Twenty Years of HAREM: A Reproducible Audit and Reassessment of Portuguese Named Entity Recognition

Rafael O. Nunes, André S. Spritzer, Carla M. D. S. Freitas and Dennis G. Balreira
Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil
{ronunes,spritzer,carla,dgbalreira}@inf.ufrgs.br

Abstract

For two decades, the HAREM corpus has served as the foundational benchmark for Portuguese Named Entity Recognition (NER), establishing its evaluation paradigm. Virtually all major progress has been measured against its fixed train/test split. This paper presents the first systematic audit of this split, revealing 153 overlapping (contaminated) sentences. We re-evaluate 13 NER models (ranging from CRFs to Transformers) on both the original and a new, decontaminated version of the corpus. Our statistical analysis reveals that decontamination has a significant ($p < 0.05$) and positive impact on the majority of models. We find that performance gains are most pronounced in the $F1_{\text{macro}}$ score (up to +4 points), demonstrating that the contamination primarily harmed generalization on rare entity types. Furthermore, our audit reveals clear evidence of overfitting in some models that benefited from data leakage. We conclude that even minor contamination can distort performance metrics and mask true model generalization. We release our decontaminated benchmark to ensure more reliable future evaluations.

1 Introduction

The HAREM evaluation campaign (Santos et al., 2006, 2007) introduced the first shared task for Named Entity Recognition (NER) in Portuguese. It established a unified taxonomy of named entities, detailed annotation guidelines, and two benchmark corpora, the First HAREM and the Mini-HAREM. Together, these resources defined a methodological foundation that has guided Portuguese NER research for nearly two decades (Albuquerque et al., 2023).

The impact of HAREM on the Portuguese NLP community is difficult to overstate. It provided the first large-scale manually annotated dataset for NER and became the default benchmark for model comparison, supporting progress from early neural models (dos Santos and Guimarães, 2015) to the advent of Transformer-based architectures (Souza et al., 2020; Tapajós et al., 2025). Despite these technological leaps, the same evaluation protocol, training on First HAREM and testing on Mini-HAREM, has remained virtually unchanged.

However, to the best of our knowledge, the canonical split between First HAREM and Mini-HAREM was never formally audited for data integrity. In particular, potential overlaps between the two subsets have never been systematically verified. Such contamination in the fine-tuning data may lead to inflated performance estimates, affecting model comparisons and reproducibility (Arango et al., 2022; Rücker and Akbik, 2023). Given the benchmark's historical importance, even small cross-set overlaps warrant careful examination.

This paper revisits the HAREM benchmark twenty years after its inception, providing the first systematic replication, audit, and statistically grounded reanalysis of Portuguese NER results based on this corpus. Our contributions are fourfold:

1. We perform a corpus-level audit of the traditional HAREM split, quantifying cross-partition overlap between First HAREM and Mini-HAREM;
2. We construct a *decontaminated* version of the corpora, removing duplicated instances;
3. We replicate a wide range of NER models, from CRF and embedding-based baselines to modern Transformer architectures, under both contaminated and clean conditions; and
4. We conduct a comprehensive evaluation based on statistical tests (Wilcoxon, Friedman, and Nemenyi) to assess the effect of decontamination on absolute and relative model performance.

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 351–359

Our results show that the canonical HAREM split contains measurable but limited contamination between training and test partitions. Removing these overlaps alters absolute performance for most models, often improving generalization, but does not disrupt the long-standing ranking hierarchy among architectures. Beyond empirical findings, this study provides a reproducible and statistically validated foundation for future Portuguese NER research, reaffirming HAREM's enduring significance while updating its methodological underpinnings for the next decade of evaluation.

2 Related Work

The First HAREM corpus (Santos et al., 2006, 2007), which encompasses both the HAREM and Mini-HAREM collections, has long been a canonical benchmark for Portuguese NER. Over the past two decades, successive studies have evaluated new architectures against this resource, shaping an evaluation paradigm whose methodological robustness has rarely been thoroughly examined.

Among the works that helped consolidate this paradigm, the one by dos Santos and Guimarães (2015) stands out for introducing the CharWNN architecture and adopting a fixed evaluation setup, where training is performed on the First HAREM corpus and testing is conducted on Mini-HAREM. This configuration proved influential and was subsequently adopted by most later studies.

Building on this setup, Quinta de Castro et al. (2018) evaluated an LSTM-CRF model under the same partition, reporting improved performance over CharWNN and establishing a new benchmark. Santos et al. (2019) followed a similar approach with a BiLSTM-CRF+Flair model, also reporting gains over previous results. In each case, the state-of-the-art claims were based on direct F1-score comparisons between models evaluated on the same static split.

