Local Anonymization of Free-Text Robbery Police Reports in Brazilian Portuguese: a Safety-Recall-Oriented Multi-Model Benchmark

Murilo G. de Almeida[1], Leonardo C. Botega[1], Paula Miwa de Paiva Lima[2]

[1] Universidade Estadual Paulista (UNESP), Brazil
murilo.goes@unesp.br, leonardo.botega@unesp.br
[2] Polícia Militar do Estado de São Paulo / Secretaria de Segurança Pública (CAP/SSP), Brazil
paulalima@policiamilitar.sp.gov.br

Abstract. Brazilian public-safety agencies release structured crime statistics but withhold the free-text narrative field of police reports, which is rich in modus operandi and contextual information, because it contains personal data protected by the Brazilian General Data Protection Law (LGPD). We address this deadlock for robbery reports with a fully local (on-premise) two-stage anonymization pipeline: a deterministic regular-expression stage for structured personally identifiable information (PII) and a pluggable semantic layer (dedicated NER, local LLMs served by Ollama, union and chained hybrids, and Microsoft Presidio as a system baseline). We contribute (i) an expert-annotated gold standard of 996 robbery narratives, released with the source code; (ii) an entity-level evaluation protocol whose headline metric is safety recall (how much PII actually did not leak), complemented by a re-identification-criticality analysis and bootstrap confidence intervals; and (iii) an empirical benchmark of nine engine configurations quantifying the safety–utility trade-off. The union of spaCy and qwen3:8b over the regex stage reaches the highest macro safety recall (0.986, 95% CI [0.972; 0.994]; micro 0.973) and is the most conservative (highest-recall) option, while qwen3:8b alone gives the best overall balance (safety recall 0.966, precision 0.80, micro-F1 0.826, text retention 96%) and is our recommendation for general use. In this configuration, direct identifiers (names, documents, phones, e-mails) are removed at 99–100%, with residual leakage concentrated in location quasi-identifiers. Safety recall measures PII removal, not residual re-identification risk: these figures bound leakage rather than certify anonymity. We recommend qwen3:8b for general supervised anonymization, and the union when leakage must be minimized, and discuss conditions for replication across Brazilian police forces.

CCS Concepts: • **Security and privacy** → **Data anonymization and sanitization**; • **Computing methodologies** → *Natural language processing*.

Keywords: anonymization, de-identification, large language models, LGPD, named entity recognition, police reports, text mining

1. INTRODUCTION

The Public Security Department of the State of São Paulo (SSP-SP) widely releases structured crime data on its statistical portal (in 2024 alone there were 205,836 robbery occurrences in the state [SSP-SP 2024]). Safely leveraging the qualitative content of police reports (*boletins de ocorrência*, BOs), however, is a major challenge. The free-text "Histórico" (narrative) field, in which the judicial police records the dynamics of the event, is the richest source for the contextual analysis of the crime; because it is written at the discretion of the police authority, it frequently mentions personal data of victims, witnesses, officers, and suspects, even though the formal identifiers reside in the structured fields. Sharing it in raw form would expose agencies to risks under the Brazilian General Data Protection Law (LGPD) [Brasil 2018], which restricts access by researchers and public managers.

Automatic anonymization of the Histórico field addresses this impasse, but the public-security context imposes an architectural constraint. Resorting to proprietary solutions running on private companies' clouds entails sending sensitive data to third parties, with the risk that it is not even processed within the country, of unauthorized retention, and of use for purposes other than intended, with the consequent loss of sovereignty over the data. All inference in the pipeline proposed here therefore runs *locally* (on-premise), so that personally identifiable information (PII) never leaves the agency. The very consolidation of reliable, standardized, and accessible public-security data in Brazil is a recurring challenge [FBSP 2025], which reinforces the value of safely enabling access to the qualitative content of BOs. Our focus here is anonymization and its evaluation. The contributions, in order of relevance:

