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.

A Leakage-Controlled Benchmark for Nya-Həba Named Entity Recognition: Statistical, Neural, and Cost-Sensitive Transformer Baselines

Chahyaandida Ishaya

m.scdsa241243@student.mau.edu.ng

Modibbo Adama University Yola

Ayuba Madu Yami

Modibbo Adama University Yola

Ndachimya Magaji Edward

Okan University

Manunyi Godfrey

Modibbo Adama University Yola

Research Article

Keywords: Nya-Həba, named entity recognition, less-resourced languages, African NLP, class imbalance, XLM-R

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11019985/v1

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

Additional Declarations: No competing interests reported.

A Leakage-Controlled Benchmark for Nya-Həba Named Entity Recognition: Statistical, Neural, and Cost-Sensitive Transformer Baselines

Full-length paper — Special Focus: Less-resourced Languages

Chahyaandida Ishaya[1]*, Ayuba Madu Yami[1], Ndachimya Magaji Edward[2], and Manunyi Godfrey[3]

[1]Department of Data Science and Analytics, Faculty of Computing, Modibbo Adama University, Yola, Nigeria

[2]Istanbul Okan University, Istanbul, Türkiye

[3]Modibbo Adama University, Yola, Nigeria

Corresponding author: Chahyaandida Ishaya, m.scdsa241243@student.mau.edu.ng

Co-author emails: Ayuba Madu Yami — amyami@mau.edu.ng; Ndachimya Magaji Edward — nedward@stu.okan.edu.tr (personal: ndachimya@gmail.com); Manunyi Godfrey — m.sccsc251856@student.mau.edu.ng

ORCID — Chahyaandida Ishaya: 0009-0003-5031-7292

ORCID — Ayuba Madu Yami: 0000-0003-0061-8356

ORCID — Ndachimya Magaji Edward: 0009-0002-1046-2543

ORCID — Manunyi Godfrey: 0009-0005-3798-7923

Abstract

Named entity recognition (NER) for extremely low-resource languages is constrained not only by limited labelled data, but also by duplicate-text leakage, annotation inconsistency, severe class imbalance, and uncertain transfer from multilingual pretrained models. This study establishes a leakage-controlled evaluation benchmark for Nya-Həba (Kilba), a Chadic language of Adamawa State, Nigeria, and compares statistical, neural, and transformer-based NER baselines. The strict benchmark contains 1,853 unique, consistently annotated sentences (24,239 tokens; 2,277 entity spans) with fixed train/validation/test partitions and zero exact sentence-text overlap. A feature-based conditional random field (CRF), BiLSTM-CRF, XLM-R, and training-only class-weighted XLM-R were evaluated with exact-span entity metrics. CRF achieved the strongest overall micro-F1 (69.03%), followed by BiLSTM-CRF (63.14 ± 1.75%) and vanilla XLM-R (60.00 ± 1.95%). Class weighting did not improve overall micro-F1 (58.32 ± 1.57%), but increased recall from 53.03% to 67.72% and macro-F1 from 32.04% to 50.74%; TIME F1 rose from 4.60% to 36.80% and ANIM from 0.00% to 42.24%. Error analysis shows that this rare-class recovery is accompanied by substantially more spurious entities. Paired sentence-level bootstrap intervals for weighted-versus-vanilla XLM-R micro-F1 cross zero for all seeds. The results show that model complexity does not guarantee stronger NER performance under extreme resource scarcity, while cost-sensitive training can recover minority categories at a precision cost. The benchmark is publicly available on Zenodo.

Keywords: Nya-Həba; named entity recognition; less-resourced languages; African NLP; class imbalance; XLM-R

Acknowledgments

The authors acknowledge the academic support of the Department of Data Science and Analytics, Faculty of Computing, Modibbo Adama University, Yola, Nigeria. The first author gratefully remembers the late Prof. Hammadiko Mu'azu Gaya for his early academic guidance during the development of this work.

1 Introduction

Named entity recognition (NER) is a core information-extraction task in which text spans are located and assigned task-defined semantic categories. Reliable NER supports search, question answering, information retrieval, knowledge-base construction, public-interest text analytics, and many other downstream applications. Much of the standard NER literature, however, has been built around a relatively small set of well-resourced languages and benchmark domains. CoNLL-style evaluation helped standardize span-based NER comparison (Tjong Kim Sang & De Meulder, 2003), but the availability of similarly usable training and evaluation resources remains highly uneven across the world's languages (Joshi et al., 2020; Hedderich et al., 2021).

African-language NER has advanced substantially through efforts such as MasakhaNER and MasakhaNER 2.0, which created annotated datasets and evaluated supervised and cross-lingual transfer methods across multiple African languages (Adelani et al., 2021, 2022). Related work has shown that Africa-focused pretraining and adaptation can be competitive with global multilingual models, particularly when language coverage and transfer relations are better aligned with the target setting (Ogueji et al., 2021; Alabi et al., 2022; Adebara et al., 2023). These developments demonstrate both the value of multilingual transfer and the continuing importance of language-specific resources. Thousands of African languages nevertheless remain absent from commonly used NLP benchmarks, and the absence is especially pronounced for languages with limited digitized text, orthographic standardization, or prior computational annotation.

Nya-Həba, commonly known as Kilba, is one such language. It is a Chadic language associated primarily with Adamawa State, Nigeria, and has descriptive linguistic documentation but very limited reusable NLP infrastructure (Mu'azu, 2009). The Nya-Həba Named Entity Recognition Corpus v1.0 was released to provide a machine-readable annotated resource for the language (Ishaya & Yami, 2026). The archival release contains 3,184 sentence instances and 44,308 tokens annotated with BIO labels for Person (PER), Location (LOC), Time (TIME), and a corpus-specific Animal (ANIM) category. A subsequent audit identified substantial exact sentence repetition and a smaller set of repeated texts carrying inconsistent BIO sequences. Rather than allowing those properties to silently influence evaluation, the release provides a strict benchmark of unique, consistently annotated sentences with fixed leakage-controlled partitions.

The distinction between an archival corpus and an evaluation benchmark is methodologically important. Training/test overlap can inflate apparent NLP performance by rewarding memorization rather than generalization, including in NER settings (Elangovan et al., 2021). More broadly, dataset composition, label distribution, and distribution shift can materially alter NER performance and therefore the conclusions drawn from model comparisons (Li & Groth, 2025). In extremely small corpora, even a modest number of repeated sentences can have an outsized effect. A useful benchmark should therefore make identity criteria, split construction, annotation conflicts, and class supports explicit rather than treating data preparation as an invisible preprocessing step.

A second methodological challenge is model selection. Contemporary NLP often defaults to multilingual pretrained transformers because they provide powerful cross-lingual representations (Conneau et al., 2020). Yet less-resourced settings can violate assumptions that make large pretrained models successful: labelled supervision may be very small, the target language may have little direct representation in pretraining corpora, orthographic and lexical variation may be high, and rare labels may receive too few updates during fine-tuning. Classical feature-based models can remain competitive when local lexical and orthographic cues are strong, while compact neural sequence models offer a middle ground between manually specified features and large-scale pretraining. These contrasts motivate direct cross-paradigm evaluation rather than assuming a priori that the most complex architecture will dominate.

Class imbalance is especially severe in the Nya-Həba benchmark. In the training partition, the non-entity O label accounts for 15,187 of 16,893 tokens, while the rare B-ANIM and I-PER labels occur only 66 and 24 times, respectively. The problem is not unique to this dataset: cost-sensitive learning has been proposed for NER when ordinary fine-tuning underweights minority entity categories (Mobasher et al., 2023). For a very small benchmark, however, increasing the contribution of rare classes may improve recall while simultaneously producing more false positives. Both micro-averaged and macro-averaged evaluation are therefore necessary to characterize the trade-off.

This paper uses the strict Nya-Həba benchmark to address four research questions. RQ1 asks how a feature-based CRF, a BiLSTM-CRF, and XLM-R compare under a leakage-controlled split. RQ2 asks how class imbalance affects category-specific performance, particularly for TIME and ANIM. RQ3 asks whether training-only cost-sensitive weighting can recover rare categories during XLM-R fine-tuning and what precision/recall trade-off it introduces. RQ4 asks which error types dominate each model and what those errors imply for future resource and model development.