The advent of Transformer-based models has not altered this evaluation tradition. Souza et al. (2020), introducing BERTimbau, explicitly adopted the same train/test configuration, comparing their results to the BiLSTM-CRF+Flair benchmark. Carmo et al. (2020) maintained this setup for PTT5, and more recent work, such as Tapajós et al. (2025), continued to evaluate models like XLM-RoBERTa under the same partition approach. Consequently, advances in Portuguese NER have largely relied on a single, static evaluation split, where training is performed on the First HAREM, and testing is conducted on Mini-HAREM.

This long-standing dependency highlights a critical methodological gap. Despite its central role in the field, the integrity of the canonical HAREM split, particularly regarding potential data contamination or overlap between training and test sets, has not, to our knowledge, been systematically examined. Similarly, statistical validation procedures, such as cross-validation or significance testing, have received limited attention in the literature. To the best of our knowledge, the present study offers the first systematic replication, corpus audit, and statistically grounded re-evaluation of the HAREM benchmark, twenty years after its inception.

3 Corpus Description

Our experiments rely on the First HAREM and Mini-HAREM corpora (Santos et al., 2006, 2007) in the *total scenario*[1] (Santos et al., 2019), which together constitute the primary benchmark for Portuguese NER. The First HAREM corpus comprises 4,640 sentences, while Mini-HAREM contains 3,339 sentences. We audited both corpora for internal duplication and cross-corpus overlap in order to quantify potential sources of data contamination in the canonical evaluation setup.

3.1 Internal duplication

An audit detected a small number of duplicated sentences within each corpus. In First HAREM, 15 sentences appear more than once; in Mini-HAREM, 25 sentences are duplicated. All duplicated instances found share identical BIO annotations and present no conflicting entity labels. Table 1 summarizes these counts.

Corpus	Duplicated sentences	Conflicting tags
First HAREM	15	0
Mini-HAREM	25	0

Table 1: Internal duplication in the HAREM corpora.

Most repetitions correspond to short or formulaic utterances (e.g., "sorry, your browser doesn't support Java (tm).", "(riso)", "não."). Although duplicates with identical annotations do not create label conflicts, repeated sentences can still influence aggregated metrics (for example, increasing the effective weight of repeated contexts). For this

[1] https://github.com/jneto04/ner-pt

reason, we report them and remove overlapping sentences when constructing the decontaminated evaluation split (see Sect. 3.2).

3.2 Overlap between First HAREM and Mini-HAREM

A cross-corpus comparison identified a set of overlapping sentences between First HAREM and Mini-HAREM. In total, we identified 153 overlapping sentences; of these, 142 appear in the training partition of First HAREM and 11 in its development partition. Table 2 reports the per-partition counts and a representative example.

Although the number of overlapping sentences is small relative to the corpus size, their presence creates a potential source of leakage under the widely used evaluation protocol (training on First HAREM and testing on Mini-HAREM). Even when annotations are identical, testing on sentences seen in training reduces the strictness of the evaluation. To promote reproducibility, we provide the full list of overlapping sentences and the decontamination script in the project repository[2].

3.3 Entity distribution and decontamination

We quantified how removing overlapping sentences ("decontamination") affects entity counts. Table 3 reports counts for First HAREM before and after decontamination, while Table 4 reports the same for Mini-HAREM, and Table 5 summarizes combined totals. In each table, the last column (Δ) indicates the absolute change.

The largest absolute reductions occur in highly frequent classes (LOC, PER, ORG), which is expected because frequent entity types are more likely to appear in overlapping sentences. Other categories showed minimal or no change. In relative terms, the reductions are modest (1–3%), but removing overlaps enforces stricter independence between train and test, and reduces the risk of inflated performance in comparative evaluations.

4 Experimental Evaluation

All experiments were executed on dedicated workstations. Transformer-based models were trained and evaluated on an NVIDIA GeForce RTX 4090 (24 GB GPU memory) with a system configuration of 64 GB RAM. BiLSTM–CRF experiments used an NVIDIA GeForce RTX 3060 (12 GB GPU memory). The CRF baseline was trained on CPU (Intel Core i7-10700). All neural models were implemented using the FLAIR framework [3] and PyTorch. The CRF baseline employed `sklearn_crfsuite`[4].

