaving AMA (including discussions about the decision with persons outside the family) or injuring hospital staff. Additionally, details about how the person was brought into the hospital or mentions of statements, requests or complaints expressed by the person.

SEC Any mention of specific information about the person's employment (e.g., *is a retired police officer*) or criminal history, health insurance (e.g., *has no health insurance* or *has a legal guard*) or social status such as homelessness or living in subsidized housing.

FAMILY All mentions of detailed family-related information about the person such as being adopted, having a twin sibling or having had an in vitro fertilization pregnancy. Furthermore, specific descriptions of the family's medical history (e.g., *parent died at age 40*) or involvement (e.g., *patient's daughter serves as her health care proxy*).

FCLT_PERSONNEL All mentions of hospital names, hospital units, labs, departments, facilities, consulting services/teams, floor and rooms, medical branches, outside doctors.

TIME Mentions of age or time-related information, e.g. *postoperative day number 2*, *day of delivery number 13*, *day of life 6*, exact mentions of times when lab values were taken, or exact times about when medications should be taken. Do not consider times related to the medical condition itself, e.g., *stopped breathing for 30 secs*.

LFSTL Hobbies and Lifestyle: such as sports or playing an instrument. Lifestyle: e.g. information about the patient's diet or private lifestyle.

DETAILS All mentions of PHIs that were not detected and de-identified automatically or an abstract/indirect description of a PHI, for instance regarding address (e.g., *lives in a halfway house* or *lives in prison*). Any information not related to PHIs such as weight or medical units are not part of this category and should be annotated as described in the other categories above. For consistency, the following are the PHIs to consider for this category: Name, email addresses, geographic details, dates directly related to the individual, telephone, fax numbers, social security numbers, medical record numbers, health plan beneficiary numbers, account numbers, certificate and license numbers, vehicle and device identifiers, biometric identifiers and facial photograph, URL, IP addresses.

OTHER Other kinds of non-medical information that may be too sensitive to keep in the data e.g. languages, ethnicity (e.g., *Caucasian*, *AAF* etc.) and sexual orientation.

B Data and Annotation Details

Data The discharge summaries we use for demonstrating our schema are randomly sampled from the Medical Information Mart for Intensive Care (MIMIC-III) dataset (Johnson et al., 2016). It comprises health-related data from over 40,000 patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. Among other types of data, such as patient demographics, the database also includes various types of textual data, such as diagnostic reports and discharge summaries. We chose discharge summaries for our study, since these are richer in information than other notes in MIMIC-III.

Annotation Tool For annotation, we used Prodigy (Montani and Honnibal), version 1.11.11. It was run on a secure, lab-internal server; access was only permitted to the authors.

Annotators The annotation team included one female and one male researcher, each with a different cultural background. Both annotators are fluent in English, though it is not their native language. One has expertise in computer science and data anonymization, and the other has experience in biomedical natural language processing. Neither has formal medical training, but both have experience in computational research and have contributed to various annotation projects in a research setting. Both annotators were compensated as part of their regular researcher roles.

Inter-Annotator Agreement The reported pairwise F_1-score is based on partial matches: a true positive exists when the compared spans overlap with at least one token and have the same label. We focus on partial matches because the exact span is not as important as in other entity recognition tasks; the main difficulty lies in finding the relevant information and removing it—anonymizing a longer span does not hurt the patient.

C Model Training and Evaluation Details

Category	$\mathbf{F_1}$-Score
DETAILS	0.41
FAMILY	0.87
APPEARANCE	0.62
CIRCUMSTANCES	0.59
SEC	0.78
FCLT_PERSONNEL	0.85
TIME	0.89
LFSTL	0.88
OTHER	0.52
micro average	0.87
macro average	0.71

Table 4: Inter-annotator agreement overall and per category using partial match pairwise F_1-scores (Hripcsak and Rothschild, 2005).

Data Preprocessing In order to train an NER model, we converted the Prodigy annotations (each represented with a span start and end) to word-level annotations. Words annotated as part of a category received label prefixes `B` when they are at the beginning of a category, `I` when they lie within the category, and finally, words that were not part of any category received the label `O` (out). Since BERT cannot handle sequences longer than 512 sub-tokens, we split the discharge summaries into sections to avoid truncation and information loss. Prodigy's annotation output is already pretokenized and we used the pre-trained BERT-base-cased tokenizer for subword tokenization.

