Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Brazilian-PHI: Benchmarking Checksum-Validated Recognizers for CPF, CRM, CNS, CNPJ, RG, CEP, and Phone in Portuguese Clinical Text

Igor Eduardo

igorcnnbd@gmail.com

independent researcher https://orcid.org/0009-0005-6288-1135

Method Article

Keywords: de-identification, Brazilian Portuguese, clinical NLP, LGPD, Microsoft Presidio, CPF, checksum validation, PII detection, large language models

Posted Date: August 21st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10757966/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare potential competing interests as follows: The author is an AI Applied Engineer and founder of DocMinds AI, a clinical AI company that uses the evaluated recognizers in a production system. The author therefore holds a financial interest in the system under evaluation.

Brazilian-PHI: Benchmarking Checksum-Validated Recognizers for CPF, CRM, CNS, CNPJ, RG, CEP, and Phone in Portuguese Clinical Text

Igor Eduardo[1]

[1]DocMinds AI, Austin, TX, USA, igoreduardo@docminds.ai

April 2026 — *Submitted to Journal of Biomedical Informatics*

Abstract

De-identification of clinical text is a prerequisite for LGPD-compliant deployment of artificial intelligence systems in Brazilian healthcare. Despite growing adoption of clinical NLP in Brazil, no benchmark exists for detecting the seven Brazilian PII types most prevalent in clinical notes: CPF, CRM, CNS, CNPJ, RG, CEP, and phone numbers. Existing tools, including stock Microsoft Presidio and large language models (LLMs), are not designed for Brazilian identifier formats and lack checksum validation for CPF and CNPJ.

We present **Brazilian-PHI**, the first benchmark for Brazilian clinical PII detection. We evaluate three methods—custom Presidio recognizers with mod-11 checksum validation, default (unconfigured) Presidio, and Groq Llama 3.3 70B zero-shot—against 500 synthetic Portuguese clinical notes containing 670 PII instances and 1,940 distractors (ICD-10 codes, dates, CNES facility codes, medication dosages).

Custom Presidio achieves macro F1 = 1.000 across all seven PII types (n = 500), versus 0.270 for default Presidio and 0.989 for Llama 3.3 (n = 100, rate-limited). Checksum validation eliminates false positives for CPF and CNPJ entirely; LLMs cannot perform this mathematical verification. Latency for custom Presidio averages 9.9 μs per note versus 801.2 ms for Llama 3.3—an 80,000$\times$ difference that renders LLMs unsuitable as a primary LGPD scrubber at production scale. McNemar's test confirms custom versus default Presidio is statistically significant ($\chi^2 = 313.003$, $p < 0.001$). All recognizers and the synthetic corpus are released as open-source resources.

Keywords: de-identification; Brazilian Portuguese; clinical NLP; LGPD; Microsoft Presidio; CPF; checksum validation; PII detection; large language models

1 Introduction

Brazil's Lei Geral de Proteção de Dados (LGPD), particularly Article 11, classifies health data as sensitive personal data subject to stricter processing requirements. Any clinical AI system that ingests or generates free-text clinical notes must remove or mask personally identifiable information (PII) before inference or storage. Failure to do so constitutes a regulatory violation exposing organisations to fines of up to R$50 million per infraction.

Brazilian clinical text presents a de-identification challenge distinct from English or European clinical text. Seven identifier types recur with high frequency: the individual tax register (CPF), the regional medical council number (CRM), the national health card (CNS), the corporate register (CNPJ), the civil identity card (RG), the postal code (CEP), and Brazilian phone numbers. Each has a unique format; CPF and CNPJ additionally incorporate mod-11 checksum digits that enable mathematical validation of detected candidates—a property that regex-only and LLM-based approaches cannot exploit.

Despite the clinical and regulatory importance of this problem, no published benchmark evaluates PII detection specifically for these seven Brazilian identifier types. The closest prior work is AnonyMed-BR [1], the first Brazilian clinical anonymisation corpus, which annotates names, dates, locations, and "healthcare identifiers" but does not benchmark CPF, CRM, CNS, CNPJ, RG, or CEP as separate entity classes with checksum validation. Spanish-language work (MEDDOCAN [2]) and English-language de-identification benchmarks (n2c2 2014 [3]) cannot evaluate Brazilian formats. Stock Microsoft Presidio [4], the dominant open-source PII framework, ships without Brazilian recognizers and has been cited as achieving only approximately 82% accuracy on non-US identifier formats [8].