Optimization and training details. For BERT-like architectures, we used a learning rate of 2×10^{-5}, a weight decay of 0.01, and the Adam optimizer for 10 epochs, selecting the best checkpoint based on the micro-F1 score on the development data. The maximum sequence length was set to 512 tokens, with truncation and padding applied as necessary to ensure consistent input sizes.

The BiLSTM–CRF models were trained with a learning rate of 0.1, a batch size of 32, and up to 150 epochs. The CRF baseline used regularization parameters $c_1 = 0.1$ and $c_2 = 0.1$, with a maximum of 100 training iterations.

Evaluation protocol. To maintain comparability with previous Portuguese NER work, we followed the common protocol of training on the First HAREM corpus and reporting final holdout results on Mini-HAREM (Santos et al., 2007, 2019; Souza et al., 2020). In addition, to obtain more reliable performance estimates, we performed a stratified five-fold cross-validation (CV) on First HAREM. Folds were created at the sentence level using the multilabel stratification procedure of Sechidis et al. (2011), which preserves the distribution of categories of named entities in folds.

For each model, we: (i) run five-fold stratified CV on First HAREM and record per-fold metrics; (ii) use the per-fold results for statistical comparisons and model selection; and (iii) train a final model (on the full First HAREM) for a single evaluation on Mini-HAREM to facilitate direct comparison with prior work. All reported scores are micro/macro precision, recall, and F1, presented as means $\pm$ standard deviations across folds.

Statistical analysis. Our statistical testing followed a conservative and reproducible pipeline. For each model, we first compute the vector of five per-fold micro-F1 scores under both the contaminated and decontaminated conditions. To decide between parametric and nonparametric pairwise tests, we assess the normality of the paired differences (decontaminated $-$ contaminated) using the Shapiro–Wilk test (Shapiro and Wilk, 1965). In practice, the Shapiro–Wilk indicated non-

[2] Source code and corpus available at https://github.com/RafaelOleques/harem_decontaminated.

[3] https://github.com/flairNLP/flair

[4] https://sklearn-crfsuite.readthedocs.io/en/latest/

Overlap type	Count	Example
Train ∩ Mini-HAREM	142	“meu pai tinha diversos irmãos...”
Dev ∩ Mini-HAREM	11	“então , nós tínhamos clientes de o brasil inteiro...”
All overlaps (total)	153	“isso foi em julho de 1988”

Table 2: Overlap between First HAREM and Mini-HAREM.

Entity category	Before	After	Δ
ORG	900	881	-19
ABS	389	378	-11
TMP	434	425	-9
LOC	1,207	1,147	-60
VAL	461	458	-3
PER	974	962	-12
ACO	127	127	0
OBR	186	186	0
OTR	40	39	-1
COI	132	127	-5
Total	4,950	4,830	-120

Table 3: Entity distribution for First HAREM before and after decontamination. Δ denotes absolute change.

Entity category	Before	After	Δ
PER	813	793	-20
LOC	841	840	-1
TMP	348	344	-4
ORG	561	542	-19
OBR	187	187	0
ABS	196	196	0
VAL	325	325	0
COI	158	158	0
ACO	48	48	0
OTR	14	14	0
Total	3,481	3,437	-44

Table 4: Entity distribution for Mini-HAREM before and after decontamination.

normality for several models; therefore, we report the Wilcoxon signed-rank test (Wilcoxon, 1945) for paired comparisons of $F1_{\text{micro}}$ (per-model). To compare multiple models simultaneously, we apply the Friedman test (Friedman, 1937) across models (paired by fold), followed by the Nemenyi post-hoc (Nemenyi, 1963) procedure to identify statistically distinguishable groups.

All statistical analyses were implemented in Python using SciPy[5] and the `scikit-posthocs`[6] package. Exact p-values, test statistics, and per-fold score vectors are available in the project repository.

Models. We evaluated three categories of NER models representing different stages of the methodology. First, **transformers** capturing contextual dependencies, including Brazilian Portuguese models (**BERTimbau** base/large, **Albertina** 100m) (Souza et al., 2020; Rodrigues et al., 2023), multilingual models (**XLM-RoBERTa** base/large, **mBERT** base) (Conneau et al., 2020; Devlin et al., 2019), and **ModernBERT** (base/large) (Warner et al., 2024), which is predominantly trained on English data, as a baseline of a more recent encoder-only model. Second, **static embedding models** (Hartmann et al., 2017), specifically **NILC FastText** and **GloVe** (Skip-gram 300) integrated with a BiLSTM-CRF for sequential modeling. Finally, a **feature-based CRF** relying on handcrafted lexical and orthographic features without pretrained embeddings.