BERT Fine-Tuning For choosing the hyperparameters, a bert-base-cased model[8] was fine-tuned for maximally 15 epochs (early stopping after two epochs' patience) on the training set and evaluated on the development set using a grid search over learning rate values (1e-5, 2e-5, 3e-5, 4e-5, 5e-5) and batch size values (4, 8, 16). After selecting the hyperparameters, we trained a BERT model on 75% of the data (training and development combined) using the best-performing hyperparameters: 8 epochs, 3e-5 as the learning rate and 8 as the batch size.

Evaluation Details We evaluated on the held-out test set using the `nervaluate` package,[9] which is a Python implementation for evaluating NER models as defined in the SemEval 2013 - 9.1 task (Segura-Bedmar et al., 2013). We report the results following the type evaluation schema, which

[8] https://huggingface.co/google-bert/bert-base-cased

[9] https://github.com/MantisAI/nervaluate

requires some overlap between the system-tagged entity and the gold-standard annotation.

Model	P	R	$\mathbf{F_1}$	Support
Llama-3.1-8B	0.08	0.40	0.13	1560
Llama-3.1-8B 3-shot	0.18	0.35	0.24	1560
Mistral-7B-v0.3	0.17	**0.47**	0.25	1560
Mistral-7B-v0.3 3-shot	0.05	0.30	0.09	1560
Qwen2.5-14B	**0.64**	0.42	**0.51**	1560
Qwen2.5-14B 3-shot	**0.64**	0.28	0.39	1560
Qwen2.5-72B*	0.48	**0.47**	0.48	1560

Table 5: Micro-averaged test results for each LLM showing precision (**P**), recall (**R**) and F_1-score ($\mathbf{F_1}$). *This is the 8-bit quantized version of this model. Values in **Bold** represent the highest performance for each metric among all tested LLMs.

Use of AI Assistants ChatGPT was partially used as an AI assistant for coding support.

Computing Environment The following packages were used for conducting the experiments:

- Transformers version 4.44.2[10]
- spacy version 3.7.5[11]
- Prodigy version 1.11.11[12]

The BERT experiments were run on a T4 GPU with 16GB. The LLMs were run on 2x NVIDIA RTX A6000 with 48GB each.