(1) **Published dataset and code.** The first Brazilian-Portuguese dataset annotated for anonymization in the police domain: 996 de-characterized robbery-BO narratives, with a gold standard of PII entities in eight types, annotated by domain experts (Section 3) and released together with the pipeline code (Availability Statement).
(2) **Risk-oriented evaluation protocol.** Per-type, entity-level evaluation under a *dual view*: per-type F1 (utility) and *safety recall* (how much PII actually did not leak), with a leakage analysis by re-identification criticality (direct vs. quasi-identifiers [Sweeney 2002]), textual-utility metrics, and bootstrap confidence intervals [Efron and Tibshirani 1993].
(3) **Fully local multi-model benchmark.** Nine configurations compared under the same protocol: dedicated NER (spaCy, BERTimbau-LeNER), local LLMs (qwen3:8b, llama3.1:8b, gemma3:12b), union and chained combinations, and Microsoft Presidio as a system *baseline*, with the safety× utility *trade-off* quantified.
(4) **Two-stage pipeline** (deterministic regex + pluggable semantic layer) as enabling engineering, with an evidence-based recommendation for supervised production use.

2. RELATED WORK

Free-text de-identification is classically modeled as named entity recognition (NER) followed by substitution. Internationally, the i2b2 shared tasks on English clinical narratives established the annotated-corpus-plus-shared-evaluation paradigm for this problem [Uzuner et al. 2007], and neural sequence models later became the standard approach [Dernoncourt et al. 2017]; that literature assumes an English clinical register and unconstrained compute. In Brazilian Portuguese, the literature concentrates on the clinical domain: the SemClinBr corpus [Oliveira et al. 2022] and the BioBERTpt model [Schneider et al. 2020] consolidated neural clinical NER, and the most direct precedent of this work, [Schiezaro et al. 2026], compares extractive (BERT-based NER) and generative (T5, GPT) strategies to anonymize medical records, introducing AnonyMed-BR (the first PT-BR *dataset* annotated for anonymization) and an *LLM-as-a-Judge* module for re-identification risk. In the legal domain, LeNER-Br [Luz de Araujo et al. 2018] provides an NER corpus over legal text, and BERTimbau [Souza et al. 2020] is the reference BERT for PT-BR; we evaluate a BERTimbau *checkpoint* fine-tuned on LeNER-Br under the hypothesis (qualified by our results) that the legal vocabulary would be close to the police one. At the system level, Microsoft Presidio[1] is the most widely used off-the-shelf *framework*; we include it in the matrix to answer, in a citable way, whether a ready-made tool would already solve the problem.

The gap this paper fills has three dimensions: the *domain* (robbery police narratives: a textual genre with operational jargon, colloquial language, and descriptions of stolen objects, not yet addressed in the anonymization literature); the *architectural constraint* (fully local execution, a sovereignty requirement absent from clinical work); and the *metric* (we elevate PII recall to the headline metric, with a criticality analysis, instead of the usual F1). These same three dimensions preclude a direct numerical

[1]Microsoft Presidio: https://microsoft.github.io/presidio/

Input (de-characterized text, in Portuguese): Comparece nesta Unidade Policial a vítima, juntamente com o representante da Empresa Viação Auto Metropolitana Ltda, Sr. Carlos da Silva Campos – telefone (11) 94354-8765, informando que é cobrador de ônibus e que, na data de ontem por volta das 20:45h, fazia a linha 205 Vila Primavera – Jardim Esperança quando foi abordado por um indivíduo que anunciou o assalto. [. . .]

Anonymized output: Comparece nesta `[LOCAL_1]` a vítima, juntamente com o representante da `[ESTABELECIMENTO_1]`, Sr. `[NOME_1]` – `[TELEFONE]`, informando que é cobrador de ônibus e que, na data de ontem por volta das 20:45h, fazia a linha 205 `[ENDERECO]` quando foi abordado por um indivíduo que anunciou o assalto. [. . .]