[5] https://scipy.org/

[6] https://scikit-posthocs.readthedocs.io/

Corpus	Before	After	Δ	% change
First HAREM	4,950	4,830	-120	-2.4%
Mini-HAREM	3,481	3,437	-44	-1.3%
Combined	8,431	8,267	-164	-1.9%

Table 5: Combined entity counts before and after decontamination.

5 Results

5.1 Overall Results and Comparative Analysis

Tables 6 and 7 present cross-validation results before and after decontamination. We report $F1_{\text{micro}}$ to assess overall performance and $F1_{\text{macro}}$ to evaluate robustness to underrepresented entity types.

Performance and Decontamination Effects. Large monolingual Transformers dominate both micro and macro metrics. Before decontamination, Albertina 900M BrWaC and 900M achieved top $F1_{\text{micro}}$ (77.35 and 76.65), followed by BERTimbau large and XLM-R large (≈ 75). Macro scores reveal sensitivity to rare entities: Albertina 900M BrWaC maintained 68.5%, while BERTimbau large and XLM-R large lagged at 62–61%, indicating that contamination suppressed learning on infrequent classes.

After decontamination, BERTimbau large surpasses Albertina 900M BrWaC in micro F1 (77.66 vs. 77.45), demonstrating that even minor contamination can affect SOTA rankings. Macro-level

Model	P_{micro}	R_{micro}	$F1_{\text{micro}}$	P_{macro}	R_{macro}	$F1_{\text{macro}}$
Albertina (900M BrWaC)	77.44 ± 1.88	77.25 ± 2.34	77.35 ± 2.10	69.86 ± 3.94	68.55 ± 4.47	68.51 ± 3.55
Albertina (900M)	77.11 ± 1.45	76.21 ± 2.49	76.65 ± 1.76	69.91 ± 1.76	67.55 ± 2.11	68.06 ± 1.59
BERTimbau (large)	74.96 ± 2.61	76.24 ± 1.85	75.57 ± 1.53	62.83 ± 4.92	63.12 ± 4.02	62.53 ± 3.55
XLM-R (large)	74.72 ± 1.55	75.47 ± 1.99	75.08 ± 1.19	62.30 ± 2.14	61.31 ± 3.53	61.15 ± 2.26
BERTimbau (base)	72.97 ± 1.97	74.42 ± 2.49	73.66 ± 1.55	58.86 ± 2.10	59.77 ± 2.91	59.08 ± 1.91
XLM-R (base)	70.90 ± 1.43	71.69 ± 3.56	71.27 ± 2.42	56.28 ± 2.03	55.73 ± 2.86	55.52 ± 2.40
Albertina (100M)	71.28 ± 4.15	71.16 ± 2.77	71.20 ± 3.21	58.45 ± 4.20	58.43 ± 4.05	57.93 ± 4.02
mBERT (base)	68.82 ± 2.09	70.69 ± 2.86	69.74 ± 2.42	54.93 ± 3.18	55.67 ± 4.16	54.93 ± 3.71
GloVe (Skip-Gram 300)	70.96 ± 0.98	55.30 ± 4.48	62.08 ± 3.00	54.95 ± 4.08	39.44 ± 4.06	43.73 ± 3.84
ModernBERT (large)	62.75 ± 2.21	58.43 ± 3.09	60.45 ± 1.75	48.77 ± 3.75	45.29 ± 2.46	45.83 ± 1.66
NILC FastText (Skip-Gram 100)	70.26 ± 1.74	49.86 ± 4.03	58.28 ± 3.18	50.91 ± 10.26	32.90 ± 2.90	36.20 ± 2.90
CRF	67.81 ± 2.82	46.10 ± 2.83	54.87 ± 2.78	56.17 ± 6.06	34.66 ± 3.24	41.25 ± 4.19
ModernBERT (base)	46.67 ± 1.32	41.68 ± 1.45	44.02 ± 0.99	32.02 ± 5.93	29.23 ± 1.30	29.16 ± 1.44

Table 6: Cross-validation results **before decontamination** (mean ± standard deviation).