We address this gap with four research questions:

- **RQ1**: Do custom Presidio recognizers with checksum validation outperform default Presidio on Brazilian clinical PII?
- **RQ2**: Can zero-shot LLMs substitute for rule-based recognizers in terms of F1 accuracy?
- **RQ3**: Does checksum validation reduce false positives for CPF and CNPJ?
- **RQ4**: What is the latency cost of LLM-based detection and does it preclude production use?

2 Related Work

2.1 Clinical De-identification in English

Automated de-identification of English clinical text is a mature field. Stubbs et al. [3] established the n2c2 2014 shared task covering 18 HIPAA categories, which remains the primary English-language reference. Rule-based systems (PhysioNet DeID, NLM Scrubber) achieve F1 above 0.90 on structured identifiers [10]. LLM-based approaches have been evaluated more recently: Liu et al. [7] reported GPT-4 achieving F1 = 0.897 in zero-shot de-identification of English clinical notes. John Snow Labs' Healthcare NLP achieved 0.96 F1 in English, outperforming GPT-4o (0.79) in a controlled comparison [8].

2.2 Non-English Clinical De-identification

For Spanish, MEDDOCAN [2] provides a shared task covering Spanish clinical PII (NIF, address, profession, contact) but no Brazilian identifiers. For Portuguese, Schiezaro et al. [1] introduced AnonyMed-BR: a dataset of real and synthetic Brazilian medical records annotated for names, dates, locations, and healthcare identifiers. While AnonyMed-BR is the most directly related prior work, it does not evaluate CPF, CNPJ, CRM, CNS, RG, or CEP as separate entity classes and does not incorporate checksum validation. The authors explicitly note structured identifier detection as future work.

BioBERTpt [5] and mmBERT [6] (micro F1 = 0.76) are the strongest Portuguese clinical NER models, but neither was designed or evaluated for PII detection.

2.3 Checksum Validation in PII Detection

To our knowledge, no published benchmark for clinical de-identification incorporates checksum-based validation as an evaluation dimension. CPF numbers use a mod-11 algorithm over the first nine digits to generate two control digits; CNPJ uses an analogous mod-11 computation over fourteen digits. Checksum validation transforms CPF/CNPJ detection from probabilistic to deterministic, eliminating false positives that pattern-only detectors cannot reject.

2.4 Microsoft Presidio

Microsoft Presidio [4] is an open-source (MIT) PII detection framework supporting pluggable recognizers. Version 2.2.362 (March 2025) ships with 50+ built-in recognizers for US, EU, and international PII types. Brazilian identifier formats (CPF, CRM, CNS, CEP, RG, CNPJ) are absent from the default recognizer set.

3 Methods

3.1 Synthetic Corpus

We generated 500 synthetic Portuguese clinical notes using a template-based generator seeded with random but algorithmically valid PII instances. Each note includes at least one instance from each of the seven PII types, for a total of 670 PII instances (Table 1). All CPF and CNPJ values pass mod-11 checksum validation. Ground truth is stored as a JSON array of exact character-offset spans per note.

Distractors test false positive rates for formats that superficially resemble PII: ICD-10 alphanumeric codes (e.g., J18.9), dates in Brazilian format, CNES seven-digit facility codes, and medication dosages (e.g., "500mg 2x/dia"). No real patient data was used. The corpus is released under CC BY 4.0.

3.2 Custom Presidio Recognizers

We implemented seven custom Presidio recognizers:

1. **CPF** — regex for `XXX.XXX.XXX-XX` plus mod-11 checksum validation.
2. **CRM** — regex for `CRM/UF NNNNN` covering all 27 state abbreviations.

Table 1: Corpus PII and distractor distribution (n = 500 notes).

Entity type	Instances
CPF	150
CRM	100
CNS	80
CEP	100
RG	80
CNPJ	60
Phone	100
Total PII	**670**
ICD-10 codes	500
Dates (DD/MM/YYYY)	500
CNES codes (7-digit)	480
Medication dosages	460
Total distractors	**1,940**

3. **CNS** — regex for the 15-digit national health card, constrained to values beginning with 1 or 2 (Datasus specification).
4. **CEP** — regex for `XXXXX-XXX`.
5. **RG** — regex for `XX.XXX.XXX-X` and state variants.
6. **CNPJ** — regex for `XX.XXX.XXX/XXXX-XX` plus mod-11 checksum validation.
7. **Phone** — regex covering `+55`, DDD, and local formats.

These recognizers are deployed as Layer 2 of a five-layer clinical safety agent in a production clinical AI system in Brazil, compliant with LGPD Article 11.

3.3 Baseline Methods

Default Presidio. Stock Presidio v2.2.362 with no custom recognizers.