The study makes four contributions. First, it establishes a reproducible evaluation protocol around the public strict benchmark, in which exact sentence-text overlap across partitions is zero. Second, it provides statistical, neural, and multilingual-transformer baselines evaluated with exact-span metrics on the same test set. Third, it adds a targeted cost-sensitive XLM-R experiment derived exclusively from training-label frequencies, allowing minority-class recovery to be studied without using test information for weighting. Fourth, it complements aggregate scores with per-class analysis, an explicit error taxonomy, and paired sentence-level bootstrap intervals. Together, these elements turn the corpus from a descriptive language resource into an auditable benchmark for less-resourced NER research.

2 Related work

2.1 NER resources and models for African and less-resourced languages

Early NER systems frequently relied on statistical sequence labelling and manually designed lexical, orthographic, and contextual features. Conditional random fields (CRFs) became a standard discriminative approach for sequence labelling because they model dependencies between adjacent output labels while supporting rich overlapping features (Lafferty et al., 2001). Neural architectures later reduced dependence on hand-crafted features. In particular, BiLSTM-CRF models combine contextual recurrent representations with structured sequence decoding and became a strong pre-transformer baseline for NER (Lample et al., 2016). These two paradigms remain useful reference points in less-resourced settings because they differ in how they consume small amounts of supervision.

Multilingual pretrained language models changed the expected baseline for cross-lingual NLP. XLM-R was pretrained on one hundred languages and demonstrated strong transfer across multiple cross-lingual tasks, including NER (Conneau et al., 2020). The success of multilingual pretraining does not remove resource disparities, however. Cross-lingual transfer quality depends on pretraining coverage, script, lexical and typological relationships, and the amount and quality of target supervision. The broader low-resource NLP literature therefore treats transfer learning as one component of a resource strategy rather than a universal substitute for target-language data (Hedderich et al., 2021).

MasakhaNER provided a major empirical foundation for African-language NER by releasing human-annotated datasets for ten languages and comparing supervised and transfer-learning methods (Adelani et al., 2021). MasakhaNER 2.0 expanded the resource to twenty African languages and showed that source-language choice can substantially affect transfer performance (Adelani et al., 2022). The finding is particularly relevant to languages outside widely studied benchmarks: transfer cannot be reduced to a simple high-resource-to-low-resource recipe, and Africa-centric language relations can matter.

Other Africa-focused multilingual modelling work reinforces this point. AfriBERTa showed that a comparatively small pretraining corpus focused on African languages can be competitive with larger multilingual models on NER and text classification (Ogueji et al., 2021). Multilingual adaptive fine-tuning has also been used to improve pretrained-model alignment to African languages while retaining cross-lingual transfer (Alabi et al., 2022). SERENGETI extended the scale of Africa-oriented multilingual modelling and highlighted the continuing underrepresentation of African languages in general-purpose pretrained models (Adebara et al., 2023). These studies motivate transformer evaluation for Nya-Həba, while also cautioning against interpreting a single general multilingual model as an upper bound for the language.

To the best of our knowledge, the literature reviewed for this study did not identify a previously published NER benchmark specifically for Nya-Həba. The contribution here is therefore not a new sequence-labelling algorithm. It is the first systematic application and comparison, within the released Nya-Həba benchmark, of established statistical, neural, and multilingual-transformer approaches, together with an imbalance-aware transformer variant. This

positioning is consistent with the need for strong first baselines in languages where reusable annotated resources have only recently become available.

2.2 Evaluation integrity, leakage, and dataset documentation

Language-resource evaluation depends on more than model architecture. Dataset documentation practices have increasingly emphasized explicit provenance, intended use, representativeness, and known limitations so that experimental claims can be interpreted in context (Bender & Friedman, 2018; Gebru et al., 2021). For NER, evaluation is additionally sensitive to entity definitions, span conventions, label supports, domain composition, and the relation between training and test text. Gritta et al. (2020), for example, show how task definition and evaluation choices affect conclusions in location-oriented NER and geoparsing.

Leakage is a direct threat to benchmark validity. Elangovan et al. (2021) demonstrate that overlap between NLP training and test data can inflate measured performance and blur the distinction between memorization and generalization. Recent work has broadened the concern to contamination in modern NLP benchmarks (Sainz et al., 2023). The problem is especially important when a corpus contains repeated sentence templates or recycled material: random instance-level splitting can place identical text in multiple partitions even when the split appears formally disjoint at the record level.

A related issue is distribution shift. Li and Groth (2025) systematically examine input and label shifts across NER datasets and show that changes in distribution can produce large performance differences. The Nya-Həba study does not estimate cross-domain shift because genre metadata are incomplete, but it adopts the same general principle that label and text distributions should be visible rather than implicit. The strict benchmark therefore removes exact duplicate-text leakage and reports the category supports on which every score is based.

2.3 Class imbalance and cost-sensitive fine-tuning

NER datasets are intrinsically imbalanced at token level because most tokens fall outside named entities. The imbalance becomes more severe when some entity types are rare. A model optimized with ordinary cross-entropy can achieve high token accuracy while neglecting minority categories, particularly when the O label dominates gradient updates. For this reason, entity-level span F1 is conventionally more informative than raw token accuracy for NER, and per-class measures are necessary when entity supports differ sharply.

Cost-sensitive learning addresses imbalance by assigning different training losses to classes. Mobasher et al. (2023) demonstrate that reweighted fine-tuning can improve recognition of underrepresented biomedical entities, although the precise weighting strategy and domain differ from the present work. We use the same general cost-sensitive principle but derive a deliberately conservative square-root inverse-frequency weighting from Nya-Həba training data only. The experiment is framed as a diagnostic: the goal is to determine whether the rare TIME and ANIM categories become learnable for XLM-R and to quantify any accompanying precision loss, not to claim a new general weighting algorithm.

3 Nya-Həba corpus and leakage-controlled benchmark

3.1 Language and resource context

Nya-Həba (Kilba) is a Chadic language used primarily in Adamawa State, Nigeria. A descriptive grammar is available (Mu'azu, 2009), but digital NLP resources remain sparse. The Nya-Həba Named Entity Recognition Corpus v1.0 was created to provide a structured resource for token classification and is publicly deposited on Zenodo (Ishaya & Yami, 2026). The present paper does not create a separate hidden benchmark; it evaluates the public strict benchmark distributed with that release.

The corpus was assembled from pre-existing Nya-Həba materials available to the first author, including oral-narrative transcriptions, transcribed conversations, locally produced texts, and other available language materials. No new participant recruitment, interviews, or recordings were conducted specifically for this benchmark study or the public

release. Source passages were segmented into sentence-level units, conservatively cleaned, and Unicode-normalized while preserving Nya-Həba orthographic variation.

3.2 Annotation and entity inventory

The first author manually assigned BIO labels in Label Studio. The inventory comprises PER, LOC, TIME, ANIM, and O. PER covers person names and person-like named references; LOC covers place names and annotated geographical/location expressions; TIME covers annotated temporal expressions. ANIM is a corpus-specific semantic category for animal references selected during annotation. It should not be interpreted as a direct equivalent of a conventional CoNLL named-entity class, and in version 1.0 all ANIM instances are single-token spans. The data are distributed in UTF-8 CoNLL-style files with sentence boundaries represented by blank lines.

Version 1.0 is a single-annotator resource. No inter-annotator agreement statistic is reported because a second independent annotator did not label overlapping material. This limitation is made explicit because single-annotator decisions can affect both apparent model ceiling and category boundaries, especially for a task-specific class such as ANIM. The public release includes annotation guidance and a conflict report to support future review or independent re-annotation.

3.3 Audit and strict benchmark construction

The cleaned archival corpus contains 3,184 sentence instances and 44,308 tokens. Audit at the exact surface-token-sequence level found 2,038 unique sentence texts. Among repeated texts, 185 unique sentence groups exhibited more than one BIO sequence, involving 566 sentence instances. These conflicting groups are retained for transparency in the archival release rather than being silently resolved by automated majority voting.

