Meddies-PII explained: multilingual de-identification of clinical text with a 350M model

25 November 2026 · 7 min read · blogs.iictp.org/meddies-pii-multilingual-clinical-text-deidentification-explained

Short answer: Meddies-PII is a dataset and a 350M-parameter token classifier for finding personally identifiable information (PII) in clinical text in seventeen languages, described in an arXiv preprint submitted on 22 November 2026 by Le Linh Uyen and colleagues. The authors report a mean exact-match F1 of 0.827 against 0.672 for the strongest of four public systems, and say they release the full generation pipeline, though the release is planned only after acceptance. This is an arXiv preprint. It has not been peer reviewed, and its numbers are the authors' own claims.

TL;DR The short version

- Meddies-PII is a dataset and a 350M-parameter token classifier for finding personal information in clinical text in seventeen languages.
- The authors report mean exact-match F1 of 0.827 against 0.672 for the strongest of four public systems.
- The release is planned only after acceptance and training data is synthetic, so test on your own text and measure recall per label.

Read online, listen to it, or change the text size: https://blogs.iictp.org/meddies-pii-multilingual-clinical-text-deidentification-explained

What is Meddies-PII?

De-identification removes names, dates, contacts and similar details from records before they are shared. The abstract says multilingual PII extraction in clinical text is limited by scarce governed data and by synthetic corpora whose construction is thinly documented. Meddies-PII covers seventeen languages, including seven Southeast Asian languages, and nine PII labels.

How it works

- **Synthetic data with documented generation.** The authors describe a generation pipeline and say they release prompts, catalogues, checks and label maps.
- **Small classifier.** A 350M-parameter BIOES token classifier labels each token as the beginning, inside, end, single or outside of a PII span.
- **Leakage check.** The authors report finding almost no near-duplicates and no text overlap between training and any evaluation set.

What is new compared with earlier approaches

The emphasis is documentation: the pipeline, prompts and checks are described so others can reproduce or audit the synthetic data, and a compact model is used rather than a large LLM.

Results reported by the source

- The model achieves the highest score on all fifteen public benchmarks, significantly ahead of the strongest of four public systems on fourteen of them.
- Mean exact-match F1 is 0.827 against 0.672 for the strongest system.
- Without adaptation, on German and Spanish clinical-style corpora it is competitive with the best public systems once annotation conventions are aligned, and best in four of ten MultiGraSCCo languages.

Limitations and open questions

- The dataset, model and code are planned for release upon acceptance, so they cannot be checked yet.
- Synthetic training data may not reflect real hospital notes.
- A missed identifier is a privacy failure; F1 does not show worst-case misses. Sources were read on 25 November 2026. Products, rules and standards change, so check the linked originals before you rely on any detail.

How a hospital or vendor could evaluate such a tool

1. Test on a sample of your own de-identified text, reviewed by a privacy officer.
2. Measure recall for each PII label separately, since a missed name is worse than a false alarm.
3. Check each language and script you hold.
4. Keep human review for release of data to third parties.

What it means for healthcare data teams

A small model that runs locally is attractive where data cannot leave the premises. The paper’s caveat about annotation conventions matters in practice: scoring depends on what counts as PII. De-identification does not by itself make data legally anonymous, so follow your own regulator’s rules.

Frequently asked questions

- **What is Meddies-PII?** A multilingual dataset and 350M model for PII extraction in clinical text, covering seventeen languages and nine labels.
- **What does it score?** A mean exact-match F1 of 0.827 against 0.672 for the strongest compared system.
- **Is it available?** The authors plan release upon acceptance.
- **Does it make data anonymous?** No. Legal anonymisation depends on your jurisdiction’s rules.

Related reading

- Gemini for healthcare
- Token pooling-aware training for smaller embeddings
- AI Practitioner certification

Related credential: AI Practitioner. Explore the exam facts and syllabus.

Sources and further reading

1. Uyen LL, Hoang C, Tri HD, Ha HH. Meddies-PII: A Multilingual Framework for Personally Identifiable Information Extraction in Clinical De-identification. arXiv 2611.29491, 22 Nov 2026 (preprint). arxiv.org/abs/2611.29491

Published by IICTP, an independent technical credential body, on 25 November 2026. The current version of this post is at https://blogs.iictp.org/meddies-pii-multilingual-clinical-text-deidentification-explained. More guides: blogs.iictp.org.