Fig. 1. Real pipeline example (union configuration). Name, phone, company, and address are replaced by typed placeholders; each index identifies a distinct entity (coreference), and [. . .] marks omitted spans. The narrative is a real Brazilian-Portuguese sample kept verbatim; names are already fictitious in the publishable de-characterized version.

comparison with the PT-BR clinical systems above: they are trained and scored on a different register, over a different entity inventory, and under F1 rather than safety recall, so transporting their reported figures to our corpus would compare protocols, not systems. We therefore compare against what is transportable — a domain-agnostic off-the-shelf *framework* (Presidio) and dedicated NER models — evaluated here under a single protocol on the same 996 narratives.

3. CORPUS, EXPERT ANNOTATION, AND EVALUATION PROTOCOL

3.1 Corpus and Gold Standard

The corpus comprises 996 robbery-BO narratives (Art. 157 of the Brazilian Penal Code [Brasil 1940]). Each record contains the original text, a *de-characterized* version (publishable, with names and locations replaced by plausible fictitious values), and eight PII annotation columns: Name(s), Locations, Document(s), License plate, Phone, E-mail/accounts, Patrol-car number (VTR), and Nickname(s). Entities are listed *verbatim* from the de-characterized text, which is the basis of all processing, ensuring an exact correspondence between annotation and text (Figure 1 shows an input and its output). The distribution is strongly imbalanced: Locations 729, Name 502, Document 65 (natural-person, after the reclassification below), Plate 45, Phone 11, VTR 5, E-mail 1, Nickname 1. The methodological consequence is direct: removing robbery PII depends mostly on the semantic layer (names and locations), not on the regex.

Annotation provenance. The de-characterization and gold-standard annotation were carried out by four public-security experts from the State Public Security Department, coordinated by a Military Police captain heading the analysis and planning section. The annotation was *disjoint* (each BO annotated by a single expert), which precludes computing inter-annotator agreement; we state this limitation explicitly (Section 8). The gold standard was subsequently audited, producing a change log (1,860 edits over 952 records: 865 formatting normalizations and ∼520 entity-level corrections, 44 of them removals of remaining PII); the audit revealed that the Document column mixed natural-person documents (RG/CPF/CNH/IMEI) with institutional codes (IML, NOC, BOPM, CNJ case number), which were *reclassified as hygiene*, removed by the pipeline but excluded from the personal-PII metrics.

3.2 Evaluation Protocol and Metrics

Evaluation is over the entire pipeline, at the per-type entity level, against the eight gold columns. For each BO and type: TP = a gold entity removed under the correct type; FN = an entity that slipped through (*leakage*, the severe error); FP = a removal that did not belong to that type (*over-removal*, a utility cost). Matching is robust (lowercase, accent-insensitive, punctuation-tolerant, bidirectional containment). Two measurement decisions are central:

(a) Dual view. Besides per-type precision, recall, and F1 (micro and macro), we report a *safety recall*: "removed under *any* placeholder = did not leak". From the standpoint of re-identification risk

and the LGPD, what matters is whether the PII left the text, even under a different label; the error matrix separates, for each FN, actual leakage from mere type confusion. **(b) Quantified utility.** We report the *over-removal rate* (FP divided by total removed), the fraction of text characters replaced by placeholders, and content *retention* (fraction of narrative characters preserved), besides per-record latency.

Uncertainty. For each configuration, 95% confidence intervals via non-parametric bootstrap [Efron and Tibshirani 1993] (10,000 replicates, resampling the 996 BOs with replacement) for macro safety recall and micro-F1; among the best configurations, a *paired* comparison (same resampling indices) with the CI of the difference.

4. ANONYMIZATION PIPELINE

The pipeline is deterministic where stable and pluggable where semantic; substitution is shared by all engines, isolating extraction quality from the other components.