For benchmarking, a sentence text was eligible for the strict view only when all occurrences of that exact token sequence agreed on the complete BIO label sequence. One copy of each eligible text was retained. This procedure yields 1,853 unique, consistently annotated sentences containing 24,239 tokens and 2,277 entity spans. The sentences were assigned to fixed 70/15/15 train/validation/test partitions with a deterministic search that preserved exact split sizes while approximating overall entity-bearing and entity-class distributions. Because every strict sentence text is unique, exact sentence-text overlap across partitions is zero.

The strict benchmark should be understood as a quality-controlled evaluation view rather than a replacement for the archival corpus. It deliberately excludes unresolved conflict groups and duplicate instances, which increases evaluation integrity but can alter the source distribution. The archival material remains available for future adjudication and studies of duplication or annotation revision.

Table 1 Composition of the strict unique benchmark

Partition	Sentences	Tokens	Entities	PER	LOC	TIME	ANIM
Training	1,297	16,893	1,611	1,060	375	110	66
Validation	278	3,588	319	210	70	24	15
Test	278	3,758	347	228	81	24	14
Total	1,853	24,239	2,277	1,498	526	158	95

3.4 Class distribution and evaluation challenge

The benchmark is strongly imbalanced both between entity and non-entity tokens and among entity categories. In the training split, O occurs 15,187 times out of 16,893 labelled tokens (89.90%). B-PER occurs 1,060 times, whereas B-TIME occurs 110 times and B-ANIM only 66 times. Continuation labels are even rarer: I-LOC occurs 42 times, I-PER 24 times, and I-TIME 29 times. Figure 1 visualizes the distribution on a logarithmic scale.

The test partition has a similar evaluation problem: O accounts for 3,396 of 3,758 tokens (90.37%). An all-O system would therefore obtain approximately 90.37% token accuracy while recognizing no entity at all. Token accuracy is

consequently reported only as a secondary diagnostic. Exact-span entity precision, recall, and F1 are the primary measures, supplemented by macro-F1 and class-specific scores.

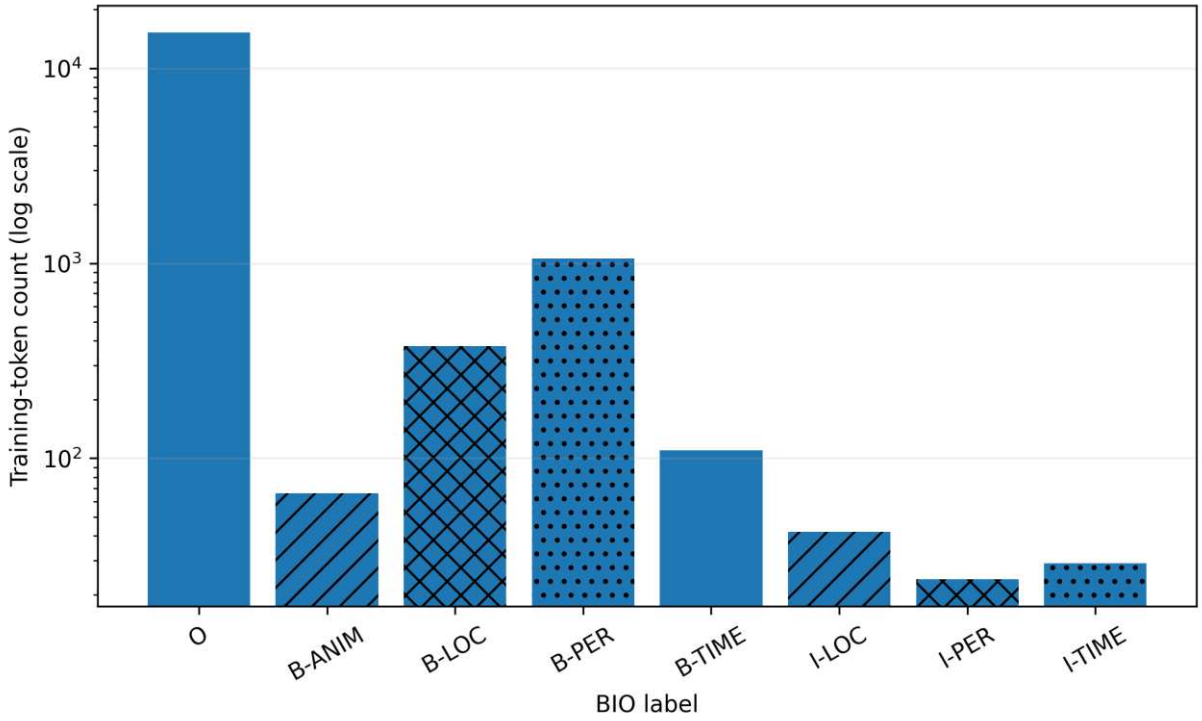


Fig. 1 Training-token distribution across the BIO label inventory on a logarithmic scale; counts are derived only from the strict benchmark training partition

4 Experimental methodology

4.1 Research design and model selection

All models use the same strict train/validation/test files from Table 1. The design intentionally spans three established modelling paradigms. CRF represents a feature-based statistical sequence model; BiLSTM-CRF represents a neural sequence model learned from the task corpus without pretrained word embeddings; and XLM-R represents a large multilingual pretrained transformer. A fourth model, class-weighted XLM-R, isolates the effect of cost-sensitive fine-tuning while holding the pretrained architecture and tokenization strategy constant.

This choice follows two principles. First, the benchmark is intended to be useful to researchers with different computational constraints, so a strong lightweight baseline is valuable. Second, the less-resourced-language setting requires a contemporary transformer comparator rather than only older statistical approaches. We therefore use CRF and BiLSTM-CRF as interpretable reference baselines and XLM-R as the state-of-the-art model family required for a meaningful modern comparison.

Table 2 Model configurations used in the common benchmark pipeline

Model	Representation / features	Optimization and selection	Runs
CRF	Token identity; lowercase; shape; length; case/digit flags; 1–3 char prefixes/suffixes; previous/next-token context; BOS/EOS	L-BFGS; c1=0.1; c2=0.1; max 200 iterations; all transitions enabled	Deterministic
BiLSTM-CRF	96-d randomly initialized word embeddings; 128-d bidirectional LSTM representation; dropout 0.30; first-order CRF	Adam, lr=0.003; batch 32; max 20 epochs; patience 3; validation micro-F1	3 seeds
XLM-R	FacebookAI/xlm-roberta-base; first subword labelled; later subwords ignored for loss/evaluation	AdamW, lr=2e-5; wd=0.01; 10% warm-up; batch 16; max 6 epochs; patience 2; validation micro-F1	3 seeds
Weighted XLM-R	Same XLM-R model and alignment; training-derived BIO loss weights	Weighted cross-entropy; AdamW, lr=2e-5; wd=0.01; 10% warm-up; batch 16; max 8 epochs; patience 3	3 seeds

4.2 Feature-based CRF

The CRF baseline was implemented with sklearn-crfsuite using L-BFGS optimization, L1 coefficient c1=0.1, L2 coefficient c2=0.1, a 200-iteration cap, and all possible label transitions enabled. Each token is represented by lexical and local-context features: raw and lowercase word forms, character shape, token length, uppercase/title/digit indicators, one- to three-character prefixes and suffixes, previous- and next-token forms and shapes, and sentence-boundary flags. These features were selected because the dataset is small and contains lexical/orthographic cues that can be exploited without learning high-dimensional embeddings. The optimizer is effectively deterministic under this setup, so repeated nominal seed values produced identical outputs. We therefore report the CRF result once rather than treating duplicate runs as independent stochastic trials.

4.3 BiLSTM-CRF

The neural baseline follows the BiLSTM-CRF sequence-labelling pattern of Lample et al. (2016). A training-derived vocabulary maps unseen tokens to an UNK symbol. Tokens are represented by 96-dimensional randomly initialized embeddings and processed by a one-layer bidirectional LSTM whose concatenated hidden representation has 128 dimensions. A dropout rate of 0.30 is applied before the emission layer, and a first-order CRF provides structured decoding. Training uses Adam with learning rate 0.003, batch size 32, gradient clipping at 1.0, a maximum of 20 epochs, and early stopping after three epochs without improvement in validation exact-span micro-F1. Results are averaged across seeds 20260912, 20260913, and 20260914.

