

# Meddies-PII: A Multilingual Framework for Personally Identifiable Information Extraction in Clinical De-identification

Le Linh Uyen<sup>1</sup>, Christian Hoang<sup>1</sup>, and Huy Hoang Ha<sup>1,†</sup>

<sup>1</sup>Meddies AI, Vietnam

<sup>†</sup>Corresponding author: hoangha@meddies.ai

Clinical de-identification depends on accurately identifying personally identifiable information (PII). However, manually annotated datasets are costly to build, and existing synthetic alternatives often provide limited generation details or use relatively simple synthesis strategies. We introduce Meddies-PII-Dataset, which contains one million synthetic clinical documents across seventeen languages and nine PII labels. Attribute-conditioned prompts generate the documents, and thirteen deterministic gates check their structural and annotation validity. To assess the dataset’s utility, we train Meddies-PII-Model, a BIOES token classifier. We compare it with existing PII extraction systems using exact-match entity-level F1. Meddies-PII-Model achieves the highest score among the evaluated systems on all reported benchmarks, with a mean F1 of 0.827 across fifteen external benchmarks compared with 0.658 for the strongest baseline. Upon acceptance, we will publicly release the dataset, benchmark, model, generation framework, and evaluation code to support research on multilingual clinical de-identification.

Keywords: Natural Language Processing; Named Entity Recognition; Personally Identifiable Information; Synthetic Data; Healthcare

## 1. Introduction

Named Entity Recognition (NER) is a fundamental task in natural language processing. In privacy-sensitive domains such as healthcare, however, the primary objective is identifying personally identifiable information (PII), a specialized subset of named entities whose disclosure may compromise patient privacy [1]. Although PII extraction is commonly formulated as an NER task, it differs from general-purpose NER in both its ontology and application objectives. Consequently, data-driven PII systems require dedicated datasets rather than relying solely on conventional NER corpora. Existing PII datasets are predominantly manually annotated, making them expensive to construct and limiting their scale, language coverage, and domain diversity [2–4]. Although several synthetic PII datasets have recently been proposed, most either do not disclose their generation pipeline or rely on relatively simple prompting strategies [5–7].

We propose a multilingual synthetic data generation framework that combines attribute-conditioned prompting with deterministic validation to construct Meddies-PII-Dataset. The framework controls language, document type, text format, scenario, ontology constraints, and challenging edge cases to generate diverse and structurally complex documents. Generated samples that violate the declared requirements are repaired or rejected before inclusion in the dataset. We further train a token-classification model to evaluate the utility of the resulting data. We evaluate the model on fif-

teen external benchmarks and the Meddies-PII Benchmark. Our benchmark combines standard and challenging examples under the Meddies-PII ontology to assess performance across different levels of difficulty.

This paper makes three contributions:

- (1) We introduce Meddies-PII-Dataset, a synthetic clinical PII dataset containing one million documents across seventeen languages and annotated with a nine-label PII ontology.
- (2) We present a controlled multilingual synthesis framework that combines multiple document and generation attributes with deterministic structural and annotation validation.
- (3) We introduce Meddies-PII-Model, a BIOES token classifier, and show that it achieves the highest performance on all evaluated external benchmarks and our benchmark.

## 2. Related Works

### 2.1. PII Datasets

**Manually annotated datasets** Manually annotated PII datasets support supervised training and evaluation, but expert annotation makes them costly to construct and difficult to scale [8]. Early clinical de-identification resources focused primarily on English medical records and supported influential shared tasks and system evaluations [2–4]. Non-English clinical resources include corpora for French clinical notes and the German GRASCCO corpus [9, 10]. Sensitive-entity datasets have also been developed for Spanish legal text, showing progress beyond the clinical domain [11]. More recent multilingual resources broaden this coverage, but manually annotated datasets still span relatively few languages and domains [12].

**Synthetic datasets** Synthetic generation provides a scalable alternative to manually annotated datasets, and several large PII datasets have recently been released [5–7]. However, their generation procedures are not always described in enough detail to reproduce the controls used for document diversity, label coverage, and difficult examples. SPY [1] documents its workflow more clearly, including iterative placeholder insertion, synthetic entity swapping, and the addition of unrelated entities. Meddies-PII-Dataset addresses this reproducibility gap through an independently designed multilingual pipeline that explicitly controls document and generation attributes and applies deterministic validation before accepting samples.