Model	P_{micro}	R_{micro}	$F1_{\text{micro}}$	P_{macro}	R_{macro}	$F1_{\text{macro}}$
BERTimbau (large)	77.27 ± 0.60	78.07 ± 1.99	77.66 ± 1.22	67.05 ± 3.31	66.93 ± 4.49	66.50 ± 3.87
Albertina (900M BrWaC)	78.28 ± 2.95	76.65 ± 3.69	77.45 ± 3.20	72.76 ± 4.50	67.75 ± 6.51	69.17 ± 4.91
XLM-R (large)	74.73 ± 1.91	76.58 ± 3.09	75.64 ± 2.41	67.65 ± 6.05	65.80 ± 4.19	65.18 ± 3.99
BERTimbau (base)	72.76 ± 2.49	75.58 ± 2.10	74.13 ± 2.02	59.31 ± 3.16	60.65 ± 2.39	59.58 ± 2.45
Albertina (100M)	73.52 ± 1.83	73.47 ± 1.97	73.49 ± 1.87	61.39 ± 2.86	59.91 ± 2.27	60.30 ± 2.27
Albertina (900M)	74.19 ± 11.17	70.80 ± 11.16	72.39 ± 10.89	67.44 ± 19.19	60.91 ± 15.62	62.96 ± 17.16
XLM-R (base)	71.16 ± 3.04	72.38 ± 2.39	71.73 ± 2.18	55.43 ± 2.87	56.11 ± 3.46	55.43 ± 2.80
mBERT (base)	70.11 ± 2.08	72.03 ± 2.77	71.05 ± 2.21	58.88 ± 3.95	58.94 ± 3.80	58.15 ± 2.83
GloVe (Skip-Gram 300)	71.05 ± 2.72	59.49 ± 2.48	64.75 ± 2.56	55.77 ± 1.35	43.66 ± 0.64	47.51 ± 0.80
ModernBERT (large)	66.51 ± 2.48	60.64 ± 4.87	63.38 ± 3.46	49.09 ± 2.34	45.50 ± 4.77	46.74 ± 3.84
NILC FastText (Skip-Gram 100)	71.33 ± 3.36	52.86 ± 1.47	60.70 ± 1.84	55.51 ± 2.85	35.64 ± 1.84	39.36 ± 1.78
CRF	70.06 ± 3.11	49.07 ± 2.32	57.69 ± 2.31	62.59 ± 3.35	37.22 ± 1.67	44.40 ± 1.31
ModernBERT (base)	48.23 ± 2.34	44.33 ± 2.42	46.19 ± 2.36	33.55 ± 3.83	30.61 ± 1.74	30.74 ± 1.79

Table 7: Cross-validation results **after decontamination** (mean ± standard deviation).

gains are more pronounced: BERTimbau large (+3.97) and XLM-R large (+4.03) improved significantly, while Albertina 900M BrWaC remained stable (+0.66). Baselines (CRF, GloVe, FastText) remain limited in macro ($\approx$ 40–47%), highlighting structural constraints in handling long-tail distributions.

Synthesis. These results reveal three key phenomena: (i) large contextual Transformers are robust and top-performing, (ii) minor data contamination can distort evaluation, particularly for rare entities, and (iii) macro-averaged metrics provide a crucial lens for assessing generalization and fairness across classes.

5.2 Statistical Impact of Decontamination

To evaluate whether the observed changes between the "contaminated" and "decontaminated" runs were statistically significant, we conducted a paired Wilcoxon signed-rank test. The test compares the $F1_{\text{micro}}$ of each model across the five cross-validation folds (see Table 8).

Interpretation. The Wilcoxon test reveals that decontamination had a statistically significant impact ($p < 0.05$) on the majority of evaluated models (9 out of 13).

Crucially, the nature of this impact is a consistent improvement in performance. Eight of the nine models with significant differences (including CRF, GloVe, FastText, mBERT, and BERTimbau-large) showed higher F1-scores after data cleaning. This strongly suggests that the contaminated data was not just noise but acted as a detriment to generalization, and its removal allowed the models to learn more robust patterns.

The only exception to this trend is Albertina (900M), which exhibited a statistically significant performance drop ($\Delta = -4.26$, $p = 0.0089$). This result suggests that the model benefited from the data leakage (overfitting on the contaminated ex-

Model	Cont.	Decont.	Δ	p-value
BERTimbau (large)*	75.57	77.66	+2.09	0.000386
Albertina (900M BrWaC)	77.35	77.45	+0.10	0.278546
XLM-R (large)*	75.08	75.64	+0.56	0.002797
BERTimbau (base)	73.66	74.13	+0.47	0.371463
Albertina (100M)*	71.20	73.49	+2.29	0.024967
Albertina (900M)*	76.65	72.39	-4.26	0.008908
XLM-R (base)	71.27	71.73	+0.46	0.585780
mBERT (base)*	69.74	71.05	+1.31	0.001151
GloVe (Skip-Gram 300)*	62.08	64.75	+2.67	0.000621
ModernBERT (large)	60.45	63.38	+2.93	0.445496
NILC FastText (Skip-Gram 100)*	58.28	60.70	+2.42	0.000322
CRF*	54.87	57.69	+2.82	0.000138
ModernBERT (base)*	44.02	46.19	+2.17	0.025984

Table 8: Paired Wilcoxon signed-rank test comparing contaminated and decontaminated runs ($F1_{\text{micro}}$). Significant ($p < 0.05$) results are marked with *.

amples), and decontamination shows a lower generalization.