4.4 Vanilla XLM-R

The transformer baseline fine-tunes FacebookAI/xlm-roberta-base (Conneau et al., 2020) for token classification. Input sentences are tokenized with the corresponding fast tokenizer. For a word split into multiple subwords, the original BIO label is assigned to the first subword and subsequent subwords are ignored in the loss using the -100 mask; test predictions are mapped back from the first subword to the original-token sequence. The maximum sequence length is 128 subwords, comfortably above the corpus maximum of 25 original tokens per sentence. Training uses AdamW with learning rate 2×10^{-5}, weight decay 0.01, 10% linear warm-up, batch size 16, gradient clipping at 1.0, and at most six epochs. Early stopping uses validation exact-span micro-F1 with patience two. Three seeds match the BiLSTM-CRF runs.

4.5 Cost-sensitive XLM-R

The follow-up model uses the same XLM-R architecture, tokenizer, label alignment, optimizer family, learning rate, weight decay, warm-up proportion, batch size, and random seeds as vanilla XLM-R. The only substantive change is the token-level training loss. BIO class weights are computed from the strict training partition only; validation and test labels do not influence weighting. For label c with training count n_c, total labelled training tokens N, and K=8 BIO labels, the raw weight is (N/(K n_c))^0.5. The square-root exponent was selected before the follow-up run as a

conservative alternative to raw inverse frequency, reducing the risk that extremely rare continuation labels dominate the loss. Weights are then normalized to mean 1.0.

The weighted model is trained for at most eight epochs with patience three on validation micro-F1. The slightly larger epoch cap and patience allow the cost-sensitive objective additional opportunity to stabilize while keeping checkpoint selection independent of the test set. This experiment is diagnostic rather than an exhaustive hyperparameter search: alpha=0.5 is not tuned on test performance, and no claim is made that it is globally optimal.

Table 3 Training-only BIO label frequencies and normalized cost-sensitive weights

BIO label	Training count	Frequency (%)	Normalized weight
O	15,187	89.897	0.076
B-ANIM	66	0.391	1.154
B-LOC	375	2.220	0.484
B-PER	1,060	6.275	0.288
B-TIME	110	0.651	0.894
I-LOC	42	0.249	1.447
I-PER	24	0.142	1.914
I-TIME	29	0.172	1.742

4.6 Evaluation metrics and reporting

Evaluation is performed on original tokens, not subword pieces. A predicted entity is counted as correct only when its type and complete token span exactly match a gold entity. We report exact-span micro-averaged precision, recall, and F1 as the principal NER measures. Macro-F1 is the unweighted mean of the four class-specific F1 scores (PER, LOC, TIME, ANIM), allowing rare classes to contribute equally to the summary. Token accuracy across all BIO labels is reported for completeness but is not used as the primary basis for model ranking because O dominates the test set.

For the stochastic BiLSTM-CRF and transformer models, tables report the arithmetic mean and sample standard deviation across the three fixed seeds. The CRF is reported as a deterministic single estimate because repeated runs under the selected L-BFGS configuration are identical. Per-class test supports are fixed at 228 PER, 81 LOC, 24 TIME, and 14 ANIM spans. Rare-class scores are therefore interpreted cautiously, particularly ANIM, where a small change in the number of correct spans can move F1 substantially.

4.7 Error taxonomy and paired bootstrap analysis

To move beyond aggregate F1, every test prediction is aligned with the gold entity spans and assigned to one of five error categories. A missed entity is a gold span with no matched prediction; a spurious entity is a prediction with no matched gold span; a type confusion has the correct span boundaries but the wrong category; a boundary error overlaps a gold span with the correct category but incorrect boundaries; and a boundary-and-type error overlaps a gold span while differing in both type and boundaries. Exact matches are excluded from error counts. Mean error events are reported over three seeds for stochastic models.

The primary follow-up comparison—weighted versus vanilla XLM-R—is additionally evaluated with paired sentence-level bootstrap resampling. For each matching seed, 1,000 bootstrap samples of the 278 test sentences are drawn with replacement, and the same sampled indices are used for both models. Exact-span micro-F1 is recalculated for each model on every sample. The 2.5th and 97.5th percentiles of the weighted-minus-vanilla F1 differences form a descriptive 95% bootstrap interval. Because this is a small benchmark and only three training seeds are available, the intervals are used as uncertainty evidence rather than as a definitive hypothesis test.

4.8 Reproducibility and AI-assisted workflow

The corpus is publicly available at https://doi.org/10.5281/zenodo.22344918. Online Resource 1 contains the Google Colab notebooks used for the common CRF/BiLSTM-CRF/XLM-R benchmark and the class-weighted XLM-R

follow-up, together with the machine-readable CSV/JSON outputs and prediction files used to generate the tables and figures in this paper. The notebooks fix all reported seed values and reconstruct evaluation from the deposited strict benchmark files.

OpenAI ChatGPT was used to assist with drafting and reviewing experimental code, organizing the manuscript, and language refinement. The authors executed the experiments in Google Colab, retained the resulting CSV/JSON outputs and predictions, checked reported values against those outputs and the public corpus, and reviewed the generated text and code. ChatGPT is not an author; responsibility for the design, execution, verification, interpretation, and final manuscript remains with the human authors.

5 Results

5.1 Overall benchmark performance

Table 4 reports the common test results. The feature-based CRF is the strongest model by exact-span micro-F1, reaching 69.03% with 78.39% precision and 61.67% recall. BiLSTM-CRF follows at 63.14 ± 1.75% F1. Vanilla XLM-R reaches 60.00 ± 1.95%, despite slightly higher precision than the BiLSTM-CRF on average. Weighted XLM-R has the lowest micro-F1, 58.32 ± 1.57%, but the highest recall of all four models at 67.72 ± 7.27%.

Accuracy is high for every model, ranging from 91.80% to 95.08%, but those values must be interpreted against the 90.37% O-token rate in the test set. The more informative contrast is between micro- and macro-F1. Vanilla XLM-R has macro-F1 of only 32.04 ± 2.62%, showing that its 60.00% micro-F1 is concentrated in higher-support categories. Weighted XLM-R raises macro-F1 to 50.74 ± 0.54%, almost matching BiLSTM-CRF (50.82 ± 1.43%) despite its lower micro-F1. Figure 2 displays the precision/recall/F1 trade-off across models.

Table 4 Overall strict-test performance. Neural and transformer results are mean ± sample SD over three seeds; CRF is deterministic under the reported configuration

Model	Accuracy (%)	Precision (%)	Recall (%)	Micro-F1 (%)	Macro-F1 (%)
CRF	95.08	78.39	61.67	69.03	58.83
BiLSTM-CRF	93.88 ± 0.81	68.35 ± 8.90	59.46 ± 4.62	63.14 ± 1.75	50.82 ± 1.43
XLM-R	94.10 ± 0.20	69.10 ± 2.32	53.03 ± 2.00	60.00 ± 1.95	32.04 ± 2.62
Weighted XLM-R	91.80 ± 1.43	52.09 ± 7.17	67.72 ± 7.27	58.32 ± 1.57	50.74 ± 0.54