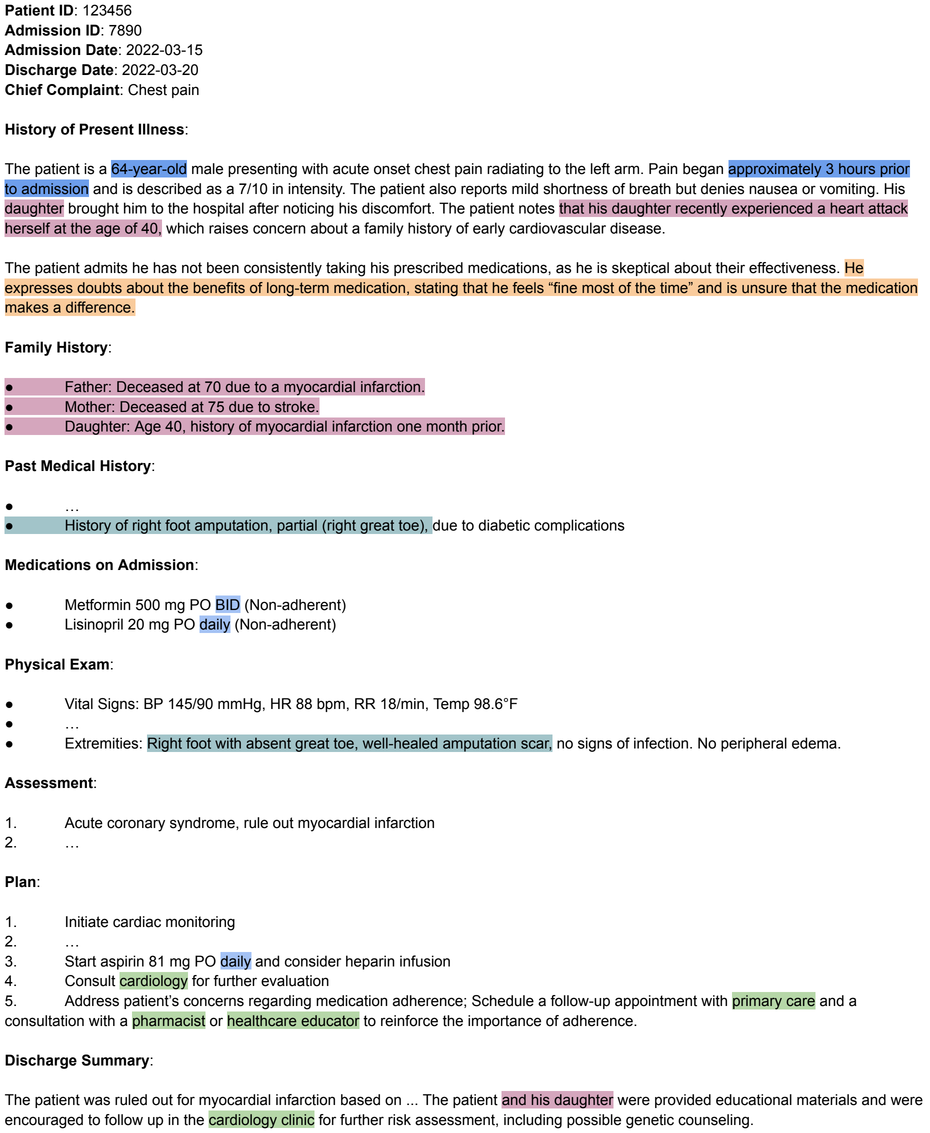
[10] https://huggingface.co/

[11] https://spacy.io/

[12] https://prodi.gy/

D Example Annotation

Figure 2 shows an example of how the discharge summaries were annotated.



Patient ID: 123456
Admission ID: 7890
Admission Date: 2022-03-15
Discharge Date: 2022-03-20
Chief Complaint: Chest pain

History of Present Illness:

The patient is a 64-year-old male presenting with acute onset chest pain radiating to the left arm. Pain began approximately 3 hours prior to admission and is described as a 7/10 in intensity. The patient also reports mild shortness of breath but denies nausea or vomiting. His daughter brought him to the hospital after noticing his discomfort. The patient notes that his daughter recently experienced a heart attack herself at the age of 40, which raises concern about a family history of early cardiovascular disease.

The patient admits he has not been consistently taking his prescribed medications, as he is skeptical about their effectiveness. He expresses doubts about the benefits of long-term medication, stating that he feels “fine most of the time” and is unsure that the medication makes a difference.

Family History:

- Father: Deceased at 70 due to a myocardial infarction.
- Mother: Deceased at 75 due to stroke.
- Daughter: Age 40, history of myocardial infarction one month prior.

Past Medical History:

- …
- History of right foot amputation, partial (right great toe), due to diabetic complications

Medications on Admission:

- Metformin 500 mg PO BID (Non-adherent)
- Lisinopril 20 mg PO daily (Non-adherent)

Physical Exam:

- Vital Signs: BP 145/90 mmHg, HR 88 bpm, RR 18/min, Temp 98.6°F
- …
- Extremities: Right foot with absent great toe, well-healed amputation scar, no signs of infection. No peripheral edema.

Assessment:

1. Acute coronary syndrome, rule out myocardial infarction
2. …

Plan:

1. Initiate cardiac monitoring
2. …
3. Start aspirin 81 mg PO daily and consider heparin infusion
4. Consult cardiology for further evaluation
5. Address patient’s concerns regarding medication adherence; Schedule a follow-up appointment with primary care and a consultation with a pharmacist or healthcare educator to reinforce the importance of adherence.

Discharge Summary:

The patient was ruled out for myocardial infarction based on ... The patient and his daughter were provided educational materials and were encouraged to follow up in the cardiology clinic for further risk assessment, including possible genetic counseling.

Figure 2: A (generated) discharge summary with annotations based on the proposed schema.

E Example Prompts

Figure 3 shows an example prompt that was used with DSPy to extract the FCLT_PERSONNEL category. Note that the format is the same for the other categories; only the descriptions vary depending on the category that the model is supposed to extract.

Example Prompt

Given the fields ‘sentence‘, produce the fields ‘extractions‘.

—

Follow the following format.
Sentence: ${sentence}
Extractions: all mentions of hospital names, hospital units, labs, departments, facilities, consulting services/teams, floor and rooms, medical branches, outside doctors and medical personnel extracted from input sentence. Do not extract anything that is between [** **]. Respond with a single JSON object. JSON Schema: {"properties": {"health_fclt": {"items": {"type": "string"}, "title": "Health Fclt", "type": "array"}}, "required": ["health_fclt"], "title": "SentenceExtraction", "type": "object"}

—

Sentence:
Extractions: "health_fclt": []

Figure 3: The final prompt used by DSPy for extracting the FCLT_PERSONNEL category.