In summary, the Wilcoxon analysis contradicted the hypothesis of a null effect. Decontamination causes real statistical changes in the results, mostly for the better, and exposes cases of overfitting due to contamination.

5.3 Relative Ranking Analysis

While the Wilcoxon test confirmed the impact on absolute performance, we used the Friedman test followed by the Nemenyi post-hoc test to evaluate whether this impact was sufficient to alter the relative ranking and hierarchy among model families. This analysis compares the average rank of all models across the five cross-validation folds, identifying statistically equivalent performance groups.

Analysis of rankings. In both settings (before and after decontamination), the Friedman test rejected the null hypothesis ($p < 0.001$), confirming that the performance differences among models were not due to chance.

The Nemenyi post-hoc test reveals a highly stable ranking structure.

- **Before decontamination:** The elite group (Albertina, BERTimbau) was statistically superior to the baselines (CRF, GloVe, FastText) and the major English model (ModernBERT).
- **After decontamination:** This same hierarchical structure is preserved. The top-tier Transformers (Albertina, BERTimbau) continue to form a statistically superior group. The baselines (CRF, GloVe, FastText) remain in the lowest performance group, significantly worse than the Transformer models.

Interpretation. The Nemenyi test demonstrates that although decontamination significantly improved the absolute score of many models (as seen in the Wilcoxon test), these improvements were not sufficient to alter the fundamental hierarchy of the architectures.

The baselines improved, but not enough to catch up to the Transformers. The top-tier Transformers (Albertina, BERTimbau) remain statistically indistinguishable from each other as a group, even as their absolute scores shifted.

We conclude that data contamination in HAREM did not distort the relative evaluation of model families. The superiority of monolingual Transformers over multilingual ones and over static baselines is a robust finding, valid in both the original and the decontaminated corpora.

5.4 Holdout Evaluation

We conducted a holdout evaluation to complement cross-validation, using a fixed train–test split to assess generalization on unseen data. Tables 9 and 10 report precision, recall, and F1 for both micro and macro averages.

Impact of Decontamination. Across models, decontamination consistently improves performance, with an average gain of approximately +2.3 F1 points (Table 11). Large monolingual Transformers (*Albertina 900M BrWaC* and *BERTimbau large*)

Model	P_{micro}	R_{micro}	$F1_{\text{micro}}$	P_{macro}	R_{macro}	$F1_{\text{macro}}$
Albertina (900M BrWaC)	76.13	76.97	76.55	73.28	69.98	71.08
XLM-R (large)	74.30	75.44	74.86	71.58	64.06	64.24
BERTimbau (large)	73.02	76.53	74.73	66.65	64.79	64.11
Albertina (900M)	76.46	72.71	74.54	73.96	66.41	66.42
Albertina (100M)	71.50	74.24	72.84	59.39	65.53	61.89
BERTimbau (base)	69.40	73.03	71.17	55.10	57.70	56.27
XLM-R (base)	69.25	72.27	70.73	56.37	58.14	56.91
mBERT (base)	68.24	71.07	69.63	53.03	56.95	54.65
ModernBERT (large)	65.09	64.74	64.92	60.66	54.76	55.21
GloVe (Skip-Gram 300)	66.01	58.73	62.16	54.50	45.03	47.93
NILC FastText (Skip-Gram 100)	70.33	51.75	59.62	48.10	35.32	38.86
CRF	66.57	47.82	55.65	56.65	35.51	42.03
ModernBERT (base)	48.41	43.23	45.67	32.63	29.76	30.48

Table 9: Holdout results before decontamination (contaminated dataset).