Stage 1: deterministic regex. Removes everything with a stable pattern, replacing it with typed placeholders in "most-specific-first" order: e-mail/@handle/URL; institutional references (hygiene); natural-person documents (CPF, RG, CNH, IMEI, CNPJ, professional registrations); phones; dates; patrol cars; license plates; GPS coordinates; and addresses, followed by a mop-up sweep that protects legal references ("Art. 157 of the Penal Code" is not anonymized). Domain decision: a stolen object's IMEI and plate *are* PII; a generic make/model (iPhone, Honda) is *not*, since removing it would degrade utility without privacy gain.

Stage 2: pluggable semantic layer. Detects person names, named locations, establishments, and nicknames, under a single contract with interchangeable *back-ends*: `ner` (spaCy `pt_core_news_lg`; BERTimbau fine-tuned on LeNER-Br, with slicing into blocks of < 512 tokens), `llm` (local Ollama[2]; the model returns only an entity JSON, temperature 0.1 and the same *prompt* for all, ensuring a fair comparison), `hybrid` (union NER $\cup$ LLM), `chained` (one engine, substitution, then the other), and `presidio` (system *baseline*, default PT configuration; its NER engine is the same spaCy evaluated in isolation, so this row measures the *ready-made package*). Substitution is local and identical for all: by decreasing length, case-insensitive and whole-word, with a consistent per-entity index (coreference: `[NOME_1]` always denotes the same person) and sentinel protection of the Stage-1 placeholders. An exclusion filter removes pronouns, procedural roles and domain generics (indivíduo, suspeito, guarnição: a police officer's surname is a name, the role is not), kinship terms, and makes/models.

5. EXPERIMENTAL DESIGN

All engines run locally, on a single workstation (Intel Xeon W-2155 @ 3.30 GHz, 128 GB RAM, one NVIDIA Quadro P4000 GPU with 8 GB), which reinforces that the pipeline is feasible on modest, widely available on-premise hardware; the reported latencies refer to this setup. Stage 1 is always applied; we vary the Stage-2 engines: (1) *baseline* `regex_only`; (2) NER-only: spaCy and BERTimbau; (3) LLM-only: qwen3:8b, llama3.1:8b, and gemma3:12b; (4) Presidio; (5) hybrid (union); and (6) chained in both orders, instantiated only for the best NER/LLM pair from the previous rounds, in a deliberately lean matrix. Each configuration runs with a resumable *checkpoint* and produces, per record, the entities removed by stage/type, latency, and tokens, consumed by the evaluation *harness* without re-running the models. The statistical and utility analyses in this version reuse exclusively these frozen per-record results; the only later re-execution was of Stage 1 (deterministic) for the GPS-coordinate fix (Section 6.2).

[2]Ollama: https://ollama.com. Models qwen3:8b, llama3.1:8b, and gemma3:12b, obtained from the Ollama library.

Table I. Per-configuration results (N=996), sorted by safety recall (Saf. rec.), with 95% bootstrap CIs (10,000 replicates). Over-rem. = FP/(TP+FP) = 1−micro precision; Reten. = fraction of narrative characters preserved; Lat. = per-record latency.

Configuration	Saf. rec. [95% CI]	Micro-F1 [95% CI]	Over-rem.	Reten.	Lat. (s)
union (spaCy∪qwen3)	**0.986** [0.972; 0.994]	0.496 [0.471; 0.522]	0.662	0.934	7.0
chained llm→ner	0.986 [0.972; 0.994]	0.504 [0.478; 0.530]	0.655	0.935	9.6
chained ner→llm	0.985 [0.971; 0.992]	0.482 [0.455; 0.508]	0.671	0.934	11.2
llm llama3.1:8b	0.974 [0.958; 0.982]	0.788 [0.750; 0.821]	0.280	0.959	3.3
llm gemma3:12b	0.970 [0.952; 0.977]	0.780 [0.738; 0.818]	0.284	0.960	4.2
ner spaCy	0.967 [0.948; 0.975]	0.470 [0.442; 0.497]	0.674	0.937	0.02
presidio (≡ spaCy)	0.967 [0.948; 0.975]	0.470 [0.442; 0.497]	0.674	0.937	0.02
llm qwen3:8b	0.966 [0.948; 0.974]	**0.826** [0.785; 0.862]	**0.202**	**0.960**	3.7
ner BERTimbau	0.832 [0.800; 0.958]	0.483 [0.456; 0.510]	0.660	0.960	0.29
regex_only (*baseline*)	0.652 [0.595; 0.752]	0.317 [0.282; 0.352]	0.279	0.980	0.001

