Resource-Efficient Anonymization of Textual Data via Knowledge Distillation from Large Language Models

Tobias Deußer[1,2], Max Hahnbück[1,2], Tobias Uelwer[2], Cong Zhao[1,2],
Christian Bauckhage[1,2], Rafet Sifa[1,2]
[1]University of Bonn, Bonn, Germany,
[2]Fraunhofer IAIS, Sankt Augustin, Germany
Correspondence: tdeusser@uni-bonn.de

Abstract

Protecting personal and sensitive information in textual data is increasingly crucial, especially when leveraging large language models (LLMs) that may pose privacy risks due to their API-based access. We introduce a novel approach and pipeline for anonymizing text across arbitrary domains without the need for manually labeled data or extensive computational resources. Our method employs knowledge distillation from LLMs into smaller encoder-only models via named entity recognition (NER) coupled with regular expressions to create a lightweight model capable of effective anonymization while preserving the semantic and contextual integrity of the data. This reduces computational overhead, enabling deployment on less powerful servers or even personal computing devices. Our findings suggest that knowledge distillation offers a scalable, resource-efficient pathway for anonymization, balancing privacy preservation with model performance and computational efficiency.

1 Introduction

In an increasingly data-driven and AI influenced world, the need to protect personal and sensitive information has become a critical concern across numerous domains, including, but not limited to, healthcare (Zuo et al., 2021; Dimopoulou et al., 2022), law (Csányi et al., 2021; Glaser et al., 2021; Campanile et al., 2022), and finance (Biesner et al., 2022), especially when leveraging large language models (LLMs) (Pan et al., 2020; Wu et al., 2024). Textual data often contains identifiable information that, if exposed, could lead to privacy violations and data breaches. Such privacy concerns might discourage the use of the most powerful LLMs, which are, at the time of writing, often only accessible by external API requests[1]. To tackle this, we introduce an approach and pipeline to anonymize textual data from arbitrary domains. By leveraging knowledge distillation, named entity recognition, and regular expressions, our approach enables the anonymization of sensitive information in a way that reduces the computational overhead while maintaining the semantic integrity of the data. While we evaluate and train on English and German financial documents, our approach can easily be adapted to any new domain or other language. We explore the trade-offs between privacy preservation, model performance, and computational efficiency, demonstrating that knowledge distillation provides a promising pathway for scalable, resource-efficient anonymization.

Traditional named entity recognition methods, though effective for anonymization, often present challenges due to their high computational costs or reliance on manually labeled data. The former is problematic because local computational resources may be limited, and using cloud-based solutions may not be feasible – due to the similar reasons that hinder the use of remote LLMs in the first place. The latter poses a challenge because in many domains where state-of-the-art LLMs could offer the most benefit (and thus, require robust anonymization), labor costs (OECD, 2014) are typically high, making manual data labeling an expensive and time-consuming process.

In this study, we shed light on the training pipeline for our anonymization framework that can take an arbitrary unannotated text corpus and annotation guideline to produce high quality anonymization models, that leverage the knowledge and performance of LLMs like GPT-4 (OpenAI et al., 2024) while being so small, that they can be deployed on significantly less powerful servers or even conventional personal computing devices.

Our contributions can be summarized as follows:

- We demonstrate how a small, lightweight

[1] See the LLM Leaderboard introduced in Chiang et al. (2024) and hosted at lmarena.ai.