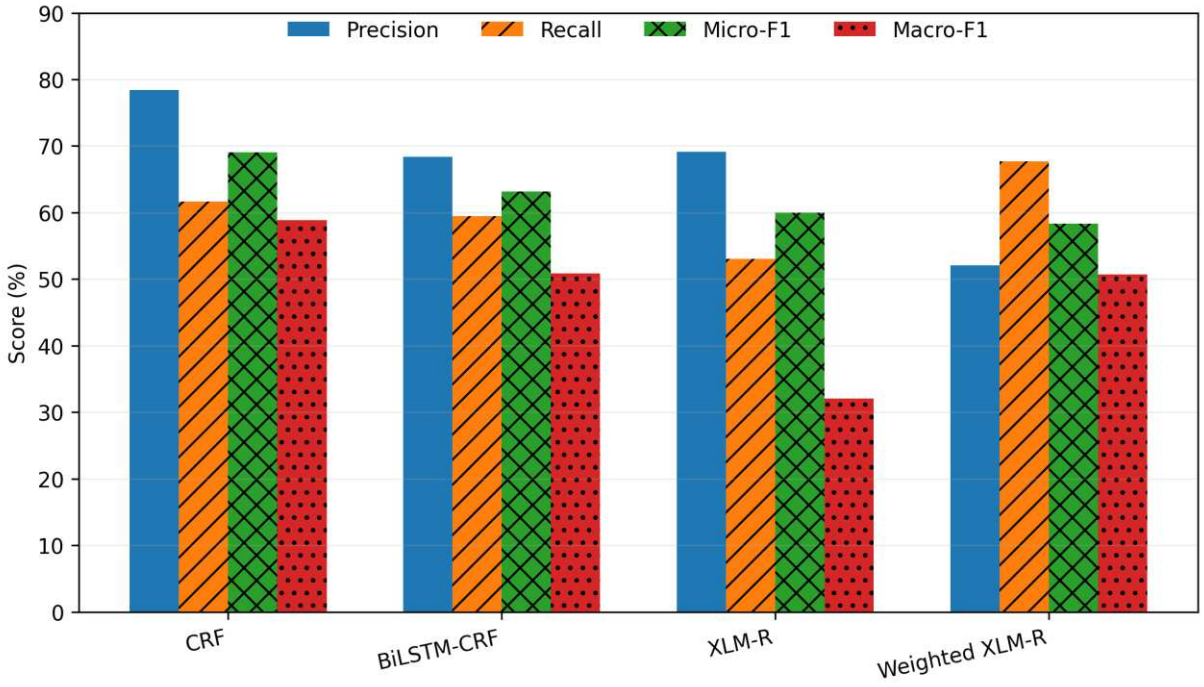


Fig. 2 Entity-level precision, recall, micro-F1, and macro-F1 on the strict test partition; CRF is deterministic and other bars show three-seed means

5.2 Entity-class performance

Table 5 and Figure 3 show that model ranking depends strongly on entity type. PER is the most consistently recognized class. CRF reaches 72.12% PER F1, while vanilla XLM-R and BiLSTM-CRF are close at 66.51% and 66.13%, respectively. For LOC, CRF again leads at 69.44%, followed by BiLSTM-CRF at 65.91%, weighted XLM-R at 61.79%, and vanilla XLM-R at 57.05%.

The largest differences occur for the rare classes. Vanilla XLM-R produces mean TIME F1 of only 4.60% and ANIM F1 of 0.00% across all three runs. CRF retains 46.15% TIME F1 and 47.62% ANIM F1, while BiLSTM-CRF obtains 31.67% and 39.60%, respectively. The very low vanilla transformer scores indicate that global micro-F1 alone would obscure a near-complete failure to recover ANIM and highly unstable recognition of TIME.

Table 5 Exact-span F1 by entity class; test supports are shown in the header. Stochastic-model values are mean ± sample SD over three seeds

Model	PER (n=228)	LOC (n=81)	TIME (n=24)	ANIM (n=14)
CRF	72.12	69.44	46.15	47.62
BiLSTM-CRF	66.13 ± 2.61	65.91 ± 1.41	31.67 ± 7.64	39.60 ± 6.07
XLM-R	66.51 ± 1.99	57.05 ± 1.69	4.60 ± 7.96	0.00 ± 0.00
Weighted XLM-R	62.12 ± 2.17	61.79 ± 1.07	36.80 ± 1.57	42.24 ± 1.94

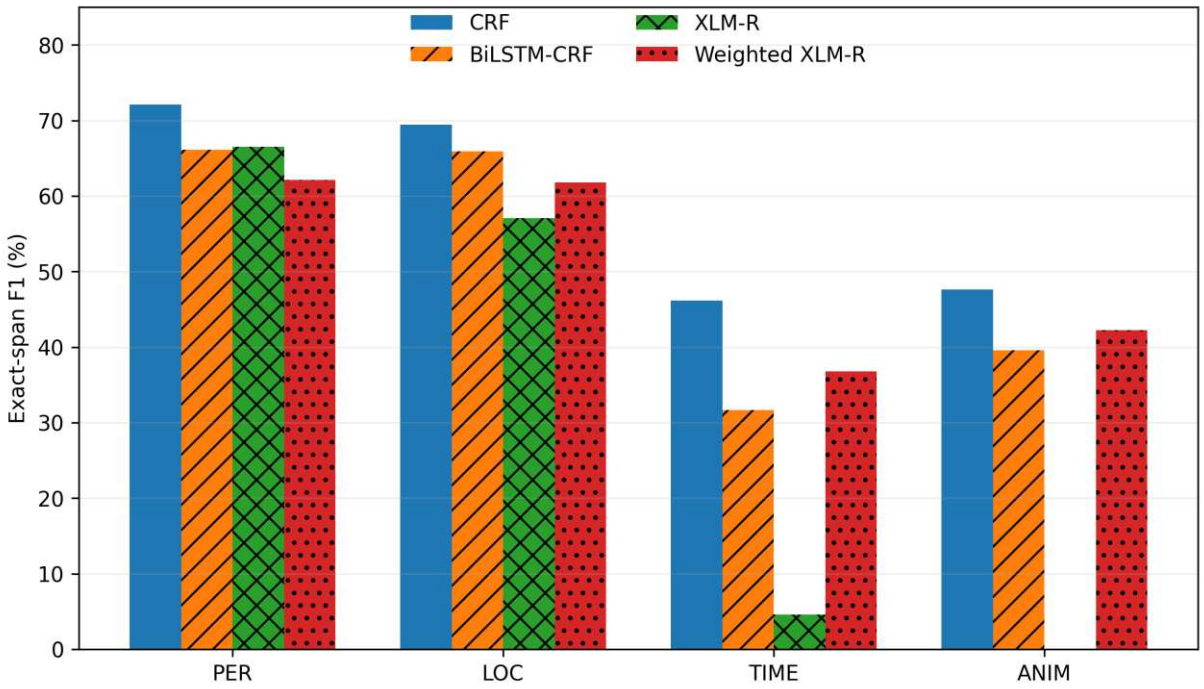


Fig. 3 Exact-span F1 by entity class. TIME and ANIM have only 24 and 14 gold test entities, respectively, and should therefore be interpreted with their small supports in mind

5.3 Effect of class-weighted XLM-R

Cost-sensitive fine-tuning changes the XLM-R operating point rather than improving every metric. Relative to vanilla XLM-R, weighted XLM-R increases mean recall from 53.03% to 67.72% (+14.70 percentage points) but reduces precision from 69.10% to 52.09% (-17.01 points). Micro-F1 consequently declines from 60.00% to 58.32% (-1.68 points). The weighted model therefore should not be characterized as superior overall.

The effect is very different under macro and class-specific evaluation. Macro-F1 rises from 32.04% to 50.74%, an increase of 18.70 points. LOC F1 improves from 57.05% to 61.79%. More importantly, TIME F1 increases from 4.60% to 36.80% (+32.20 points), and ANIM rises from 0.00% to 42.24% (+42.24 points). PER decreases from 66.51% to 62.12%. The cost-sensitive objective therefore redistributes model capacity toward categories that vanilla fine-tuning largely ignores, but the redistribution is not free: more non-entity tokens are incorrectly promoted to entities.

The result directly answers RQ3. Training-only weighting can recover rare Nya-Həba categories for XLM-R, but the resulting precision/recall trade-off means the preferred model depends on application priorities. A high-precision extraction setting would favor the CRF, whereas a screening or candidate-generation setting in which rare-entity recall is important might find the weighted transformer useful as a first stage.

5.4 Error profiles

Table 6 quantifies the error mechanism behind the aggregate scores. CRF is comparatively conservative: it averages 119 missed entities and 45 spurious entities. BiLSTM-CRF misses a similar number of entities but produces more spurious predictions (84.33 on average). Vanilla XLM-R has the highest missed-entity count among the three unweighted baselines (135.00) and more type confusions than either CRF or BiLSTM-CRF.

Weighted XLM-R changes this pattern sharply. Missed entities fall from 135.00 for vanilla XLM-R to 78.67, consistent with the recall gain. At the same time, spurious entities increase from 54.33 to 192.33 and type confusions rise from 21.33 to 27.67. Boundary errors remain uncommon for every model. Figure 4 makes the

sensitivity/specificity trade-off visually clear: weighting succeeds primarily by predicting more entities, not by improving boundary precision across the board.