### 2.2. PII Models

**NER Techniques** NER systems identify text spans and assign each span an entity label. Many neural approaches formulate this task as sequence labeling, predicting a label for each token and optionally using a Conditional Random Field (CRF) to model dependencies between adjacent labels [13]. Transformer encoders such as BERT improve these predictions by representing each token in its surrounding context [14]. Span-based methods instead score candidate spans directly, allowing the model to predict entity boundaries and labels together [15]. More recent generalist models, including GLiNER and GLiNER2, use schema-driven interfaces to recognize entity types beyond a single fixed taxonomy [16, 17].

**PII Techniques** PII extraction specializes NER by mapping detected spans to privacy-related labels used for de-identification. Generalist systems such as GLiNER2 can treat these labels as a requested extraction schema, while dedicated systems such as OpenMed-PII and OpenAI Privacy Filter are designed specifically for PII detection [17–19]. OpenAI Privacy Filter uses bidirectional token classification and span decoding over a fixed taxonomy of privacy labels [19]. This formulation combines contextual token representations with explicit entity-boundary prediction. Following this approach, we train Meddies-PII-Model on Meddies-PII-Dataset using BIOES tagging and our nine-label PII ontology.

### 3. Meddies-PII-Dataset

#### 3.1. Overview

Manually annotated PII datasets provide expert-reviewed labels for training and evaluation, but their cost limits their scale, language coverage, and domain diversity [2, 8]. Synthetic generation offers greater scalability, although insufficiently documented controls and validation procedures can make existing datasets difficult to reproduce [5–7]. Meddies-PII-Dataset addresses this trade-off through a multilingual generation framework that makes its generation attributes and deterministic validation process explicit. Figure 1 provides an overview of the complete pipeline.

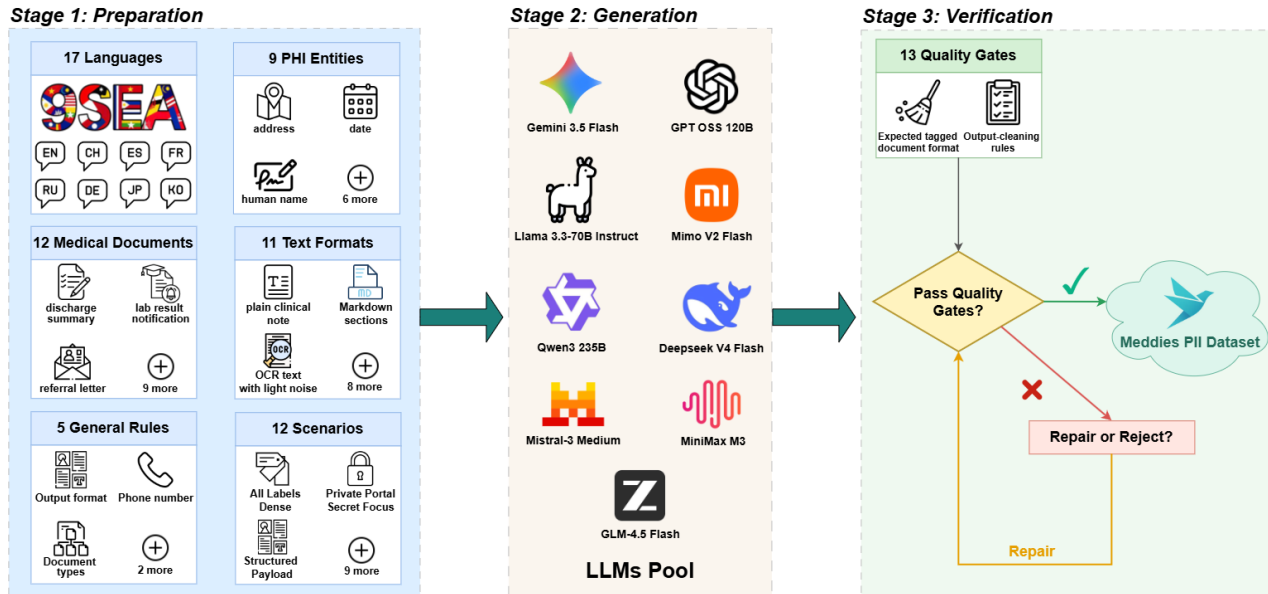


Fig. 1. Synthetic data generation and verification pipeline to create Meddies-PII-Dataset.