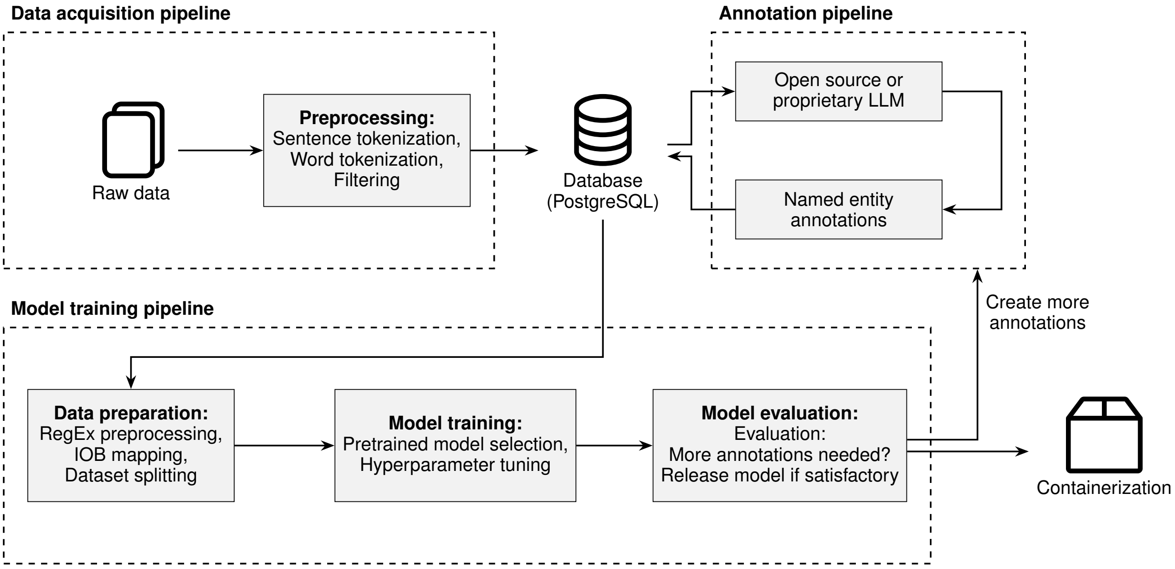
Data acquisition pipeline
Raw data
Preprocessing: Sentence tokenization, Word tokenization, Filtering
Database (PostgreSQL)
Annotation pipeline
Open source or proprietary LLM
Named entity annotations
Create more annotations
Model training pipeline
Data preparation: RegEx preprocessing, IOB mapping, Dataset splitting
Model training: Pretrained model selection, Hyperparameter tuning
Model evaluation: Evaluation: More annotations needed? Release model if satisfactory
Containerization

Figure 1: The different pipelines for our anonymization framework.

model that is trained on text annotated by an LLM can be used to solve the underlying named entity recognition (NER) task of anonymization.

- We build a production-ready anonymization system that can either be deployed locally or as a service to handle API requests.
- We compare the effectiveness of distilling knowledge from different LLMs and benchmark our anonymization system against existing solutions, namely Presidio (Mendels et al., 2018) and GLiNER (Zaratiana et al., 2024).

2 Related Work

Early approaches to automatic anonymization of textual data relied on rule-based named entity recognition models (Sweeney, 1996; Graliński et al., 2009). In contrast, NER with recurrent neural networks (RNNs) was done by Chiu and Nichols (2016). RNNs specifically for the purpose of anonymizing data were proposed by Dernoncourt et al. (2017). Recently, LLMs became a major driver of NER, as seen in Wang et al. (2023), Deußer et al. (2023) or Keloth et al. (2024). Furthermore, Bogdanov et al. (2024) developed their NER-specific foundation model NuNER that is trained on the output of an LLM, whereas Zaratiana et al. (2024) developed GLiNER, an encoder-only model, competing with LLMs for zero-shot NER. Zhou et al. (2024) and Huang et al. (2024) developed a distillation approach for smaller models from LLMs for general NER tasks. Mendels et al. (2018) described an open-source anonymization toolbox called Presidio. For a more in-depth overview on other advances in anonymization techniques, we refer to the work of Lison et al. (2021).

3 Methodology

Our method involves three steps, detailed below:

1. We collect a large number of paragraphs from publicly available documents, which are then pre-processed using traditional methods (Section 3.1).
2. We generate training data by prompting large language models, i.e., GPT-4o and GPT-4o mini, to annotate the pre-processed paragraphs (Section 3.2).
3. We train a NER model on these annotated paragraphs (Section 3.3).

If the performance of step 3 is not satisfactory, we generate more training data by repeating step 2. Results for step 1 and 2 are stored in a PostgreSQL database (Stonebraker and Rowe, 1986), whereas the final model of step 3 gets shipped in the form of a containerized environment after hyperparameter tuning is completed. Figure 1 gives an overview of our approach. In the following subsections we give more details about these three steps.