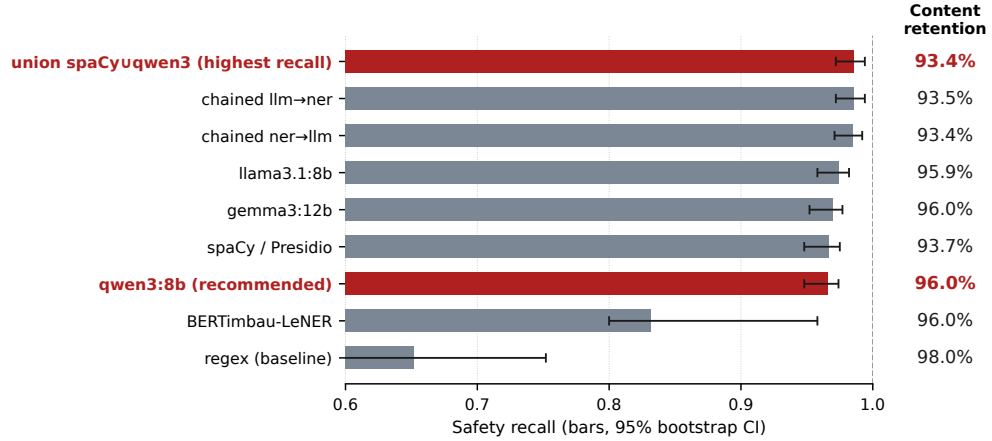

Fig. 2. Safety×utility trade-off per configuration: safety recall (bars, with 95% bootstrap CIs) and content retention (right-hand column). The union is the highest-recall option; standalone qwen3:8b is the best balance and our recommendation for general use; the *baseline* regex retains the most text but leaks far more PII.

6. RESULTS

Table I consolidates the nine configurations plus the *baseline*, sorted by safety recall, with 95% bootstrap CIs, over-removal rate, and content retention (N = 996; micro-aggregated F1 and precision; macro safety recall). Safety recall is a macro average over the eight types; the union's micro (entity-weighted) safety recall is 0.973 (36/1,359), the gap due to rare types removed at 100%.

6.1 Main Findings

The safety top is statistically comparable. The union (regex + spaCy∪qwen3) and the chained llm→ner tie on safety recall (0.986), and the paired bootstrap comparison confirms indistinguishability (0.0000; 95% CI [−0.0005; +0.0006]). Among the top three, the union has the lowest latency (7.0 s/record vs. 9.6–11.2 s for the chained variants), making it the *highest-recall* option. Its safety advantage over the best standalone models is small but real (about +0.02 over qwen3:8b; +0.018 over gemma3:12b, CI [+0.012; +0.024]), at the cost of much lower precision (0.34 vs. 0.80 for qwen3:8b).

The union's utility cost is real but contained. Its micro precision is 0.34 (over-removal of 0.66): by uniting two detectors, it removes more ambiguous strings than either engine alone. The textual impact, however, is small: the union replaces 6.6% of the narrative characters (93.4%

Table II. Residual leakage by criticality in the highest-recall configuration (union), frozen evaluation. None of the three plates is a valid plate: two occurrences of a truncated code and one chassis/VIN annotated as a plate. Rows with $n \leq 5$ are listed for completeness only (Section 8).