Table 6 Mean test error events. Standard deviations are shown for stochastic models; CRF is deterministic

Model	Missed	Spurious	Type confusion	Boundary	Boundary + type
CRF	119	45	11	3	0
BiLSTM-CRF	123.67 ± 19.60	84.33 ± 47.44	11.67 ± 3.06	5.00 ± 0.00	0.33 ± 0.58
XLM-R	135.00 ± 1.73	54.33 ± 7.57	21.33 ± 5.77	6.33 ± 0.58	0.33 ± 0.58
Weighted XLM-R	78.67 ± 23.46	192.33 ± 81.30	27.67 ± 4.04	5.67 ± 1.53	0.00 ± 0.00

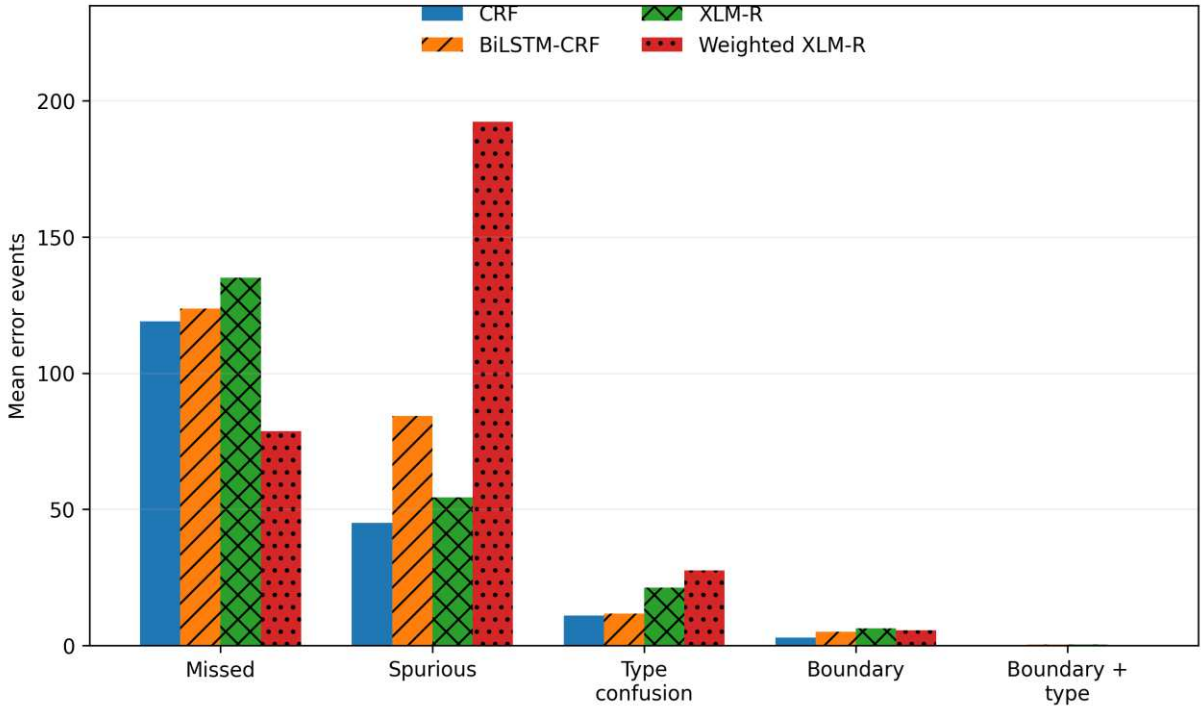


Fig. 4 Mean test error events across models. Cost-sensitive XLM-R markedly reduces missed entities but produces many more spurious entities

5.5 Paired bootstrap comparison of vanilla and weighted XLM-R

The paired bootstrap analysis does not support a clear overall micro-F1 advantage for weighting. For seed 20260912, weighted-minus-vanilla micro-F1 is -0.65 percentage points with a 95% bootstrap interval from -5.88 to +4.24. For seed 20260913, the difference is -1.63 points with interval -6.79 to +2.90, and for seed 20260914 it is -2.76 points with interval -7.74 to +2.64. Every interval crosses zero (Table 7; Figure 5).

These intervals reinforce the descriptive interpretation above. The central contribution of class weighting is not a statistically clear increase in overall micro-F1; it is a redistribution of errors that substantially increases rare-class recall and macro-F1. Claims in the discussion are therefore framed around this trade-off rather than significance or universal superiority.

Table 7 Paired sentence-level bootstrap comparison of weighted and vanilla XLM-R micro-F1; 1,000 resamples per matched seed

Seed	Vanilla F1	Weighted F1	Difference (points)	95% bootstrap interval (points)
20260912	57.89	57.25	-0.65	[-5.88, +4.24]
20260913	61.74	60.12	-1.63	[-6.79, +2.90]
20260914	60.35	57.59	-2.76	[-7.74, +2.64]

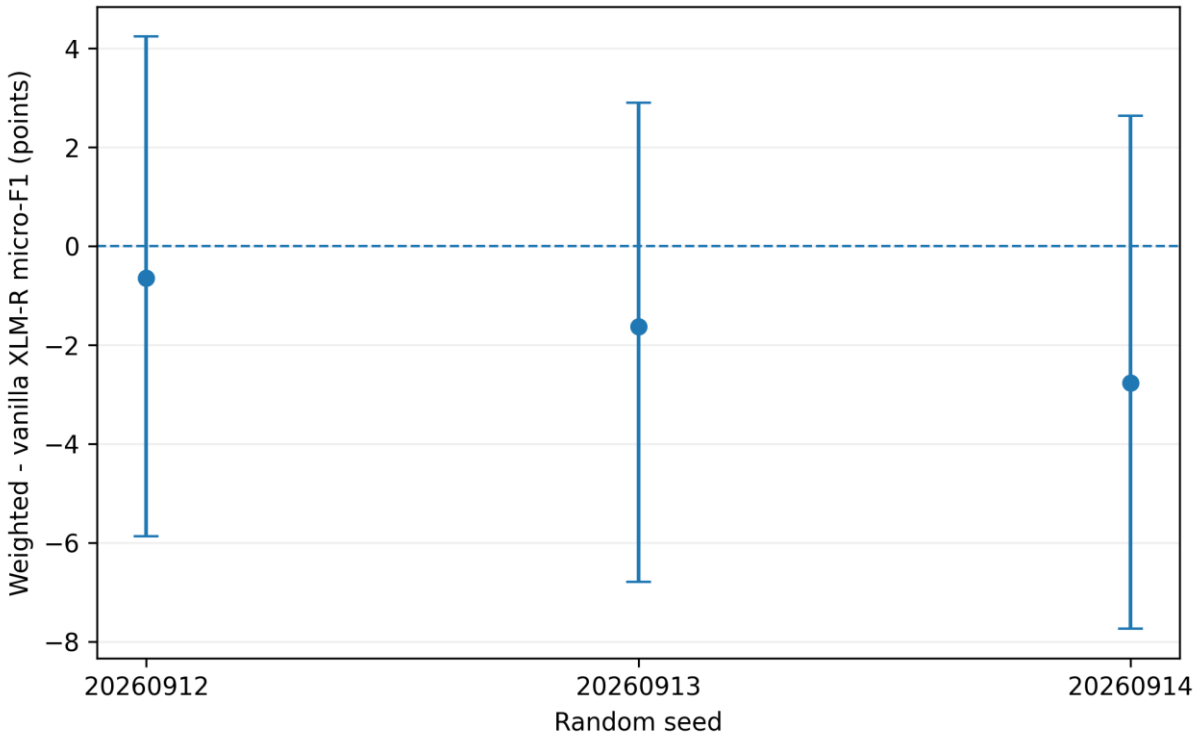


Fig. 5 Weighted-minus-vanilla XLM-R exact-span micro-F1 differences with percentile 95% paired sentence-level bootstrap intervals; all intervals include zero

6 Discussion

6.1 Why the simpler CRF leads overall

The strongest overall result is the feature-based CRF's 69.03% exact-span micro-F1. This finding should not be read as a general claim that statistical models are superior to transformers. It is instead evidence that model complexity does not automatically overcome extreme scarcity. The strict training split contains only 1,297 sentences and 1,611 entity spans. In that regime, a CRF can immediately exploit repeated lexical forms, prefixes and suffixes, token shape, capitalization patterns, and local context without needing to learn an embedding space from the task corpus or adapt a very large pretrained representation to a minimally represented target language.