3.1 Data Acquisition

We start with collecting documents from five different sources in English and German. The documents are then split into sentences and subsequently into words to allow for filtering. In detail, we remove sentences that contain an excessive number of special characters or other textual artifacts, as such features suggest the sentence may not have been parsed correctly or may not actually be a valid sentence. The preprocessed sentences are then stored in a PostgreSQL database to be easily accessible for the following steps.

3.2 Annotation

A central idea of our approach is to employ an LLM to annotate the collected sentences, thereby generating training data to train our lightweight model. We rely on GPT-4o and GPT-4o mini (OpenAI et al., 2024), which we prompted using the provided API. However, we also tested Llama-3 70B (Dubey et al., 2024), Mixtral 8x7B (Jiang et al., 2024), and Mistral Large (Mistral AI Team, 2024), which we found to be inferior to the GPT-4o models.

To find an optimal prompt, we use a comparatively small, annotated dataset composed of around 1,000 paragraphs and iteratively improve our prompt until we achieve satisfactory results. In the final prompt, we provide the model with nine different examples of input sentences and their corresponding expected outputs. For the German datasets, we manually translate the prompt to German and adjust the examples. The annotated paragraphs are stored in the same database as the one they were pulled from. The entity classes that were used to train the model described in the following Section are shown in Table 2. It is important to note that the set of entity classes is flexible and can be defined in advance, allowing customization for any specific use case.

3.3 Model Training

During the model training phase, we first parse the previously created paragraphs and split them into training, validation, and test sets. We then tokenize the text and convert the entity annotations into the Inside-Outside-Beginning (IOB, see Ramshaw and Marcus, 1995) format, so that it can be used in the downstream task. IOB is a tagging scheme used in sequence labeling tasks, where each token in a sentence is tagged as either the beginning (B), inside (I), or outside (O) of a named entity.

The data preparation is followed by the actual training of an *encoder-only* model, e.g., BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2019), with a classification head, i.e., a multilayer perceptron, on top. The encoder choice, depth, and layer size of the classification head, and more general model settings are tuneable hyperparameters in this setup.

During training, we leverage the focal loss (Lin et al., 2017) to allow for a better control of how we can weight recall and precision, which is defined as

$$\mathrm{FL}(p_k) = -\alpha_k(1 - p_k)^\gamma \log(p_k), \quad (1)$$

where α_k is used to balance an entity class k, $\gamma \geq 0$ is the focusing parameter of the modulating factor, and $p_k \in [0, 1]$ is the model's estimated probability of entity class k. We theorize that with this loss we can address the imbalance between the outside and actual entity classes. In an anonymization framework, it is paramount to identify as many entities as feasible without penalizing precision too much, thus focusing on improving recall more than precision. This favors underweighting the outside class, which is overrepresented in anonymization (and many NER) datasets. To achieve this, we assign a smaller weight to α_o compared to all α_e, where e represents any entity class other than the outside class o.

If we find that the performance after training is insufficient, we generate more annotations using the methodology previously described in Section 3.2, followed by repeating the model training step.

3.4 Application Development and Deployment

The model trained in Section 3.3 is combined with rule-based pre- and post-processing. This processing consists of the optional RegEx-based recognition of monetary values, email addresses, IBANs, phone numbers, and websites. IBANs are validated using `schwifty`[2] and only valid IBANs are anonymized.

The anonymization model, i.e., the model trained in Section 3.3 combined with the post-processing discussed above, is exposed as an API via FastAPI[3] and containerized with Docker[4]. We also serve an optional simple frontend with Streamlit[5], which we plan to replace with a more advanced version based on another software stack in the near future.

[2] schwifty.readthedocs.io
[3] fastapi.tiangolo.com
[4] docker.com
[5] streamlit.io

Name	Language	# Paragraphs	# Annotated paragraphs	Reference	URL
Edgar	English	151k	96k	-	sec.gov/search-filings
Financial News Articles	English	3.97M	172k	-	huggingface.co/datasets/ashraq/financial-news-articles
Bundesanzeiger	German	415k	38k	Hillebrand et al. (2024)	bundesanzeiger.de
German News	German	201k	40k	Schabus et al. (2017)	huggingface.co/datasets/community-datasets/gnad10
Tagesschau	German	754k	39k	-	huggingface.co/datasets/bjoernp/tagesschau-2018-2023

Table 1: The datasets and sources we used for training the NER model.