Groq Llama 3.3 70B (zero-shot). Llama 3.3 70B via the Groq API with a structured zero-shot prompt requesting Brazilian PII spans in JSON format. Evaluated on 100 of 500 notes due to free-tier rate limits during the benchmark run.

3.4 Evaluation Metrics

We compute precision, recall, and F1 per PII type using exact span matching. A detection is a true positive (TP) only if both entity type and character offsets match exactly. Additional metrics: macro F1 (unweighted mean), micro F1 (instance-weighted), checksum accuracy (mod-11 validation rate for detected CPF/CNPJ), distractor false positive rate, and wall-clock latency per note (mean and P95). McNemar's test [9] assesses statistical significance at $\alpha = 0.05$.

4 Results

4.1 RQ1: Custom vs Default Presidio

Table 2 shows F1 per PII type. Custom Presidio achieves macro F1 = 1.000 across all seven types (n = 500). Default Presidio achieves macro F1 = 0.270, driven by zero recall on five types (CRM, CNS, CEP, RG, CNPJ) absent from its built-in recognizer set.

Table 2: F1 score per PII type and method.

PII Type	Custom Presidio (n = 500)	Default Presidio (n = 500)	Groq Llama 3.3 (n = 100)
CPF	1.000	0.889	1.000
CRM	1.000	0.000	1.000
CNS	1.000	0.000	1.000
CEP	1.000	0.000	1.000
RG	1.000	0.000	1.000
CNPJ	1.000	0.000	0.923
Phone	1.000	1.000	1.000
Macro F1	**1.000**	0.270	0.989
Micro F1	1.000	0.494	0.996

Default Presidio's partial performance on CPF (F1 = 0.889) reflects its built-in generic tax-ID recognizer, which matches some CPF patterns without Brazilian-specific validation, yielding 30 false negatives (recall = 0.800) and no false positives.

4.2 RQ2: LLM Substitution

Llama 3.3 achieves macro F1 = 0.989 on 100 notes. The single failure mode is CNPJ (F1 = 0.923, precision = 0.857, recall = 1.000): one false positive was observed, likely a CNPJ-formatted string in a distractor context. The McNemar comparison on 100 notes is underpowered ($\chi^2 = 0.000$, $p = 1.000$); the critical limitation of LLMs is latency, not accuracy (see §4.4).

4.3 RQ3: Checksum Validation

Table 3 shows checksum accuracy. Custom Presidio performs mod-11 verification, achieving 100% checksum validity for all detected CPF and CNPJ instances. Default Presidio and Llama 3.3 match by pattern only; neither reports checksum-valid detections.

Table 3: Checksum validation accuracy for detected CPF and CNPJ instances.

Method	CPF checksum valid	CNPJ checksum valid
Custom Presidio	100.0%	100.0%
Default Presidio	N/A	N/A
Groq Llama 3.3	N/A	N/A

In real clinical text, strings matching the CPF pattern but failing the mod-11 check (e.g., document numbers, form codes, typographical errors) would generate false positives

in pattern-only detectors. Checksum validation deterministically eliminates this class of errors.

Table 4 shows distractor false positive rates. All regex-based methods produce zero false positives. Llama 3.3 produces a 1.0% CNES false positive rate (1/100 notes), indicating the model occasionally misclassifies seven-digit facility codes as PII.

Table 4: Distractor false positive rate by method and category.

Method	ICD-10	Dates	CNES	Dosages
Custom Presidio	0.0%	0.0%	0.0%	0.0%
Default Presidio	0.0%	0.0%	0.0%	0.0%
Groq Llama 3.3	0.0%	0.0%	1.0%	0.0%

4.4 RQ4: Latency

Table 5 shows processing latency. Custom Presidio processes a clinical note in 9.9 μs on average. Llama 3.3 requires 801.2 ms—an **80,000× latency difference**. At a throughput of 100 notes/second required for real-time clinical inference, Llama 3.3 would require approximately 80,000 parallel inference slots—an infrastructure cost prohibitive for most Brazilian healthcare deployments.

Table 5: Processing latency per note (single-threaded wall clock).

Method	Mean	P50	P95
Custom Presidio	9.9 μs	5.6 μs	19.8 μs
Default Presidio	7.4 μs	5.5 μs	8.3 μs
Groq Llama 3.3	801.2 ms	805.4 ms	1271.3 ms

4.5 Statistical Significance

McNemar's test on paired note-level outcomes:

- **Custom vs Default Presidio** (n = 500): $\chi^2 = 313.003$, $p < 0.001$. Custom Presidio succeeds on 315 notes where default fails; default succeeds on 0 notes where custom fails.
- **Custom vs Groq Llama 3.3** (n = 100): $\chi^2 = 0.000$, $p = 1.000$. Not significant on the current sample (underpowered).