Model	P_{micro}	R_{micro}	$F1_{\text{micro}}$	P_{macro}	R_{macro}	$F1_{\text{macro}}$
Albertina (900M BrWaC)	78.98	79.59	79.28	77.63	72.74	74.25
BERTimbau (large)	77.29	80.24	78.74	73.80	75.40	74.00
XLM-R (large)	76.36	78.28	77.30	71.96	74.03	72.62
Albertina (900M)	76.03	78.28	77.14	75.12	74.74	74.50
Albertina (100M)	73.38	74.34	73.86	70.89	64.46	63.84
BERTimbau (base)	71.81	75.66	73.68	59.52	64.24	61.71
XLM-R (base)	67.16	73.69	70.28	54.86	58.90	56.37
mBERT (base)	69.85	70.31	70.08	68.44	58.50	59.75
GloVe (Skip-Gram 300)	71.25	62.77	66.74	69.25	49.43	53.60
ModernBERT (large)	65.61	65.83	65.72	51.27	50.41	50.58
NILC FastText (Skip-Gram 100)	71.63	50.44	59.19	42.77	32.69	35.35
CRF	67.53	51.09	58.17	58.87	38.49	44.30
ModernBERT (base)	49.32	47.49	48.39	33.97	33.65	33.49

Table 10: Holdout results after decontamination (clean dataset).

achieve the highest micro F1 (79.28 and 78.74) and substantial macro gains (+3.17 and +9.89), confirming that removing train–test overlap enhances generalization, particularly for underrepresented entities.

Multilingual Transformers (XLM-R large and mBERT) show moderate improvements (+2.44 and +0.45 micro F1), while smaller or base variants (XLM-R base, NILC FastText) exhibit minimal or slightly negative changes, reflecting capacity limitations. Baselines (CRF, GloVe) benefit from decontamination (+2.5–4.6 F1) due to reduced noise from contaminated instances, but remain substantially below top models, highlighting structural constraints in handling long-tail entity distributions.

Macro-level Insights. Macro F1 accentuates the effect of contamination on rare classes. Before decontamination, even high-performing Transformers show reduced macro performance (BERTimbau large 64.11, XLM-R large 64.24). After decontamination, macro scores increase significantly (BERTimbau large 74.00, XLM-R large 72.62), demonstrating that leakage disproportionately suppresses the learning of infrequent entities. Baselines remain low ($\approx$ 35–53%), indicating limited representational capacity regardless of data quality.

Synthesis. The holdout results corroborate cross-validation findings: (i) large contextual Transformers are robust and top-performing, (ii) minor contamination can substantially distort evaluation, especially for rare entities, and (iii) macro-averaged metrics provide a sensitive measure of generalization and fairness, capturing effects that micro-averaged scores alone may obscure.

Model	Cont.	Decont.	Δ
Albertina (900M BrWaC)	76.55	79.28	+2.73
BERTimbau (large)	74.73	78.74	+4.01
XLM-R (large)	74.86	77.30	+2.44
Albertina (900M)	74.54	77.14	+2.60
Albertina (100M)	72.84	73.86	+1.02
BERTimbau (base)	71.17	73.68	+2.51
XLM-R (base)	70.73	70.28	-0.45
mBERT (base)	69.63	70.08	+0.45
GloVe (Skip-Gram 300)	62.16	66.74	+4.58
ModernBERT (large)	64.92	65.72	+0.80
NILC FastText (Skip-Gram 100)	59.62	59.19	-0.43
CRF	55.65	58.17	+2.52
ModernBERT (base)	45.67	48.39	+2.72

Table 11: Holdout comparison of $F1_{\text{micro}}$ before and after decontamination.

6 Conclusion

This paper presents the first systematic audit and reproducibility study of the canonical HAREM benchmark, twenty years after its creation. We identified 153 overlapping sentences between First HAREM (train) and Mini-HAREM (test) and evaluated 13 NER models, from CRFs to large Transformers, on both the original and a decontaminated version to measure the impact of this long-standing data leakage.

Our results show that even limited contamination has a statistically significant and practical effect on evaluation:

- **Decontamination matters:** Removing contaminated instances led to significant performance changes for 9 out of 13 models ($p < 0.05$).
- **Contamination hinders generalization:** Most models improved (+2.7 to +4.0 $F1_{\text{micro}}$) after cleaning, indicating the leakage actively harmed model learning.
- **Infrequent classes are most affected:** Top models gained +4 $F1_{\text{macro}}$, showing contamination primarily impacted underrepresented entity types.
- **SOTA rankings shift:** BERTimbau (large) surpassed Albertina (900M) on clean data, establishing a new state-of-the-art.
- **Evidence of overfitting:** Albertina's performance dropped post-cleaning, confirming it had benefited from data leakage.
- **Architecture hierarchy is stable:** Large Transformers remain superior over baselines despite shifts in absolute scores.