Label	Description	Support en	Support de
<PER>	Person	75,433	28,498
<LOC>	Location	95,538	41,799
<ORG>	Organization	159,434	36,857
<PROD>	Product	20,865	4,603
<DATE>	Date or time	113,876	27,418
<MISC>	Miscellaneous	216,871	91,050

Table 2: Entity classes in our dataset and their support in English (en) and German (de).

4 Experiments

In this section, we describe our experimental protocol, review the data and results, and discuss the key advantages and limitations. All training runs were conducted on a GPU node equipped with eight Nvidia V100 GPUs (each with 32GB of VRAM), an Intel Xeon 6148 CPU, and 1 TB of RAM.

4.1 Data

During the data acquisition step, described in Section 3.1, we collect roughly 5.5 million paragraphs with a focus on the financial domain. From that pool of raw, unannotated paragraphs, we sample 385,657 paragraphs, of which 268,756 are English and 116,901 are German, to annotate with GPT-4o and GPT-4o mini (see Section 3.2). Table 1 gives an overview on each dataset and Table 2 shows all entity classes considered and their respective support in English and German after synthetic annotation. We split our dataset into 80% training data and 10% validation data, which are used for model training and hyperparameter tuning. The remaining 10% was reserved as a hold-out test set, on which we report the results presented in Table 3.

4.2 Results

When working with synthetic data generation, a key question arises: At what point is the amount of data generated sufficient? To address this, Figure 2 illustrates that our validation set performance jumps significantly from zero to approximately 70% after using just about 2% of our English dataset, which is roughly five thousand paragraphs. Beyond this point, each additional paragraph yields diminishing returns and the performance plateaus when approximately 80% of the dataset is utilized.

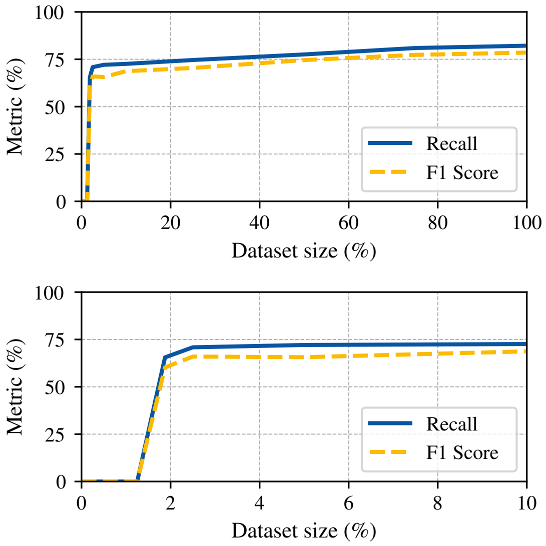

Figure 2: Diminishing Effect of dataset size on model performance. The underlying dataset is the English split of our data, as described in Table 1, totalling 268 thousand paragraphs. Note that both graphs show the same data, only with a differently scaled X-axis.

Table 3 shows the results of our experiments on the test set. We test four different configurations of our *Anonymizer* system, each with a different pretrained encoder backbone and various total model sizes. Our framework can easily outperform the two baselines, Presidio (Mendels et al., 2018) and GLiNER (Zaratiana et al., 2024).

An expected outcome is that larger models tend to exhibit superior performance. Nevertheless, even our smaller models with fewer than 200 million parameters, demonstrate satisfactory performance. Based on these findings, we propose a clear deployment strategy: smaller models are well-suited for on-device deployment due to their efficiency, while larger models, given their superior performance, are better positioned for server-based deployment.