**Dataset composition** Meddies-PII-Dataset contains one million synthetic documents across seventeen languages: English, Vietnamese, Chinese, Japanese, Lao, Tamil, Thai, Burmese, Filipino, Russian, Spanish, Malay, Indonesian, Portuguese, French, Korean, and German. After generating the initial collection, we reviewed its samples and identified opportunities to improve data quality. We then refined the generation process and created a second collection of higher-quality samples. The final dataset mixes selected samples from both collections to reach one million documents while preserv-

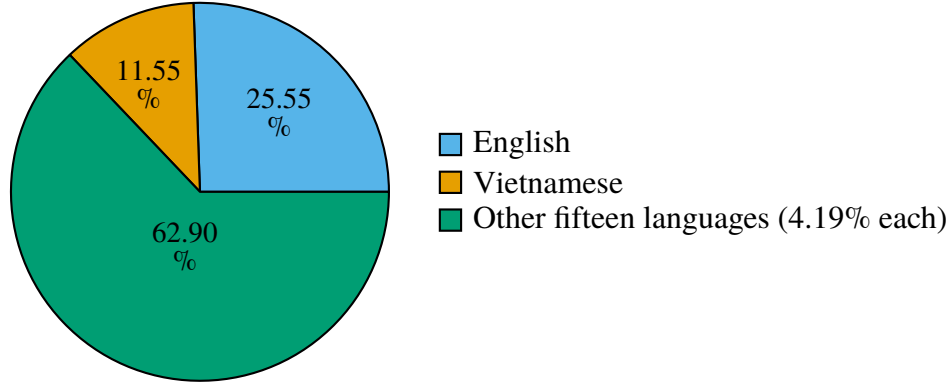


Fig. 2. Language distribution of Meddies-PII-Dataset. The other fifteen languages each contain 41,934 samples (4.19%).

ing the target language distribution and broader coverage. This design is consistent with sample-wise pretraining data mixing research that jointly considers quality and diversity when determining the resulting data distribution [20]. The language selection emphasizes Southeast Asian languages, several of which have limited representation in existing PII datasets. Because widely spoken international languages are also used across Southeast Asian countries, we extend the dataset with prominent languages beyond the region, yielding coverage across different scripts, language families, and resource levels. Figure 2 summarizes the language distribution.

**PII labels** We define a nine-label PII ontology informed by the identifiers and distinguishing attributes described in ISO/IEC 29100:2024 [21]:

- **address**: street addresses, postal codes, coordinates, and care locations;
- **company\_name**: hospitals, clinics, departments, companies, and other organizations;
- **date**: birth dates, admission and discharge dates, and other specific datetimes linked to a person or encounter;
- **email\_address**: email addresses;
- **human\_name**: patient, clinician, and other person names;
- **id\_number**: medical record numbers, government identifiers, passport and licence numbers, IP addresses, and account numbers;
- **phone\_number**: telephone and fax numbers;
- **private\_url**: access-bearing patient-portal, result, signed, or private-record URLs;
- **secret**: passwords, API keys, authentication or session tokens, cookies, personal identification numbers, and one-time passwords.

**Meddies-PII Benchmark** Our benchmark contains 5,100 validated documents distributed evenly across the seventeen languages, with 300 documents per language. For each language, 100 documents assess PII extraction under standard conditions, while 200 emphasize adversarial surface forms. Each adversarial document combines four to five of thirteen perturbation techniques. These include spoken digits and email symbols, phonetic spelling, unusual spacing, entity-internal line breaks, truncation, partial masking, OCR-like substitutions, Unicode confusables, and concatenated values. These per-

turbations preserve the annotated PII spans while testing robustness to noise commonly encountered in speech transcripts, scanned records, messages, and administrative documents.

