Towards Privacy-Compliant Healthcare Data: Developing an Efficient Multilingual Filter for PII and PHI under Safe Harbor Principles

Andrin Benz[1] , Jannis Schönleber[2], Bella Fernández[3] and Bernhard Bermeitinger[3]

[1]Institute of Computer Science, University of St. Gallen (HSG), St. Gallen (CH) | [2]44ai AG, Industriestrasse 78, 4600 Olten (CH) |

[3]Institute of Computer Science Vorarlberg, University of St. Gallen (HSG), Dornbirn (AT)

IC3K 2026 · KMIS · Angers, France · 28-30 October 2026

University of St.Gallen
Institute for Computer Science in Vorarlberg

Introduction

Data drives healthcare quality

- Switzerland leads in healthcare quality and social foundation
- Further gains in quality and efficiency need comprehensive, accessible data

PHI and PII face threats and strict laws

- Breaches, ransomware and patient-safety risks target PHI and PII
- Penalties and lost public trust call for anonymization
- GDPR, HIPAA and the Swiss DPA restrict data processing

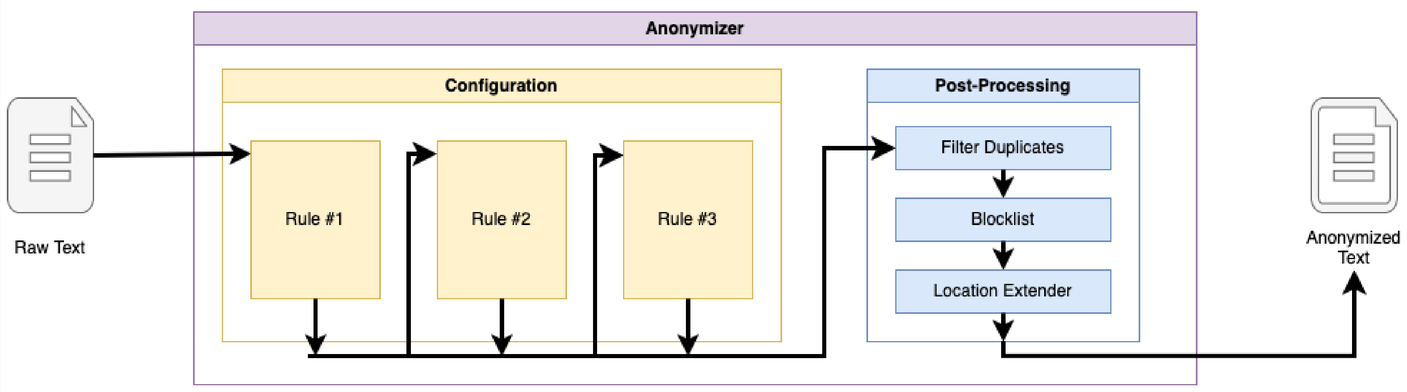
Local open-source LLMs open new options

- Open-source LLMs run locally without extensive hardware
- With guided generation, smaller models still achieve accurate results
- Basis for efficient, secure, privacy-compliant solutions with robust anonymization

This work: multilingual PII and PHI filter

- Masks text with placeholders and entity labels that keep context
- Runs locally on open-source LLMs
- Safe Harbor: removes the 18 HIPAA identifiers, as in i2b2, adapted to Swiss regulation

Model Architecture



1 Configuration
User-defined options run on the raw text or on the output of the previous configuration step

2 Post-processing
Follow-up cleanup and manual blocking rules, e.g. duplicate removal or a user-defined blocklist (Swiss city names, Swiss-known surnames)

3 Output
The final output is the anonymous (or pseudonymous) document

Dataset

Swiss synthetic clinical dataset (44ai AG[2])

50	17	7	1 012	19 663
documents	patients	entity types	entities	characters

- Synthetic Swiss German texts without local dialects, mirroring common reasons for hospital visits; about 13 records per patient

i2b2 benchmark dataset

520	8	125 966
documents	entity types	characters

- Yang et al., 2019; Stubbs and Uzuner, 2015; Norgeot et al., 2020. A sample of 520 documents suffices for benchmarking (Tannier et al., 2024)

Results

Swiss dataset, character-level F1

Rank	Configuration	F1
=1	Word-Level (Gemma 27B)	0.98
=1	Word-Level (Gemma 12B)	0.98
=1	Word-Level + Regex (Gemma 27B)	0.98
=1	Word-Level + Regex (Gemma 12B)	0.98
5	Prompt-only (Gemma 27B)	0.91
6	Prompt-only (Gemma 12B)	0.85
=7	Word-Level + Regex (Gemma 4B)	0.83
=7	Word-Level (Gemma 4B)	0.83
9	Regex Only	0.55
10	Prompt-only (Gemma 4B)	0.43
11	Prompt-only (Gemma 1B)	0.10

i2b2, character-level F1

Rank	Configuration	F1
1	Word-Level + Regex (Gemma 27B)	0.79
2	Prompt-only (Gemma 27B)	0.78
=3	Word-Level + Regex (Gemma 12B)	0.74
=3	Prompt-only (Gemma 12B)	0.74
5	Word-Level (Gemma 27B)	0.71
6	Word-Level (Gemma 12B)	0.64
=7	Regex Only	0.57
=7	Word-Level + Regex (Gemma 4B)	0.57
9	Word-Level (Gemma 4B)	0.49
10	Prompt-only (Gemma 4B)	0.47
11	Prompt-only (Gemma 1B)	0.02

Latency per document, synthetic dataset

Prompting strategy	Gemma 4B (R = 125)	Gemma 27B (R = 18.5)
Prompt-only	24 s	69 s
Word-Level Guided Generation	17 s	115 s

Best: word-level guided generation, F1 0.98 (Swiss) and 0.79 (i2b2)
Regex is fast but weak: 0.61 s per document, F1 0.55 (Swiss)
Guided beats prompt-only: F1 +0.07 (27B) to +0.40 (4B), Swiss data

Conclusions and future work

Conclusions:

- **Near-perfect on Swiss data:** exact-match F1 0.94, relaxed-match F1 0.98
- **Hybrid pipeline:** regex, rule-based filters and word-level guided generation with open-source LLMs
- **Transfers to English:** best character-level F1 0.79 on i2b2
- **Checked on real data:** a random subset of real Swiss documents showed no systematic errors

Limitations and future work:

- **Limits:** synthetic benchmark, mostly unstructured German text, no tables or images
- **Trade-off:** larger models are more accurate but slower (17 s vs 115 s per document)
- **Next:** weight entities by privacy risk, confidence thresholds from logprobs, multi-institution data

Hybrid filter reaches F1 0.98 on Swiss healthcare text with local open-source LLMs