Furthermore, we can observe that leveraging the focal loss (Lin et al., 2017) described in Equation 1 achieves our goal of favoring recall while keeping

(a) **F$_1$** Scores in %

Model	Person	Location	Organization	Product	Date	Miscellaneous	Micro avg.	excl. Misc.
				English split				
Presidio	74.39	66.59	-	-	52.62	-	39.01	48.97
GLiNER	68.40	62.85	60.62	12.17	75.87	03.89	51.20	61.54
Anony N 146M	93.61	90.90	87.88	62.74	86.52	54.63	77.69	87.95
Anony S 163M	93.49	90.34	88.37	59.84	85.40	52.22	77.28	87.58
Anony R 377M	93.51	91.11	88.69	64.89	87.31	55.03	78.07	88.63
Anony L 456M	**94.47**	**91.45**	**89.33**	**66.60**	**87.82**	**55.47**	**78.98**	**89.32**
				German split				
Presidio	06.11	25.94	-	-	41.81	-	11.83	13.94
GLiNER	60.41	65.49	47.65	23.33	68.39	04.35	45.48	56.70
Anony N 146M	87.11	84.48	79.62	55.82	88.82	49.36	69.71	83.60
Anony S 163M	88.05	86.13	80.96	55.58	88.37	47.11	70.20	84.58
Anony R 377M	89.00	86.58	82.38	60.58	89.53	49.24	70.86	85.77
Anony L 456M	**92.62**	**89.84**	**85.69**	**68.19**	**93.57**	**53.50**	**74.43**	**89.33**
				English & German split				
Presidio	30.69	50.34	-	-	51.02	-	29.05	35.62
GLiNER	66.86	63.50	57.29	21.09	74.44	04.03	49.71	56.64
Anony N 146M	91.20	88.95	86.48	62.14	87.29	52.77	75.90	87.24
Anony S 163M	92.68	89.84	87.89	63.11	88.27	54.10	76.82	88.18
Anony R 377M	**92.80**	**90.26**	**88.41**	**64.67**	88.21	54.05	76.62	**88.62**
Anony L 456M	92.69	90.10	88.37	63.21	**88.49**	**55.55**	**77.78**	88.51

(b) **Recall** Scores in %

Model	Person	Location	Organization	Product	Date	Miscellaneous	Micro avg.	excl. Misc.
				English split				
Presidio	78.95	70.62	-	-	67.16	-	30.14	43.97
GLiNER	91.37	78.04	85.74	52.80	76.49	02.45	56.54	81.35
Anony N 146M	95.40	**94.47**	91.95	**65.64**	89.12	54.54	79.58	91.15
Anony S 163M	95.73	93.16	90.91	63.95	88.77	48.49	77.29	90.46
Anony R 377M	95.23	93.58	**92.21**	61.97	89.30	**56.00**	79.88	90.91
Anony L 456M	**96.02**	94.35	91.68	64.84	**89.99**	55.94	**80.45**	**91.39**
				German split				
Presidio	31.61	41.16	-	-	33.36	-	15.46	25.60
GLiNER	86.65	79.52	79.33	61.96	75.32	02.54	49.01	79.49
Anony N 146M	88.54	86.96	84.20	59.96	90.67	**52.07**	72.65	86.41
Anony S 163M	92.12	89.98	83.01	59.00	90.61	46.37	71.32	87.69
Anony R 377M	89.59	87.68	84.17	57.30	90.25	51.34	72.51	86.67
Anony L 456M	**92.83**	**91.45**	**85.54**	**66.48**	**93.76**	50.60	**72.85**	**89.66**
				English & German split				
Presidio	65.39	62.02	-	-	60.29	-	26.38	39.69
GLiNER	89.75	78.87	84.00	53.27	76.42	02.48	54.49	80.71
Anony N 146M	94.82	91.53	91.49	62.72	90.43	52.38	77.69	90.68
Anony S 163M	94.31	92.23	89.75	63.15	89.82	54.82	78.19	89.94
Anony R 377M	94.87	92.70	89.94	65.15	**90.76**	**57.19**	79.29	90.63
Anony L 456M	**95.38**	**93.95**	**92.21**	**68.52**	90.51	54.95	**79.44**	**91.83**

Table 3: Results on the hold-out test set. Anony N, S, R, and L refers to our *Anonymizer* framework, as described in Section 3.3, with different encoder models. The number following the model identifier is the corresponding total model parameter count. Anony S and L feature the GLiNER model variant (and *only* the actual, raw transformer model without the classification head) introduced by Törnquist and Caulk (2024) in the respective small and large size, whereas Anony R represents the setup with a RoBERTa-Large previously finetuned with the OntoNotes dataset (Pradhan et al., 2013) introduced by Ushio and Camacho-Collados (2021) and Anony N has the NuNER-v2.0 model (Bogdanov et al., 2024) as its encoder. Each setup was subjected to hyperparameter tuning on the validation set before being evaluated on the test set. We add results from Presidio (Mendels et al., 2018) and GLiNER (Zaratiana et al., 2024) as a baseline. Note that Presidio only supports anonymizing persons, locations and dates out-of-the-box.

the overall F_1 score high, which is of significant importance when anonymizing data.