### 3.2. Dataset Creation

As shown in Figure 1, dataset creation proceeds through three stages: Preparation, Generation, and Verification. For each sample, the framework prepares a controlled prompt, sends it to one model from the generation pool, and validates the resulting document before accepting it into the dataset.

**Stage 1: Preparation** Each generation prompt combines six attribute groups that control the content and structure of a sample. Medical documents determine the workflow and content, while text formats determine how that content is expressed. General rules maintain annotation consistency, and scenarios control the operational context and difficulty. Together, these groups capture complementary sources of variation in clinical PII data. Randomly combining attribute values produces varied prompts at scale, an approach supported by evidence that attributed prompts improve synthetic-data diversity and downstream performance while reducing query cost [22]. We refined the final set of options through project discussions to balance realistic clinical and administrative use cases with broad label and structural coverage. Figure 3 shows the prompt template.

- (1) Languages: Specifies the target language of the generated document.
- (2) PII labels: Nine labels specify the PII categories and required label coverage for each document.
- (3) Medical documents: Represents PII in common clinical and administrative workflows. The twelve documents are discharge summary; outpatient note; lab-result notification; patient portal access log; insurance claim note; referral letter; pharmacy refill request; appointment reminder; telehealth triage note; medical bill or invoice; FHIR-like Patient and Encounter resource; and HL7-style admission or update message.
- (4) Text formats: Captures how equivalent information changes across prose, structured records, logs, transcripts, and noisy text. The eleven formats are plain clinical note; Markdown sections; messy nurse or administrative note; table-like rows; JSON-like object; FHIR-like JSON; HL7-like pipe-delimited text; patient portal audit log; voice-scribe transcript; lightly degraded OCR text; and clinical dialogue transcript.
- (5) General rules: Maintains consistent generation and annotation across combinations of attributes. The five rule groups govern the output format, phone-number construction, private URLs, secrets, and compatibility between medical documents and text formats.
- (6) Scenarios: Places PII within realistic operational contexts while controlling label density and difficulty. The twelve scenarios are All Labels Dense; Private Portal Secret Focus; Adversarial Obfuscation; Public Negative Contrast; Structured Payload; General Consumer Admin Support; HR, Employer, School, Insurance, or Telecom Admin; Code, Log, Security, or Account Recovery; Adversarial Formatting; Structured General Payload; One-Hop Deferred Clue; and Public Negative URL and Data Contrast.

```

user_prompt = f"""
Generate one synthetic {language} {document_type} in {text_format} format.

Scenario: {scenario_name}
{scenario_description}

Hard constraints:
1. Use at least {min_spans} tagged PII spans.
2. Use at least {min_unique_labels} unique Meddies labels.
3. Required labels for this sample: {required_labels}.
etc.

Required-label coverage before you answer: {"", ".join(f"{label}=present" for label in resolved_required_labels) or "none
"}.
Minimum coverage before you answer: at least {min_spans} bracketed tags and at least {min_unique_labels} unique labels.

Remember: label syntax must be `[value]<label>` and labels must be one of: {"", ".join(f"<{label}>" for label in
PII_LABELS)}.
"""

```

Fig. 3. User-prompt template for Meddies-PII-Dataset generation. Braced fields are populated from the sampled generation attributes and validation requirements.

**Hard constraints** The hard-constraint block converts the sampled attributes into explicit generation requirements. It specifies the minimum number of tagged spans and unique labels, lists labels that must appear, and controls when private URLs, secrets, negative examples, and adversarial perturbations are required. It also enforces the exact `[value]<label>` syntax, complete entity boundaries, document-only output, and a final coverage self-check. These constraints guide the model toward drafts that can be checked by the deterministic gates in Stage 3, although a draft is accepted only after passing verification.

**Stage 2: Generation** The generation pool contains nine large language models (LLMs) from different model families. These are Gemini 3.5 Flash [23], GLM-4.5-Flash [24], Llama 3.3 70B Instruct [25], MiMo-V2-Flash [26], Mistral Medium 3 [27], gpt-oss-120b [28], DeepSeek-V4-Flash [29], Qwen3-235B-A22B [30], and MiniMax M3 [31]. For each sample, the framework sends the assembled prompt to one selected LLM from this pool. Distributing generation across model families reduces the pipeline’s dependence on the behavior of a single generator.