The BiLSTM-CRF result supports the same interpretation from another angle. It has structured decoding like the CRF but learns 96-dimensional word embeddings from the Nya-Həba training split itself. Its 63.14% micro-F1 is respectable yet lower than CRF. The gap suggests that the small corpus does not provide enough labelled context for randomly initialized embeddings to dominate explicit lexical/orthographic features. This is consistent with the broader low-resource literature: neural methods can be powerful, but their data requirements and inductive biases matter (Hedderich et al., 2021).

Vanilla XLM-R provides a different challenge. Multilingual pretraining offers rich cross-lingual representations, and XLM-R is a strong general baseline (Conneau et al., 2020). Nevertheless, successful transfer is not guaranteed for

every less-resourced language. African-language work has repeatedly shown that language coverage, adaptation, and transfer-language choice can materially affect downstream performance (Adelani et al., 2022; Ogueji et al., 2021; Alabi et al., 2022). The present result therefore complements, rather than contradicts, transformer success: a general multilingual model fine-tuned on 1,297 Nya-Həba sentences is not automatically a stronger benchmark than a feature-based model tailored to the observed tokens.

A further possibility is that the strict benchmark rewards precisely the types of local regularities captured by the CRF while being too small to expose advantages that a transformer might show across broader genres or unseen lexical distributions. Because detailed source-genre metadata are not available, the study cannot separate lexical memorization from deeper contextual generalization within the strict benchmark. Future cross-domain evaluation is needed before any architecture-level conclusion is generalized beyond this resource.

6.2 Class imbalance changes the transformer story

If evaluation stopped at micro-F1, XLM-R would appear simply weaker than the two structured baselines. Per-class analysis reveals a more specific failure: vanilla XLM-R largely learns PER and LOC but does not reliably learn the rare classes. ANIM has zero F1 in every vanilla run; TIME averages only 4.60%. This is exactly the kind of behavior that micro averaging can hide because 228 PER and 81 LOC spans dominate a test set containing only 24 TIME and 14 ANIM spans.

The class-weighted experiment demonstrates that the rare categories are not wholly inaccessible to the transformer. With no new training sentences and no test-informed tuning, weighted XLM-R raises TIME F1 to 36.80% and ANIM to 42.24%. The improvement is accompanied by higher LOC F1 and a very large macro-F1 gain. In practical terms, the model becomes willing to assign labels that ordinary cross-entropy had learned to avoid.

The cost is equally important. Precision falls by 17.01 points and spurious entities more than triple relative to vanilla XLM-R. This behavior is consistent with the mechanics of loss reweighting: increasing the penalty for rare-class mistakes changes decision boundaries and makes false-positive rare/entity predictions less costly relative to false negatives. Cost-sensitive NER research has reported benefits from reweighting in other imbalanced settings (Mobasher et al., 2023), but the Nya-Həba experiment shows why such methods must be evaluated with both recall and precision. Rare-class recovery is a meaningful success only when its false-positive burden is reported alongside it.

Macro-F1 is especially informative here. Weighted XLM-R's micro-F1 falls below vanilla XLM-R, yet macro-F1 increases from 32.04% to 50.74%. Neither summary is intrinsically 'correct'; they answer different questions. Micro-F1 reflects performance over the entity population as distributed in the test set, whereas macro-F1 asks whether each entity class is being served comparably. For resource development, reporting both prevents a high-frequency class from masking complete failure on a documented minority category.

6.3 Implications for less-resourced NER benchmarking

Three practical implications follow. First, leakage control should be treated as part of benchmark design, not as an optional post-hoc check. The archival Nya-Həba resource contains substantial exact repetition, and some repeated texts carry conflicting labels. A random record-level split could therefore evaluate memorization of identical sentence text or mix annotation states across partitions. The strict benchmark's zero exact-text overlap provides a more defensible basis for model comparison, consistent with evidence that train/test overlap can inflate NLP evaluation (Elangovan et al., 2021).

Second, accuracy should not be used as the headline metric for this task. The all-O test accuracy of approximately 90.37% makes it possible for a model to appear numerically strong without recognizing a single entity. Exact-span micro-F1, macro-F1, per-class supports, and error counts provide a more faithful picture. The same principle applies more broadly to less-resourced datasets in which a small number of categories or templates dominate the label distribution.

Third, a useful baseline suite should include both computationally accessible and modern models. The CRF result is valuable for researchers who need a strong CPU-friendly reference. BiLSTM-CRF shows what can be learned from

the task corpus without external pretraining. XLM-R tests the standard multilingual-pretraining assumption. The weighted variant reveals a specific imbalance intervention. This layered design is more informative than reporting a single transformer number and is directly reusable as the corpus grows.

The results also suggest clear next steps. Language-adaptive pretraining on unlabelled Nya-Həba text, Africa-focused multilingual models such as AfriBERTa or SERENGETI, transfer from related languages, character/subword-aware structured models, focal or calibrated cost-sensitive losses, and controlled data augmentation are natural extensions. MasakhaNER 2.0 shows that Africa-centric transfer relations can matter (Adelani et al., 2022), while AfriBERTa and related work indicate that focused pretraining can be competitive despite smaller pretraining corpora (Ogueji et al., 2021). The present benchmark provides a stable test bed for evaluating those strategies without reintroducing exact sentence-text leakage.

6.4 Resource-development implications

The error profiles identify where additional annotation effort is likely to be most useful. TIME and ANIM need greater support, but simply adding more examples is not enough if category definitions remain ambiguous. ANIM is deliberately corpus-specific and may cover semantically salient animal references beyond conventional proper-name NER. A future version should therefore pair additional examples with explicit positive and negative boundary cases, followed by independent overlapping annotation to estimate agreement.

The 185 conflicting duplicate-text groups excluded from the strict benchmark are another high-value target for human adjudication. Resolving them would increase the clean unique benchmark and provide concrete evidence about which distinctions are intrinsically context-sensitive versus annotation errors. Versioned public release is preferable to silently modifying v1.0, so future corrections should preserve the current DOI record and document how adjudication changes counts and splits.

Finally, genre and provenance metadata should be strengthened. The current resource combines pre-existing materials of multiple types, but does not provide a balanced genre design. Explicit genre labels would permit controlled out-of-domain evaluation and help determine whether the CRF advantage is tied to lexical regularity within particular source types. This would also enable more rigorous study of distribution shift, an issue shown to have substantial effects across NER datasets (Li & Groth, 2025).

7 Limitations and threats to validity

The first limitation is annotation reliability. BIO labels were assigned by one annotator, so the benchmark has no inter-annotator agreement estimate. Structural validation can verify BIO consistency but cannot establish that every entity decision is linguistically uncontested. The conflict audit improves transparency, yet independent re-annotation of a representative subset is needed to quantify human agreement and refine difficult categories.

Second, the strict benchmark is small. It contains 1,853 sentences and only 347 gold entities in the test split, including 24 TIME and 14 ANIM spans. Per-class F1 for those categories is therefore high-variance. Three stochastic seeds provide some evidence of training sensitivity, but more seeds would better characterize variance. The bootstrap analysis addresses test-sentence sampling uncertainty only for the vanilla-versus-weighted XLM-R comparison and should not be interpreted as a complete significance analysis across all models.

Third, the benchmark is a cleaned subset of the archival corpus. Excluding duplicate text and conflicting annotation groups improves evaluation integrity, but may make the strict data less representative of the original source distribution. The study does not compare naïve random-split scores against strict-split scores because the goal is to establish defensible baselines, not to quantify an intentionally leaky condition. A dedicated leakage-impact experiment could be useful future work if conducted without encouraging use of contaminated scores as benchmarks.