4.3 Limitations

The "Miscellaneous" (MISC) category poses a unique challenge due to its highly heterogeneous nature. It serves as a catch-all for tokens that do not fit into other predefined categories, leading to a mix of relevant and irrelevant data, stemming from its definition: "Miscellaneous encompasses any significant information not covered by the other categories that might be used to de-anonymize". This lack of clear boundaries makes it difficult for the model to consistently identify which tokens belong to this class. Although dividing the MISC category into more detailed categories might be possible, some tokens will always resist clear classification. Additionally, classification is subjective, depending on the user's context and model application. Despite these challenges, we have chosen to retain the MISC category in our six-class schema for its balance of manageability and relevance.

This fuzzy nature is illustrated by the following example sentence from the *financial-news-articles* dataset, with annotations below each entity:

> "$\underset{\text{PER}}{\text{Francisco Palmieri}}$, acting $\underset{\text{MISC}}{\text{Assistant Secretary of State}}$ for $\underset{\text{MISC}}{\text{Western Hemisphere Affairs}}$, said the $\underset{\text{LOC}}{\text{Cuban}}$ government was responsible for the security of $\underset{\text{LOC}}{\text{U.S.}}$ diplomatic personnel on the $\underset{\text{MISC}}{\text{island}}$ 'and they have failed to live up to that responsibility.' Asked whether it was possible that the $\underset{\text{ORG}}{\text{Cuban government}}$ would have been unaware of any attacks, he said: 'I find it very difficult to believe that.'"

The entities tagged as MISC illustrate the ambiguous nature of this class, highlighting the difficulty for models to learn this entity class. One could also argue that they may not necessarily require anonymization, as they lack definitive identifying information.

Another limitation of our approach is the actual requirement to train a model. Other approaches incorporating large language models or solutions like GLiNER (Zaratiana et al., 2024) or Presidio (Mendels et al., 2018) are designed to function in a zero-shot environment without any additional training. Nevertheless, such solutions are either computationally intensive, accessible only via an API, and/or lacking in performance (see Table 3).

5 Conclusion

We have introduced a novel text anonymization approach that balances privacy preservation with computational efficiency by distilling knowledge from large language models into smaller, encoder-only models using named entity recognition and rule-based algorithms. Our lightweight system operates without the need for manually labeled data or extensive computational resources and is suitable for deployment on less powerful servers or personal computing devices. It can easily be adapted to any domain and is currently deployed for the anonymization of financial documents and texts.

Our experiments demonstrate that our method outperforms existing solutions like GLiNER (Zaratiana et al., 2024) or Presidio (Mendels et al., 2018), achieving higher F_1 scores and, more importantly, higher recall overall and in all entity classes. Even our smaller models with fewer than 200 million parameters showed still satisfactory and superior performance, indicating their practicality for on-device deployment where computational resources are limited and anonymization is paramount.

In conclusion, our findings suggest that knowledge distillation offers a scalable, customizable, and resource-efficient pathway for text anonymization. By harnessing the capabilities of LLMs, our approach holds significant promise for enhancing privacy preservation in textual data across various domains. Furthermore, with the continuous development of new LLMs, we can enhance our framework by updating the teacher, i.e., the LLM, of our NER models.

Future work could shift the focus from the financial domain onto different languages or domains, like social media, healthcare, or law, which require a different set of entities, but can likely be solved with the same framework as introduced here. Additionally, one could test if we see a performance degradation after replacing the raw, real-world data (see Section 3.1 and 4.1) with synthetic data generated by a LLM, as seen in Watson et al. (2024) for example. Another interesting venue is exploring the effect anonymization has on the performance of LLM-powered downstream tasks like contradiction detection (Deußer et al., 2023), factual consistency evaluation (Gekhman et al., 2023), or automated regulatory compliance verification (Berger et al., 2023) or on the direct, actual performance of LLMs, evaluated by benchmarks like the Open LLM Leaderboard (Fourrier et al., 2024).