**Stage 3: Verification** Each generated draft enters deterministic verification. Verification uses regular expressions, a deterministic character-level parser, rule-based checks, and SHA-256 hashing. Regular expressions check annotation syntax and language-specific Unicode ranges. The parser reconstructs the raw document and verifies span offsets, while rule-based checks validate phone numbers. SHA-256 hashes detect duplicate documents. Together, thirteen gates check label coverage, span count, document length, offsets, phone-number structure, native-script presence, and duplication. Drafts that pass are accepted, while failed drafts enter the repair workflow. API failures and empty content are rejected because they contain no recoverable document, while unknown scenarios are rejected because the metadata needed to interpret and validate the document is unavailable. For the remaining drafts, conservative regex repair fixes only explicit labels with unambiguous value boundaries, such as misplaced brackets, spacing around tags, or unwrapped structured-field values. Repaired drafts are then revalidated and accepted only if they pass every gate and are unique. Drafts with missing

or unknown labels, too few spans, invalid lengths or offsets, templated phone numbers, or missing native script are rejected because correcting them would require inventing content, labels.

## 4. Evaluation

### 4.1. Evaluation Settings

**Evaluation Protocol** We compare Meddies-PII-Model with four publicly available PII extraction systems: OpenAI Privacy Filter [19], GLiNER2 [17], SuperClinical-Large-434M-v1 [18], and LFM2.5-Encoder-350M-PII-Detector [32]. The external evaluation includes twelve language subsets from OpenPII 1.5M [5], together with CredData [33], Gretel PII Masking [6], and Nemotron-PII [7]. We additionally evaluate every system on the Meddies-PII Benchmark. Labels produced by each system are mapped to the shared PII ontology before scoring. We report exact-match entity-level micro-F1 by pooling entity counts across documents. A predicted entity is a true positive only when both its PII label and character boundaries exactly match a reference entity. Predictions with an incorrect label or boundary are counted as false positives, while unmatched reference entities are counted as false negatives. Using the pooled counts, the metric is computed as

$$P = \frac{TP}{TP + FP}, \quad R = \frac{TP}{TP + FN}, \quad F_1 = \frac{2PR}{P + R} = \frac{2TP}{2TP + FP + FN}. \quad (1)$$

**Model Development** Our initial model development treated PII extraction as autoregressive structured generation. We first applied supervised fine-tuning (SFT) and subsequently reinforcement learning (RL), but sequential decoding made inference slow and the resulting extraction quality was not competitive. We therefore adopted the one-pass token-classification approach used by OpenAI Privacy Filter [19]. This approach classifies each token using bidirectional document context instead of generating entities token by token. BIOES marks tokens as Beginning, Inside, End, Single-token, or Outside (O) relative to an entity. Meddies-PII-Model fine-tunes LFM2.5-350M with a classifier containing 37 output classes: O and four states for each of the nine PII labels. A linear classification head predicts a BIOES state for each token. Constrained Viterbi decoding then converts the predicted sequence into character-level entity spans while enforcing valid BIOES transitions [34].

### 4.2. Results

**Overall comparison** Table 1 reports exact-match entity-level micro-F1 for the evaluated systems. Meddies-PII-Model ranks first on every benchmark shown. Across the fifteen external benchmarks, it achieves an unweighted mean F1 of 0.827, exceeding OpenAI Privacy Filter, the strongest baseline by external mean, by 0.169. When the Meddies-PII Benchmark is included, the overall means are 0.833 and 0.644, respectively. The consistent advantage across both external and in-domain evaluations suggests that the improvement is not confined to the distribution used to develop Meddies-PII-Model.