Type	Criticality	Gold	Leaked	% removed
Document (CPF/RG/CNH/IMEI)	Direct	65	0	100
Phone	Direct	11	0	100
E-mail/accounts	Direct	1	0	100
Name	Direct	502	5	99.0
VTR	Quasi-id.	5	0	100
Nickname	Quasi-id.	1	0	100
Location	Quasi-id.	729	28	96.2
Plate	Quasi-id.	45	3	93.3

retention), against 4.0% for qwen3:8b (96.0%) and 2.0% for the *baseline* regex (Figure 2). Over-removal concentrates on short, ambiguous strings (numbers, capitalized terms), not on extensive narrative portions.

Standalone qwen3:8b is the best balance. It combines the highest micro-F1 (0.826) and the highest precision (0.80) with high safety recall (0.966) and 96% retention, at 3.7 s/record and using a single model, making it our recommendation for general use. llama3.1:8b (F1 0.79) and gemma3:12b (F1 0.78) follow closely; the NER engines and Presidio sit at F1 $\approx$ 0.47. **BERTimbau-LeNER was the worst semantic engine** (safety recall 0.832; wide CI [0.800; 0.958], reflecting a tail of BOs with many leaks): the legal *checkpoint* is less adherent to the names and locations of police narratives, qualifying the vocabulary-adherence hypothesis. **Presidio $\equiv$ spaCy** on every metric, confirming it measures the ready-made package over the same engine. **Document has a per-type recall of 0.26 but zero actual leakage:** the 48 FNs were removed under another placeholder (`[CODIGO]`/`[TELEFONE]`), with safety recall 1.0 for the type; the low figure is an attribution artifact. In Stage 1, plate has F1 0.97, VTR 1.00, and phone recall 1.00.

6.2 Leakage by Criticality and Production Fix

The union's residue was inspected entity by entity and classified by re-identification criticality [Sweeney 2002]: *direct identifiers* (identifying on their own: name, document, phone, e-mail) vs. *quasi-identifiers* (identifying only in combination: location, plate, VTR, nickname). Table II summarizes: direct identifiers total 5 leaks in 579 entities (99.1% removed; personal document: 100%); quasi-identifiers, 31 in 780 (96.0%), concentrated in location. The structured direct identifiers (document, phone, e-mail) stay near 100% also under qwen3:8b; its difference from the union is in names (recall 0.93 vs. 0.98), motivating the union when name protection is paramount.

The qualitative inspection shows that part of the "leakage" is annotation noise: among the 5 name leaks, $\sim$2 are genuine misses (lowercase surname; a 10+-name BO) and the rest are edge cases (attached rank, concatenated PIX beneficiary); among the 28 location leaks, $\sim$1/3 is noise (loose numbers, a bank account, a country) and two are GPS coordinates, the rest being low-power quasi-identifiers (bus lines, stations, neighborhoods). Discounting noise, actual location leakage is $\sim$2%. An independent word-boundary scan of the released texts confirms these figures (names 99.0%; locations $\sim$97.7%), so the safety-recall metric is not optimistic.

Production fix (GPS). A DMS coordinate recognizer was incorporated into Stage 1 *after* the evaluation was frozen; Table I keeps the frozen numbers for a fair comparison among engines. Re-running only Stage 1 with it active captures the two coordinates (the only truly sensitive location leaks): in production the union reaches 0.986 (0.9856 $\rightarrow$ 0.9860), total leaks 36 $\rightarrow$ 34, location removal 96.2% $\rightarrow$ 96.4%, F1/precision unchanged.