In summary, our contributions are twofold: we released a reproducible, decontaminated HAREM benchmark with robust baselines, and we showed that even "minor" contamination can distort evaluation, particularly for rare classes and SOTA rankings. This reassessment strengthens the benchmark's reliability for future Portuguese NER research.

As future work, we plan to extend this auditing methodology to other Portuguese corpora and analyze which rare entity types are most affected by contamination.

Acknowledgements

This work has been partially funded by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. We also acknowledge the financial support from the Brazilian funding agency CNPq and Petrobras. Some experiments in this work used the PCAD infrastructure, http://gppd-hpc.inf.ufrgs.br, at INF/UFRGS.

References

Hidelberg O Albuquerque, Ellen Souza, Carlos Gomes, Matheus Henrique de C Pinto, Ricardo PS Filho, Rosimeire Costa, Vinícius Teixeira de M Lopes, Nádia FF da Silva, André CPLF de Carvalho, and Adriano LI Oliveira. 2023. Named entity recognition: a survey for the portuguese language. *Procesamiento del Lenguaje Natural*, 70:171–185.

Aymé Arango, Jorge Pérez, and Barbara Poblete. 2022. Hate speech detection is not as easy as you may think: A closer look at model validation (extended version). *Information Systems*, 105:101584.

Diedre Carmo, Marcos Piau, Israel Campiotti, Rodrigo Nogueira, and Roberto Lotufo. 2020. Ptt5: Pretraining and validating the t5 model on brazilian portuguese data. *arXiv preprint arXiv:2008.09144*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Cicero dos Santos and Victor Guimarães. 2015. Boosting named entity recognition with neural character embeddings. In *Proceedings of the Fifth Named Entity Workshop*, pages 25–33.

Milton Friedman. 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200):675–701.

Nathan Hartmann, Erick Fonseca, Christopher Shulby, Marcos Treviso, Jessica Rodrigues, and Sandra Aluisio. 2017. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. *arXiv preprint arXiv:1708.06025*.

Peter Bjorn Nemenyi. 1963. *Distribution-free multiple comparisons.* Princeton University.

Pedro Vitor Quinta de Castro, Nádia Félix Felipe da Silva, and Anderson da Silva Soares. 2018. Portuguese named entity recognition using lstm-crf. In *International Conference on Computational Processing of the Portuguese Language*, pages 83–92. Springer.

João Rodrigues, Luís Gomes, João Silva, António Branco, Rodrigo Santos, Henrique Lopes Cardoso, and Tomás Osório. 2023. Advancing neural encoding of portuguese with transformer albertina pt. In *EPIA Conference on Artificial Intelligence*, pages 441–453. Springer.

Susanna Rücker and Alan Akbik. 2023. Cleanconll: A nearly noise-free named entity recognition dataset. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8628–8645.

Diana Santos, Nuno Cardoso, Nuno Seco, and Rui Vilela. 2007. Breve introdução ao harem. *HAREM, a primeira avaliaçao conjunta de sistemas de reconhecimento de entidades mencionadas para português: documentaçao e actas do encontro, Linguateca*.

Diana Santos, Nuno Seco, Nuno Cardoso, and Rui Vilela. 2006. Harem: An advanced ner evaluation contest for portuguese. In *quot; In Nicoletta Calzolari; Khalid Choukri; Aldo Gangemi; Bente Maegaard; Joseph Mariani; Jan Odjik; Daniel Tapias (ed) Proceedings of the 5 th International Conference on Language Resources and Evaluation (LREC'2006)(Genoa Italy 22-28 May 2006)*.

Joaquim Santos, Bernardo Consoli, Cicero dos Santos, Juliano Terra, Sandra Collonini, and Renata Vieira. 2019. Assessing the impact of contextual embeddings for portuguese named entity recognition. In *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 437–442. IEEE.

Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. 2011. On the stratification of multi-label data. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011, Proceedings, Part III 22*, pages 145–158. Springer.

Samuel Sanford Shapiro and Martin B Wilk. 1965. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4):591–611.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: pretrained bert models for brazilian portuguese. In *Brazilian conference on intelligent systems*, pages 403–417. Springer.

Guilherme Tapajós, Tiago de Melo, Elloá B Guedes, and Fábio Santos. 2025. Reconhecimento de entidades nomeadas em português: Comparação de modelos pré-treinados com fine-tuning. *Revista Eletrônica de Iniciação Científica em Computação*, 23:111–117.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, and 1 others. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*.

F Wilcoxon. 1945. Individual comparisons by ranking methods. biom. bull., 1, 80–83.