**Multilingual performance** On the twelve OpenPII 1.5M language benchmarks, Meddies-PII-Model scores between 0.739 and 0.880 and ranks first in every language. Its largest margins over the strongest competing score occur for Vietnamese (0.880 versus 0.634), Chinese (0.816 versus 0.572), and Korean (0.739 versus 0.554). This pattern indicates that the external mean is not driven only by English and that the model transfers effectively across several languages and scripts. However, the lower ab-

Table 1. Exact-match entity-level micro-F1 scores. The best result for each benchmark is shown in bold.

| Benchmark             | Meddies PII Model | OpenAI Privacy Filter | GLiNER2 | SuperClinical Large 434M | LFM2.5 Encoder 350M PII Detector |
|-----------------------|-------------------|-----------------------|---------|--------------------------|----------------------------------|
| External evaluation   |                   |                       |         |                          |                                  |
| OpenPII 1.5M (de)     | 0.827             | 0.716                 | 0.663   | 0.620                    | 0.610                            |
| OpenPII 1.5M (en)     | 0.878             | 0.745                 | 0.687   | 0.656                    | 0.568                            |
| OpenPII 1.5M (es)     | 0.870             | 0.794                 | 0.732   | 0.683                    | 0.655                            |
| OpenPII 1.5M (fil)    | 0.848             | 0.667                 | 0.690   | 0.584                    | 0.600                            |
| OpenPII 1.5M (fr)     | 0.844             | 0.738                 | 0.674   | 0.604                    | 0.611                            |
| OpenPII 1.5M (id)     | 0.848             | 0.664                 | 0.688   | 0.566                    | 0.627                            |
| OpenPII 1.5M (ja)     | 0.750             | 0.549                 | 0.366   | 0.493                    | 0.479                            |
| OpenPII 1.5M (ko)     | 0.739             | 0.554                 | 0.322   | 0.545                    | 0.460                            |
| OpenPII 1.5M (ms)     | 0.839             | 0.640                 | 0.692   | 0.547                    | 0.587                            |
| OpenPII 1.5M (pt)     | 0.855             | 0.756                 | 0.692   | 0.687                    | 0.651                            |
| OpenPII 1.5M (vi)     | 0.880             | 0.634                 | 0.628   | 0.499                    | 0.615                            |
| OpenPII 1.5M (zh)     | 0.816             | 0.572                 | 0.453   | 0.428                    | 0.548                            |
| CredData (en)         | 0.590             | 0.370                 | 0.357   | 0.150                    | 0.263                            |
| Gretel (en)           | 0.872             | 0.697                 | 0.588   | 0.715                    | 0.729                            |
| Nemotron (en)         | 0.956             | 0.778                 | 0.692   | 0.878                    | 0.574                            |
| External mean         | 0.827             | 0.658                 | 0.595   | 0.577                    | 0.572                            |
| Meddies-PII Benchmark | 0.878             | 0.540                 | 0.476   | 0.244                    | 0.269                            |
| Overall mean          | 0.833             | 0.644                 | 0.581   | 0.538                    | 0.536                            |

solute scores for Korean and Japanese show that multilingual performance is not uniform, leaving room for language-specific analysis and improvement.

**Cross-dataset performance** Meddies-PII-Model achieves F1 scores of 0.590 on CredData, 0.872 on Gretel, and 0.956 on Nemotron. The strongest baseline scores on the same benchmarks are 0.370, 0.729, and 0.878, giving margins of 0.220, 0.143, and 0.078, respectively. Positive margins across all three datasets provide evidence of transfer beyond a single annotation source. CredData remains substantially more difficult for every system, indicating that differences in domain, document structure, or annotation policy still present a meaningful challenge.

**Meddies-PII Benchmark** The Meddies-PII Benchmark combines standard examples with examples emphasizing adversarial surface forms. On this benchmark, Meddies-PII-Model achieves an F1 of 0.878, compared with 0.540 for the strongest baseline. The 0.338 margin shows that the model handles the ontology and perturbations represented in its target setting more effectively than the compared systems. Because this is an in-domain evaluation, we report and interpret it separately from the external mean rather than treating it as direct evidence of generalization.

Taken together, the results show consistent gains across languages and datasets, while the lower scores on CredData, Korean, and Japanese identify important directions for further evaluation.

**Data and code availability** Upon acceptance, we will publicly release Meddies-PII-Dataset, the Meddies-PII Benchmark, Meddies-PII-Model, the generation framework, and the evaluation code.

