Designing and Contextualising Probes for African Languages

Wisdom Aduah
African Institute for Mathematical Sciences
South Africa
wizzy@aims.ac.za

Francois Meyer
Department of Computer Science
University of Cape Town
francois.meyer@uct.ac.za

Abstract

Pretrained language models (PLMs) for African languages are continually improving, but the reasons behind these advances remain unclear. This paper presents the first systematic investigation into probing PLMs for linguistic knowledge about African languages. We train layer-wise probes for six typologically diverse African languages to analyse how linguistic features are distributed. We also design control tasks, a way to interpret probe performance, for the MasakhaPOS dataset. We find PLMs adapted for African languages to encode more linguistic information about target languages than massively multilingual PLMs. Our results reaffirm previous findings that token-level syntactic information concentrates in middle-to-last layers, while sentence-level semantic information is distributed across all layers. Through control tasks and probing baselines, we confirm that performance reflects the internal knowledge of PLMs rather than probe memorisation. Our study applies established interpretability techniques to African-language PLMs. In doing so, we highlight the internal mechanisms underlying the success of strategies like active learning and multilingual adaptation.

1 Introduction

The past few years have seen the proliferation of pretrained language models (PLMs) across various domains including education, healthcare, and finance (Hadi et al., 2024). The blackbox nature of these models, paired with their increasing size and complexity, has prompted the growing subfield of NLP interpretability (Luo and Specia, 2024). These methods aim for insights into the internal mechanisms underlying the success and failures of PLMs. One of the earliest interpretability methods to gain traction in NLP was probing (Alain and Bengio, 2017), which trains a classifier on intermediate PLM representations. Probing measures to what extent specific linguistic features, such as part-of-speech (POS) categories or semantic concepts, are encoded in hidden layers.

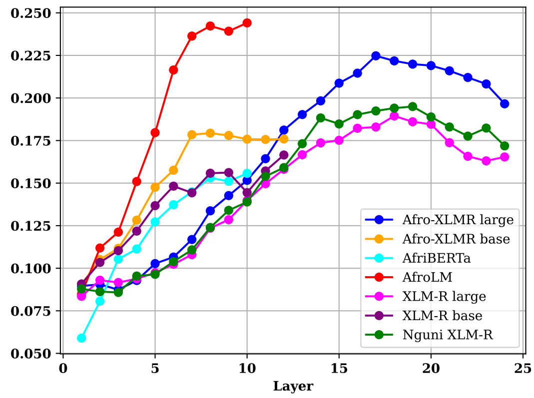

Figure 1: POS probe performance (selectivity), averaged over 6 African languages.

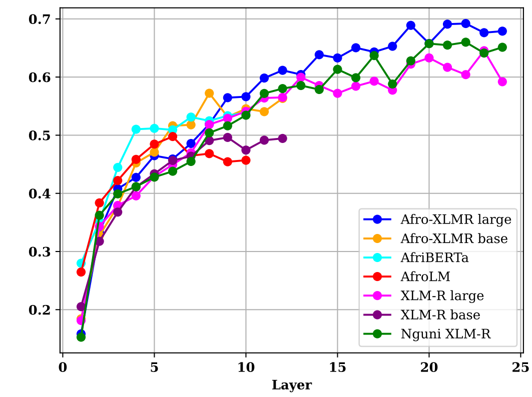

Figure 2: NER probe gains (over random baselines) across layers, averaged over 6 African languages.

Probing provides insights into the internal mechanisms of PLMs by revealing how models acquire, store, and leverage linguistic information in hidden layers. This allows NLP practitioners to better understand the mechanisms by which PLMs succeed in certain tasks, and can also point to the underlying reasons for failing in others. Besides contributing to a greater, linguistically grounded understanding of PLM computations, probing also has the poten-

Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025), pages 39–51
July 31, 2025 ©2025 Association for Computational Linguistics

Model	Layers	Params	Swahili	Igbo	Hausa	Luganda	isiXhosa	Naija
XLM-R-base (Conneau et al., 2020)	12	270M	★★	☆☆	★★	☆☆	★★	☆☆
XLM-R-large (Conneau et al., 2020)	24	550M	★★	☆☆	★★	☆☆	★★	☆☆
AfroXLMR-base (Alabi et al., 2022)	12	270M	★★	★★	★★	☆☆	★★	★★
AfroXLMR-large (Alabi et al., 2022)	24	550M	★★	★★	★★	☆☆	★★	★★
Nguni-XLMR (Meyer et al., 2024)	24	550M	★☆	★☆	★☆	☆☆	★★	☆☆
AfriBERTa (Ogueji et al., 2021)	10	126M	★★	★★	★★	☆☆	☆☆	★★
AfroLM (Dossou et al., 2022)	10	264M	★★	★★	★★	★★	★★	★★

Table 1: Language coverage of PLMs. ☆☆ indicates no data from the language was included in pretraining or adaptation. ★☆ shows the language was included in the base model but not in the adapted model. ★★ shows the model was either pretrained or adapted for the language.

tial to contribute to performance gains by guiding the finetuning process for downstream tasks. For example, knowing which layers encode specific properties can inform which layers should be targetted for finetuning, optimising both performance and efficiency (Katinskaia and Yangarber, 2024).

Probing is an established tool in NLP interpretability, having been extensively applied and studied across different settings. One area where it has yet to be applied is the growing body of work on PLMs for African languages. Most African languages are under-represented in the pretraining data of multilingual PLMs, which limits their performance. Efforts to address this gap have led to the development of PLMs targeting African languages, such as AfriBERTa (Ogueji et al., 2021), AfroLM (Dossou et al., 2022), and AfroXLMR (Alabi et al., 2022). These models leverage strategies such as cross-lingual transfer (Conneau et al., 2020), active learning (Dossou et al., 2022), and multilingual adaptive fine-tuning (MAFT) (Alabi et al., 2022) to improve performance for low-resource languages.