7. DISCUSSION

Safety is governed by free text. The *baseline* regex, faithful on structured PII, reaches only 0.652 safety recall because it captures neither names (recall 0.000) nor most locations (0.274); it is the semantic layer that lifts it to ∼0.99. **Why the union beats chaining:** the union maximizes recall by construction (a single engine detecting suffices), and spaCy and qwen3:8b err at different points (NER is strong on locations; the LLM, on colloquial names and fine-grained types); chaining hides context by substituting before the second engine and pays for two sequential inferences, being slower without a safety gain. **Choice by usage profile:** under the LGPD, where leakage (FN) is the dominant error, standalone qwen3:8b is the general-use recommendation, combining high non-leakage (0.966), the best precision (0.80), and operational simplicity (a single model, fully local); when minimizing leakage is the absolute priority and over-removal is tolerable under review, the union raises safety recall to 0.986, under *supervised anonymization* with human review of the ∼1.4% residue, concentrated in location quasi-identifiers (Section 6.2). All figures reported here are *removal* rates over a single-annotator gold standard: safety recall does not estimate the residual risk of the quasi-identifiers that survive, which identify only in combination [Sweeney 2002] and which we did not attack empirically (Section 8). The dual view of metrics makes the *trade-off* explicit (Table I) and leaves the operating-point choice to the manager, with numbers rather than judgment.

The feasibility of fully local processing preserves sovereignty over sensitive data and eliminates the risk of third-party retention. Since the record systems of all Brazilian police forces have a field equivalent to the Histórico, the approach may be *adaptable* to other forces; we make no claim of validated national-scale transfer, which would require recalibrating the exclusion lexicons and regular expressions to each force's writing style and, crucially, new local validation (Section 8). We make no claim of automatic legal compliance with the LGPD: the pipeline is a technical instrument that substantially reduces the risk of PII exposure; the legal framing of the anonymized data rests with the controller.

8. THREATS TO VALIDITY

Rare types. Document, Phone, VTR, E-mail, and Nickname have few instances (Nickname: one); per-type metrics there are unstable, so strong conclusions are restricted to Name, Location, and aggregate safety recall. **Disjoint annotation.** Each BO was annotated by a single expert; there is no overlap for inter-annotator agreement. The audit log mitigates, but does not replace, this measure. **Gold under-annotation.** The audit identified phones present in the text but absent from the annotation, which the pipeline removes and that count as FP: the reported precision is a *lower bound* for some types, and recall is mildly upward-biased (PII absent from the gold cannot be scored as a leak); the gold standard was not altered. **NER label mapping.** spaCy/BERTimbau output PER/LOC/ORG; nickname is a subcase of PER, favoring LLM configurations on the Nickname column: a declared limitation, not an error. **LLM non-determinism.** Even at temperature 0.1, local LLMs are not bit-for-bit reproducible; the reported results correspond to a frozen controlled run, and the bootstrap CIs quantify sampling uncertainty. Re-running qwen3:8b three times gave a macro safety recall of 0.9665 ± 0.0003 (SD), far below the bootstrap interval, i.e., negligible. **Example echoing by the LLM.** In some BOs the LLM echoed *prompt*-example entities absent from the text; since substitution requires an exact match, they never alter the output or safety recall, but inflated false positives, so we excluded them before computing precision/F1 and fixed the *prompt* in the released code. **Generalization.** Corpus from a single jurisdiction/period; style and vocabulary vary across forces. **Re-identification risk not measured directly.** We measure entity removal, not residual risk from combining quasi-identifiers; an *LLM-as-a-Judge* module [Schiezaro et al. 2026] is left as future work.