## 5. Limitations

Meddies-PII-Dataset is entirely synthetic and may not capture all linguistic variation, documentation practices, and annotation ambiguity found in real-world clinical records. Its deterministic gates enforce the declared structural and annotation requirements, but they do not directly measure the semantic naturalness of every generated document. In addition, the present evaluation focuses on synthetic PII benchmarks and Meddies-PII Benchmark data; evaluation on appropriately governed real clinical records remains an important direction for future work.

## 6. Conclusion

We presented Meddies-PII-Dataset, a synthetic clinical PII dataset containing one million documents across seventeen languages and nine PII labels. After reviewing the initial generation, we refined the process, generated a second collection of higher-quality samples, and mixed both collections to preserve the target distribution and broader coverage. The improved process controls multiple document and scenario attributes and selects one generator per sample from a pool of nine models. Thirteen deterministic gates then validate each sample before acceptance. Meddies-PII-Model achieves the highest performance among the evaluated systems across all reported external benchmarks and the Meddies-PII Benchmark, demonstrating the downstream utility of the generated data. By making the dataset, benchmark, model, generation framework, and evaluation code publicly available upon acceptance, we aim to support further research on multilingual clinical de-identification and synthetic PII generation.

## Acknowledgments

We thank Modal for providing the compute resources used to run our experiments.

## References

- [1] M. Savkin, T. Ionov and V. Konovalov, SPY: Enhancing privacy with synthetic PII detection dataset, in Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop), eds. A. Ebrahimi, S. Haider, E. Liu, S. Haider, M. L. Pacheco and S. Wein (Association for Computational Linguistics, Albuquerque, USA, April 2025).
- [2] I. Neamatullah, M. M. Douglass, L.-W. H. Lehman, A. Reisner, M. Villarroel, W. J. Long, P. Szolovits, G. B. Moody, R. G. Mark and G. D. Clifford, Automated de-identification of free-text medical records, BMC medical informatics and decision making 8, p. 32 (2008).
- [3] Ö. Uzuner, Y. Luo and P. Szolovits, Evaluating the state-of-the-art in automatic de-identification, Journal of the American Medical Informatics Association 14, 550 (2007).
- [4] 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 58, S11 (2015).
- [5] Ai4Privacy, OpenPII 1.5M: Multilingual PII Masking Dataset with Asia Pacific Extension Hugging Face dataset, (2026), Accessed: 2026-07-25.
- [6] G. AI, Gliner models for pii detection through fine-tuning on gretel-generated synthetic documents (10 2024).
- [7] A. Steier, A. Manoel, A. Haushalter and M. V. Segbroeck, Nemotron-pii: Synthesized data for privacy-preserving ai (2025).