6 Ethical Considerations

Our work focuses on enhancing privacy by anonymizing sensitive information in textual data across various domains. While our approach aims to protect personal data and mitigate the risk of privacy breaches, it is important to acknowledge that no anonymization method, even manual anonymization, can provide a 100% guarantee of complete confidentiality, and our method is no exception, as shown in Table 3.

Additionally, if one applies the same approach as the one in our model, the complete opposite is possible: The identification of sensitive information and entities from arbitrary chunks of text, leading to easier retrieval of said personal information, which is an inherent risk of all named entity recognition models.

Acknowledgments

This research has been partially funded by the Federal Ministry of Education and Research of Germany and the state of North-Rhine Westphalia as part of the Lamarr-Institute for Machine Learning and Artificial Intelligence.

References

Armin Berger, Lars Hillebrand, David Leonhard, Tobias Deußer, Thiago Bell Felix De Oliveira, Tim Dilmaghani, Mohamed Khaled, Bernd Kliem, Rudiger Loitz, Christian Bauckhage, et al. 2023. Towards automated regulatory compliance verification in financial auditing with large language models. In *Proc. BigData*, pages 4626–4635. IEEE.

David Biesner, Rajkumar Ramamurthy, Robin Stenzel, Max Lübbering, Lars Hillebrand, Anna Ladi, Maren Pielka, Rüdiger Loitz, Christian Bauckhage, and Rafet Sifa. 2022. Anonymization of german financial documents using neural network-based language models with contextual word representations. *International Journal of Data Science and Analytics*, pages 1–11.

Sergei Bogdanov, Alexandre Constantin, Timothée Bernard, Benoit Crabbé, and Etienne Bernard. 2024. NuNER: Entity recognition encoder pre-training via LLM-annotated data. *Preprint*, arXiv:2402.15343.

Lelio Campanile, Maria Stella de Biase, Stefano Marrone, Fiammetta Marulli, Mariapia Raimondo, and Laura Verde. 2022. Sensitive information detection adopting named entity recognition: A proposed methodology. In *Proc. ICCSA Workshops*, pages 377–388.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating LLMs by human preference. *Preprint*, arXiv:2403.04132.

Jason PC Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the association for computational linguistics*, 4:357–370.

Gergely Márk Csányi, Dániel Nagy, Renátó Vági, János Pál Vadász, and Tamás Orosz. 2021. Challenges and open problems of legal document anonymization. *Symmetry*, 13(8).

Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. 2017. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, 24(3):596–606.

Tobias Deußer, David Leonhard, Lars Hillebrand, Armin Berger, Mohamed Khaled, Sarah Heiden, Tim Dilmaghani, Bernd Kliem, Rüdiger Loitz, Christian Bauckhage, et al. 2023. Uncovering inconsistencies and contradictions in financial reports using large language models. In *Proc. BigData*, pages 2814–2822. IEEE.

Tobias Deußer, Lars Hillebrand, Christian Bauckhage, and Rafet Sifa. 2023. Informed named entity recognition decoding for generative language models. *Preprint*, arXiv:2308.07791.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL-HLT*, pages 4171–4186.

Stella Dimopoulou, Chrysostomos Symvoulidis, Konstantinos Koutsoukos, Athanasios Kiourtis, Argyro Mavrogiorgou, and Dimosthenis Kyriazis. 2022. Mobile anonymization and pseudonymization of structured health data for research. In *Proc. MobiSecServ*, pages 1–6.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Clémentine Fourrier, Nathan Habib, Alina Lozovskaya, Konrad Szafer, and Thomas Wolf. 2024. Open LLM leaderboard v2. `https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard`.

Zorik Gekhman, Jonathan Herzig, Roee Aharoni, Chen Elkind, and Idan Szpektor. 2023. TrueTeacher: Learning factual consistency evaluation with large language models. In *Proc. EMNLP*, pages 2053–2070.