9. CONCLUSION

This paper presented an expert-annotated *dataset* for anonymizing PT-BR robbery-BO narratives, a safety-recall-oriented evaluation protocol with criticality analysis and bootstrap CIs, and a fully local multi-model benchmark. The union regex + spaCy∪qwen3:8b maximizes non-leakage (0.986 [0.972; 0.994]; 0.9860 with the production GPS fix), statistically tied with the best chained variant but cheaper; standalone qwen3:8b is the best balance between safety and utility (safety recall 0.966; F1 0.826; precision 0.80; retention 96%) and our recommendation for general use. In the highest-recall configuration, direct identifiers are removed at 99–100%, with the residue concentrated in location quasi-identifiers. These are removal rates, not residual re-identification risk; we therefore recommend supervised anonymization, with no claim of automatic LGPD compliance or of anonymity in the strict sense. Future work: an *LLM-as-a-Judge* module for residual risk and information loss; expanding the gold standard for rare types, with overlapping annotation for agreement; transfer evaluation to other states and crime types; and integration into the situational-awareness workflow that motivated this work.

Data and Code Availability Statement. The de-characterized *dataset* with the annotated gold standard (with the original text removed) the complete source code of the pipeline, the evaluation *harness*, and the statistical analyses, together with an open-source local web application (Streamlit) for single-text and batch (spreadsheet) anonymization, are available under open licenses (code under MIT, dataset under CC BY 4.0) at: `https://github.com/murilogoes/bo-anonymizer-ptbr`. It includes the verbatim extraction *prompt*, the exact Ollama model tags, and the frozen per-record outputs.

REFERENCES

Brasil. Decreto-Lei n. 2.848, de 7 de dezembro de 1940. Código Penal. https://www.planalto.gov.br/ccivil_03/decreto-lei/del2848compilado.htm, 1940.

Brasil. Lei n. 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm, 2018.

Dernoncourt, F., Lee, J. Y., Uzuner, Ö., and Szolovits, P. De-identification of Patient Notes with Recurrent Neural Networks. *Journal of the American Medical Informatics Association* 24 (3): 596–606, 2017.

Efron, B. and Tibshirani, R. J. *An Introduction to the Bootstrap*. Chapman & Hall, New York, 1993.

FBSP. Anuário Brasileiro de Segurança Pública 2025. Fórum Brasileiro de Segurança Pública (FBSP), 19. ed., São Paulo, 2025.

Luz de Araujo, P. H., de Campos, T. E., de Oliveira, R. R. R., Stauffer, M., Couto, S., and Bermejo, P. LeNER-Br: A Dataset for Named Entity Recognition in Brazilian Legal Text. In *Computational Processing of the Portuguese Language – PROPOR 2018*. Lecture Notes in Computer Science, vol. 11122. Springer, Cham, pp. 313–323, 2018.

Oliveira, L. E. S. e. et al. SemClinBr – a Multi-Institutional and Multi-Specialty Semantically Annotated Corpus for Portuguese Clinical NLP Tasks. *Journal of Biomedical Semantics* 13 (1): 13, 2022.

Schiezaro, M., Rosa, G., Campos, B. A. G., and Pedrini, H. Guardians of the Data: NER and LLMs for Effective Medical Record Anonymization in Brazilian Portuguese. *Frontiers in Public Health* vol. 13, pp. 1717303, 2026.

Schneider, E. T. R., de Souza, J. V. A., Knafou, J., e Oliveira, L. E. S., Copara, J., Gumiel, Y. B., de Oliveira, L. F. A., Paraiso, E. C., Teodoro, D., and Barra, C. M. C. M. BioBERTpt – A Portuguese Neural Language Model for Clinical Named Entity Recognition. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*. Association for Computational Linguistics, Online, pp. 65–72, 2020.

Souza, F., Nogueira, R., and Lotufo, R. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In *Intelligent Systems – BRACIS 2020*. Lecture Notes in Computer Science, vol. 12319. Springer, Cham, pp. 403–417, 2020.

SSP-SP. Estatísticas de Criminalidade. https://www.ssp.sp.gov.br/estatistica, 2024.

Sweeney, L. k-Anonymity: A Model for Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10 (5): 557–570, 2002.

Uzuner, Ö., Luo, Y., and Szolovits, P. Evaluating the State-of-the-Art in Automatic De-identification. *Journal of the American Medical Informatics Association* 14 (5): 550–563, 2007.