- [8] L. Deleger, T. Lingren, Y. Ni, M. Kaiser, L. Stoutenborough, K. Marsolo, M. Kouril, K. Molnar and I. Solti, Preparing an annotated gold standard corpus to share with extramural investigators for de-identification research, *Journal of biomedical informatics* 50, 173 (2014).
- [9] C. Grouin and A. Névéol, De-identification of clinical notes in french: towards a protocol for reference corpus development, *Journal of biomedical informatics* 50, 151 (2014).
- [10] L. Modersohn, S. Schulz, C. Lohr and U. Hahn, Grasco—the first publicly shareable, multiply-alienated german clinical text corpus, in *German Medical Data Sciences 2022–Future Medicine: More Precise, More Integrative, More Sustainable!*, (IOS Press, 2022) pp. 66–72.
- [11] O. de Gibert, A. García-Pablos, M. Cuadros and M. Melero, Spanish datasets for sensitive entity detection in the legal domain, in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, eds. N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk and S. Piperidis (European Language Resources Association, Marseille, France, June 2022).
- [12] I. Baroud, C. Otto, V. Czehmann, C. Hovhannisyan, L. Raithel, S. Möller and R. Roller, Multigrasco: A multilingual anonymization benchmark with annotations of personal identifiers, *arXiv preprint arXiv:2603.08879* (2026).
- [13] F. Souza, R. Nogueira and R. Lotufo, Portuguese named entity recognition using BERT-CRF *arXiv preprint arXiv:1909.10649*, (2019).
- [14] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, eds. J. Burstein, C. Doran and T. Solorio (Association for Computational Linguistics, Minneapolis, Minnesota, June 2019).
- [15] M. Eberts and A. Ulges, Span-based joint entity and relation extraction with transformer pre-training, *arXiv preprint arXiv:1909.07755* (2019).
- [16] U. Zaratiana, N. Tomeh, P. Holat and T. Charnois, Gliner: Generalist model for named entity recognition using bidirectional transformer, *arXiv preprint arXiv:2311.08526* (2023).
- [17] U. Zaratiana, G. Pasternak, O. Boyd, G. Hurn-Maloney and A. Lewis, Gliner2: An efficient multi-task information extraction system with schema-driven interface (2025).
- [18] O. Science, Openmed-pii-superclinical-large-434m-v1: Pii detection model (2026).
- [19] OpenAI, Introducing openai privacy filter <https://openai.com/index/introducing-openai-privacy-filter/> (April, 2026), Accessed: 2026-07-22.
- [20] X. Xi, D. Kong, J. Yang, J. Yang, Z. Chen, W. Wang, J. Wang, X. Cai, S. Zhang and W. Ye, SampleMix: A sample-wise pre-training data mixing strategy by coordinating data quality and diversity, in *Findings of the Association for Computational Linguistics: EMNLP 2025*, eds. C. Christodoulopoulos, T. Chakraborty, C. Rose and V. Peng (Association for Computational Linguistics, Suzhou, China, November 2025).
- [21] International Organization for Standardization and International Electrotechnical Commission, Information technology—Security techniques—Privacy framework, International Standard ISO/IEC 29100:2024, ISO/IEC (2024).
- [22] Y. Yu, Y. Zhuang, J. Zhang, Y. Meng, A. J. Ratner, R. Krishna, J. Shen and C. Zhang, Large language model as attributed training data generator: A tale of diversity and bias, *Advances in neural information processing systems* 36, 55734 (2023).
- [23] Google DeepMind, Gemini 3.5 flash model card Model card, (2026), Accessed: 2026-08-04.
- [24] A. Zeng, X. Lv, Q. Zheng, Z. Hou, B. Chen, C. Xie, C. Wang, D. Yin, H. Zeng, J. Zhang et al., GLM-4.5: Agentic, reasoning, and coding (ARC) foundation models, *arXiv preprint arXiv:2508.06471* (2025).
- [25] Meta AI, Llama 3.3 70b instruct model card Hugging Face model card, (2024), Accessed: 2026-08-04.
- [26] B. Xiao, B. Xia, B. Yang, B. Gao, B. Shen, C. Zhang, C. He, C. Lou, F. Luo, G. Wang et al., MiMo-V2-Flash technical report, *arXiv preprint arXiv:2601.02780* (2026).

- [27] Mistral AI, Mistral medium 3 model card Model card, (2025), Accessed: 2026-08-04.
- [28] S. Agarwal, L. Ahmad, J. Ai, S. Altman, A. Applebaum, E. Arbus, R. K. Arora, Y. Bai, B. Baker, H. Bao et al., gpt-oss-120b & gpt-oss-20b model card, arXiv preprint arXiv:2508.10925 (2025).
- [29] DeepSeek-AI, DeepSeek-V4-Flash model card Hugging Face model card, (2026), Accessed: 2026-08-04.
- [30] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv et al., Qwen3 technical report, arXiv preprint arXiv:2505.09388 (2025).
- [31] MiniMax, MiniMax M3 model card Model card, (2026), Accessed: 2026-08-04.
- [32] Liquid AI, LFM2.5-Encoder-350M-PII-Detector Hugging Face model card, (2026), Accessed: 2026-08-04.
- [33] J. Yun, S. Choi, Y. Lee, O. Sokol, W. Shim, A. Melkonyan and D. Kuzmenko, Project creddata: A dataset of credentials for research <https://github.com/Samsung/CredData>, (2021).
- [34] B. Lester, D. Pressel, A. Hemmeter, S. Ray Choudhury and S. Bangalore, Constrained decoding for computationally efficient named entity recognition taggers, in Findings of the Association for Computational Linguistics: EMNLP 2020, eds. T. Cohn, Y. He and Y. Liu (Association for Computational Linguistics, Online, November 2020).