Ingo Glaser, Tom Schamberger, and Florian Matthes. 2021. Anonymization of german legal court rulings. In *Proc. ICAIL*, page 205–209.

Filip Graliński, Krzysztof Jassem, Michał Marcińczuk, and Paweł Wawrzyniak. 2009. Named entity recognition in machine anonymization. *Recent Advances in Intelligent Information Systems*, pages 247–260.

Lars Hillebrand, Prabhupad Pradhan, Christian Bauckhage, and Rafet Sifa. 2024. Pointer-guided pre-training: Infusing large language models with paragraph-level contextual awareness. In *Proc. ECML-PKDD*, pages 386–402. Springer.

Yining Huang, Keke Tang, and Meilian Chen. 2024. Leveraging large language models for enhanced nlp task performance through knowledge distillation and optimized training strategies. *Preprint*, arXiv:2402.09282.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts.

Vipina K Keloth, Yan Hu, Qianqian Xie, Xueqing Peng, Yan Wang, Andrew Zheng, Melih Selek, Kalpana Raja, Chih Hsuan Wei, Qiao Jin, et al. 2024. Advancing entity recognition in biomedicine via instruction tuning of large language models. *Bioinformatics*, 40(4).

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proc. ICCV*, pages 2999–3007.

Pierre Lison, Ildikó Pilán, David Sánchez, Montserrat Batet, and Lilja Øvrelid. 2021. Anonymisation models for text data: State of the art, challenges and future directions. In *Proc. ACL-IJCNLP*, pages 4188–4203.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *Preprint*, arXiv:1907.11692.

Omri Mendels, Coby Peled, Nava Vaisman Levy, Sharon Hart, Tomer Rosenthal, Limor Lahiani, et al. 2018. Microsoft Presidio: Context aware, pluggable and customizable pii anonymization service for text and images.

Mistral AI Team. 2024. Mistral large. Accessed: 2024-09-19.

OECD. 2014. Productivity and unit labour cost by industry, isic rev. 4.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, et al. 2024. GPT-4 technical report. *Preprint*, arXiv:2303.08774.

Xudong Pan, Mi Zhang, Shouling Ji, and Min Yang. 2020. Privacy risks of general-purpose language models. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 1314–1331. IEEE.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using OntoNotes. In *Proc. CoNLL*, pages 143–152, Sofia, Bulgaria. Association for Computational Linguistics.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Dietmar Schabus, Marcin Skowron, and Martin Trapp. 2017. One million posts: A data set of german online discussions. In *Proc. SIGIR*, pages 1241–1244.

Michael Stonebraker and Lawrence A Rowe. 1986. The design of postgres. *ACM Sigmod Record*, 15(2):340–355.

Latanya Sweeney. 1996. Replacing personally-identifying information in medical records, the scrub system. In *Proc. AMIA*, page 333.

Elin Törnquist and Robert Alexander Caulk. 2024. Curating grounded synthetic data with global perspectives for equitable ai. *Preprint*, arXiv:2406.10258.

Asahi Ushio and Jose Camacho-Collados. 2021. T-NER: An all-round python library for transformer-based named entity recognition. In *Proc. EACL*, pages 53–62.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023. GPT-NER: Named entity recognition via large language models. *Preprint*, arXiv:2304.10428.

Alex Watson, Yev Meyer, Maarten Van Segbroeck, Matthew Grossman, Sami Torbey, Piotr Mlocek, and Johnny Greco. 2024. Synthetic-PII-Financial-Documents-North-America: A synthetic dataset for training language models to label and detect pii in domain specific formats.

Xiaodong Wu, Ran Duan, and Jianbing Ni. 2024. Unveiling security, privacy, and ethical concerns of chatGPT. *Journal of Information and Intelligence*, 2(2):102–115.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. GLiNER: Generalist model for named entity recognition using bidirectional transformer. In *Proc. NAACL*, pages 5364–5376.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2024. UniversalNER: Targeted distillation from large language models for open named entity recognition. In *Proc. ICLR*.

Zheming Zuo, Matthew Watson, David Budgen, Robert Hall, Chris Kennelly, and Noura Al Moubayed. 2021. Data anonymization for pervasive health care: systematic literature mapping study. *JMIR medical informatics*, 9(10).