5 Discussion

5.1 Rule-Based vs LLM for Structured Identifiers

Our results suggest a principled division of labour. For structured identifiers with fixed formats and checksum validation (CPF, CNPJ), rule-based recognizers are superior: deterministic, 80,000× faster, and able to mathematically reject false positives. For unstructured PII (names, free-form addresses), LLMs are likely superior due to their ability to

generalise to unseen surface forms. The recommended deployment pattern is hybrid: custom Presidio recognizers for structured Brazilian identifiers, with an LLM or fine-tuned NER layer for unstructured PII.

5.2 The CNPJ Format Change

The Receita Federal do Brasil will introduce alphanumeric CNPJ format from July 2026 (Normativa RFB 2229/2024), replacing the all-numeric format with a 12-character alphanumeric prefix. The current recognizer handles only the existing numeric format; an update is required before July 2026 to maintain coverage. We flag this as a concrete, time-bounded limitation.

5.3 Limitations

First, the evaluation corpus is entirely synthetic. Real clinical notes may contain abbreviations, misspellings, and formatting variations not captured by template-generated text. Second, Llama 3.3 was evaluated on only 100 of 500 notes; the McNemar comparison is underpowered. Third, BioBERTpt and mmBERT were not evaluated and represent a natural comparison point for future work.

6 Conclusion

We presented Brazilian-PHI, the first benchmark for detecting the seven Brazilian PII types most prevalent in clinical text. Four findings emerge:

1. Custom Presidio recognizers with mod-11 checksum validation achieve macro F1 = 1.000 across all seven types on 500 synthetic clinical notes.
2. Default Presidio achieves macro F1 = 0.270 because it lacks recognizers for five of seven Brazilian identifier types.
3. Checksum validation deterministically eliminates false positives for CPF and CNPJ; LLMs and pattern-only methods cannot provide this guarantee.
4. Llama 3.3 70B is competitive in F1 (0.989) but 80,000× slower, precluding use as a primary LGPD scrubber at scale.

A hybrid architecture—rule-based recognizers for structured identifiers, LLM or NER for unstructured PII—is the recommended pattern for LGPD-compliant clinical AI deployment in Brazil.

Data Availability

The synthetic evaluation corpus, all seven custom Presidio recognizers, and evaluation scripts are available at `https://github.com/nomad-link-id` under MIT license (code) and CC BY 4.0 (corpus). No real patient data was used.

Ethics Statement

This study uses entirely synthetic data. No real patient records were used. The work supports patient privacy protections under LGPD Article 11.

Declaration of Competing Interests

The author is co-founder of a clinical AI company that uses the evaluated recognizers in a production system.

References

[1] M. Schiezaro, G. Rosa, B.A.G. Campos, H. Pedrini, "Guardians of the data: NER and LLMs for effective medical record anonymization in Brazilian Portuguese," *Frontiers in Public Health*, vol. 13, 2026. https://doi.org/10.3389/fpubh.2025.1717303

[2] M. Marimon et al., "Automatic de-identification of medical texts in Spanish: the MEDDOCAN shared task," in *IberLEF 2019*, 2019.

[3] A. Stubbs, C. Kotfila, O. Uzuner, "Automated systems for the de-identification of longitudinal clinical narratives: overview of 2014 i2b2/UTHealth shared task Track 1," *J. Am. Med. Inform. Assoc.*, vol. 22, pp. e307–e315, 2015.

[4] Microsoft, "Presidio — Data Protection and Anonymization API," 2020. https://github.com/microsoft/presidio

[5] A. Schneider et al., "BioBERTpt — A Portuguese neural language model for clinical NER," in *Proc. ACL ClinicalNLP Workshop*, 2020.

[6] V.A. de Almeida et al., "Clinical named entity recognition in the Portuguese language: a benchmark of modern BERT models and LLMs," arXiv:2603.26510, 2026. https://arxiv.org/abs/2603.26510

[7] Z. Liu et al., "Evaluating GPT models for clinical note de-identification," *JMIR Medical Informatics*, 2024. https://pmc.ncbi.nlm.nih.gov/articles/PMC11785955/

[8] IntuitionLabs, "Open Source PHI De-Identification: A Technical Review," Mar. 2026. https://intuitionlabs.ai/articles/open-source-phi-de-identification-tools

[9] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947.

[10] Censinet, "2025 Benchmark: De-Identification Tools," 2025. https://censinet.com

[11] I. Eduardo, "BM25 and Dense Retrieval Are Complementary for Portuguese Clinical Text," Zenodo, 2026. https://doi.org/10.5281/zenodo.19686739