Fourth, the model search is deliberately limited. We evaluate one feature-based CRF, one randomly initialized BiLSTM-CRF, one general multilingual transformer, and one fixed weighting scheme. We do not conduct a broad hyperparameter optimization, language-adaptive pretraining, data augmentation, an ensemble, or an Africa-specific pretrained model. The reported scores are therefore baselines rather than state-of-the-art claims.

Fifth, the cost-sensitive weighting is token-label based. The square-root inverse-frequency rule with alpha=0.5 was fixed before the follow-up run and not optimized on test data, but other weighting functions could yield a better precision/recall balance. Extremely rare continuation labels receive comparatively large weights, and the resulting false-positive growth shows that weighting can overcorrect. Future work should select cost-sensitive objectives on validation criteria that explicitly combine rare-class recall and precision.

Sixth, source materials are not balanced by genre and the benchmark does not test geographic, dialectal, temporal, or domain transfer. The strongest model on this test partition may not be strongest on new Nya-Həba data from another source. The findings should therefore be interpreted as benchmark-specific evidence about extreme data scarcity and imbalance, not as a universal ranking of CRF, BiLSTM-CRF, and XLM-R for all Nya-Həba applications.

8 Conclusion

This study establishes reproducible NER baselines for the leakage-controlled strict view of the Nya-Həba Named Entity Recognition Corpus v1.0. On 1,853 unique, consistently annotated sentences with zero exact sentence-text overlap across partitions, a feature-based CRF achieves the strongest overall exact-span micro-F1 at 69.03%. BiLSTM-CRF reaches 63.14 ± 1.75%, while vanilla XLM-R reaches 60.00 ± 1.95%. The result demonstrates that a large multilingual transformer does not automatically dominate a lightweight statistical sequence model when labelled data are extremely limited.

The transformer analysis also shows why one aggregate score is insufficient. Vanilla XLM-R largely fails on TIME and ANIM, but training-only class weighting raises their F1 from 4.60% to 36.80% and from 0.00% to 42.24%, respectively. That recovery increases recall and macro-F1 but creates many more spurious entities, reducing overall micro-F1 to 58.32 ± 1.57%. Paired bootstrap intervals cross zero for the weighted-versus-vanilla micro-F1 difference, so the contribution is best understood as a change in error distribution rather than an overall performance gain.

For less-resourced NER, the broader lesson is methodological: benchmark construction, class supports, leakage control, per-class reporting, and error analysis are as important as architecture choice. The public Nya-Həba benchmark and accompanying reproducibility materials provide a foundation for future work on language-adaptive pretraining, Africa-focused transfer, independent annotation validation, improved cost-sensitive objectives, and larger versioned releases.

Statements and Declarations

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing interests

The authors declare that they have no financial or non-financial interests that are directly or indirectly related to the work submitted for publication.

Ethics approval

No institutional ethics committee review was undertaken. The study did not recruit human participants or collect new interviews, recordings, social-media data, animal data, or personally identifiable participant data specifically for this research. The corpus was compiled from pre-existing Nya-Həba language materials available to the first author. Before public deposition, the released files were reviewed for obvious contact identifiers and privacy-sensitive content.

Data availability

The Nya-Həba Named Entity Recognition Corpus v1.0, including the strict unique benchmark used in this study, is publicly available on Zenodo: https://doi.org/10.5281/zenodo.22344918 (Ishaya & Yami, 2026).

Code and supplementary material availability

Online Resource 1 (ESM_1_Nya_Heba_NER_Experiments.zip) contains the Google Colab notebooks used for the baseline and class-weighted experiments, machine-readable summary tables, per-class metrics, error-analysis outputs, bootstrap results, and saved prediction files. The notebooks expect the strict benchmark files from the public Zenodo release.

Author contributions

Chahyaandida Ishaya: Conceptualization, Methodology, Software, Investigation, Data curation, Formal analysis, Validation, Visualization, Writing – original draft, Writing – review and editing. Ayuba Madu Yami: Supervision, Methodology, Validation, Writing – review and editing. Ndachimya Magaji Edward: Validation (typographic and orthographic review of Nya-Həba alphabet forms), Writing – review and editing. Manunyi Godfrey: Validation (proofreading of Nya-Həba words and lexical forms), Writing – review and editing.

References

Adebara, I., Elmadany, A., Abdul-Mageed, M., & Alcoba Inciarte, A. (2023). SERENGETI: Massively multilingual language models for Africa. Findings of the Association for Computational Linguistics: ACL 2023, 1498–1537. https://doi.org/10.18653/v1/2023.findings-acl.97

Adelani, D. I., Abbott, J., Neubig, G., et al. (2021). MasakhaNER: Named entity recognition for African languages. Transactions of the Association for Computational Linguistics, 9, 1116–1131. https://doi.org/10.1162/tacl_a_00416

Adelani, D. I., Neubig, G., Ruder, S., et al. (2022). MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 4488–4508. https://doi.org/10.18653/v1/2022.emnlp-main.298

Alabi, J. O., Adelani, D. I., Mosbach, M., & Klakow, D. (2022). Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. Proceedings of the 29th International Conference on Computational Linguistics, 4336–4349.

Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. Transactions of the Association for Computational Linguistics, 6, 587–604. https://doi.org/10.1162/tacl_a_00041

Conneau, A., Khandelwal, K., Goyal, N., et al. (2020). Unsupervised cross-lingual representation learning at scale. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

Elangovan, A., He, J., & Verspoor, K. (2021). Memorization vs. generalization: Quantifying data leakage in NLP performance evaluation. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, 1325–1335. https://doi.org/10.18653/v1/2021.eacl-main.113

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86–92. https://doi.org/10.1145/3458723

Gritta, M., Pilehvar, M. T., & Collier, N. (2020). A pragmatic guide to geoparsing evaluation: Toponyms, named entity recognition and pragmatics. Language Resources and Evaluation, 54, 683–712. https://doi.org/10.1007/s10579-019-09475-3

Hedderich, M. A., Lange, L., Adel, H., Strötgen, J., & Klakow, D. (2021). A survey on recent approaches for natural language processing in low-resource scenarios. Proceedings of NAACL-HLT 2021, 2545–2568. https://doi.org/10.18653/v1/2021.naacl-main.201

Ishaya, C., & Yami, A. M. (2026). Nya-Həba Named Entity Recognition Corpus v1.0 [Data set]. Zenodo. https://doi.org/10.5281/zenodo.22344918

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 6282–6293. https://doi.org/10.18653/v1/2020.acl-main.560

Lafferty, J., McCallum, A., & Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proceedings of the Eighteenth International Conference on Machine Learning, 282–289.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. Proceedings of NAACL-HLT 2016, 260–270. https://doi.org/10.18653/v1/N16-1030

Li, X., & Groth, P. (2025). How different is different? Systematically identifying distribution shifts and their impacts in NER datasets. Language Resources and Evaluation, 59, 1111–1150. https://doi.org/10.1007/s10579-024-09754-8

Mobasher, G., Müller, W., Krebs, O., & Gertz, M. (2023). WeLT: Improving biomedical fine-tuned pre-trained language models with cost-sensitive learning. Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, 427–438. https://doi.org/10.18653/v1/2023.bionlp-1.40

Mu'azu, M. A. (2009). A grammar of the Kilba language. LINCOM Studies in African Linguistics, 76. LINCOM.

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. Lingvisticae Investigationes, 30(1), 3–26. https://doi.org/10.1075/li.30.1.03nad

Ogueji, K., Zhu, Y., & Lin, J. (2021). Small data? No problem! Exploring the viability of pretrained multilingual language models for low-resourced languages. Proceedings of the 1st Workshop on Multilingual Representation Learning, 116–126. https://doi.org/10.18653/v1/2021.mrl-1.11

Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is multilingual BERT? Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 4996–5001. https://doi.org/10.18653/v1/P19-1493

Sainz, O., Campos, J., García-Ferrero, I., Etxaniz, J., Lopez de Lacalle, O., & Agirre, E. (2023). NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. Findings of the Association for Computational Linguistics: EMNLP 2023, 10776–10787. https://doi.org/10.18653/v1/2023.findings-emnlp.722

Tjong Kim Sang, E. F., & De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, 142–147.

Wolf, T., Debut, L., Sanh, V., et al. (2020). Transformers: State-of-the-art natural language processing. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6