Despite this progress, there is limited understanding of how these PLMs encode African languages internally, which is where probing holds promise. Most probing research targets higher-resourced languages such as English, French, and Russian (Arps et al., 2024; Katinskaia and Yangarber, 2024; Conneau et al., 2018; Hou et al., 2024). Previous works have explored some low-resource languages, such as Tagalog, Hindi and Tamil (Arora et al., 2023; Li et al., 2024), but to the best of our knowledge, there has been no research targeting African languages.

In this paper, we conduct the first systematic probing study for PLMs focussed on African languages. We design probes for POS tagging, named entity recognition (NER), and news topic classification (NTC), using the MasakhaPOS (Dione et al., 2023), MasakhaNER (Adelani et al., 2022), and MasakhaNEWS (Adelani et al., 2023) datasets respectively. We train probes on seven masked PLMs (listed in Table 1), representing different approaches to developing PLMs for low-resource languages. We evaluate how syntactic and semantic information is encoded for six African languages, which cover different language families and varying levels of data availability, as shown in Table 1.

To interpret probe accuracies, one has to isolate the contribution of model-encoded knowledge, as opposed to the probe itself learning the task. To enable such probe interpretability for African languages, we design a control task (Hewitt and Liang, 2019) for MasakhaPOS. Control tasks are synthetic tasks that measure to what extent probes can learn a task without model-encoded knowledge. Our control task enables researchers to contextualise probing results for MasakhaPOS.

Our main findings can be summarised as follows:

1. Word-level linguistic knowledge (POS, NER) concentrates in middle layers, while sentence-level information (NTC) is spread out.

2. The inclusion of target languages in pretraining or multilingual adaptation improves probe performance across all tasks.

3. Cross-lingual transfer improves probe performance for languages not in pretraining, but less so for low-resource language families.

2 Background

PLMs for African Languages Multilingual modelling has been leveraged in different ways to build PLMs for African languages. The massively multilingual XLM-R (Conneau et al., 2020) is trained on 100 languages, of which only 8 are African languages. AfroXLMR (Alabi et al., 2022) improves performance by adapting XLM-R for 17 African languages, while Nguni-XLMR (Meyer et al., 2024) adapts XLM-R for the four Nguni languages (isiXhosa, isiZulu, isiNdebele, and Siswati). AfriBERTa (Ogueji et al., 2021) is a smaller model pretrained from scratch on 11 African languages on less than 1GB data. AfroLM (Dossou et al., 2022) is also trained from scratch on 23 African languages, using self-active learning (the model learns to identify beneficial training samples).

Contextualising Probe Performance Probes are not direct measures of model-encoded knowledge, since the probe itself can contribute to performance by learning the task. Probing studies use baselines, such as majority class prediction (Belinkov et al., 2017; Conneau et al., 2018) or probes trained on random representations (Zhang and Bowman, 2018; Conneau et al., 2018; Chrupała et al., 2020; Tenney et al., 2019b), to contextualize performance.

However, even random baselines may encode information that a sophisticated classifier could exploit. As an alternative, Hewitt and Liang (2019) propose control tasks: pairing word types with random labels to neutralise the linguistic information required for the original task. They define selectivity as the difference between original task and control task accuracy. Selectivity captures the contribution of linguistic knowledge to probe performance, so it is a more reliable measure of model knowledge than raw probe accuracies. To enable probe contextualisation for African languages, we design a control task for the MasakhaPOS dataset.

3 Probing Framework

3.1 Probe Design

Some works advocate for linear probes (Alain and Bengio, 2017; Hewitt and Liang, 2019), arguing that they are less prone to memorisation, while others argue that some linguistic features might not be linearly separable in the representation space (Conneau et al., 2018; Pimentel et al., 2020).

For our experiments, we select a probe to strike a balance between complexity and simplicity. Our probe classifier is a multilayer perceptron (MLP) with a single hidden layer of 50 neurons, which we formally define as

$$\mathbf{y} = f(\mathbf{W_2}\,\sigma(\mathbf{W_1}\mathbf{x} + \mathbf{b_1}) + \mathbf{b_2}), \qquad (1)$$

where $\mathbf{x} \in \mathbb{R}^n$ is the input representation, $\mathbf{W_1} \in \mathbb{R}^{m\times n}$ is the weight matrix for the hidden layer, $\mathbf{W_2} \in \mathbb{R}^{k\times m}$ is the weight matrix for the output layer, $\mathbf{b_1} \in \mathbb{R}^m$ and $\mathbf{b_2} \in \mathbb{R}^k$ are bias vectors, $\sigma(\cdot)$ is the ReLu activation function, and $f(\cdot)$ is a softmax function for label prediction.

For POS tagging and NER, we define a word-level task as a function f that maps an input sequence X to an output sequence Y. That is $f : X \longrightarrow Y$, where X is a sequence of contextualized hidden representations (embeddings) of the input text, and Y is the sequence of output labels corresponding to the words encoded by X. Given that some words are tokenized into multiple subwords, we use the first subword in each word to represent the word in the classifier.

For news topic classification (NTC), we define a sentence-level task similarly. Instead of passing word embeddings to the probe classifier, we pass the embedding of the special token for sequence classification (e.g. <s> for XLM-R). We truncate inputs consisting of more tokens than the maximum sequence length of PLMs.

3.2 MasakhaPOS Control Task

As discussed in section 2, control tasks (Hewitt and Liang, 2019) can be used to contextualise probe results. A probe could achieve a high raw accuracy by learning to map word types to labels, without relying on linguistic knowledge. For example, a probe classifier could learn to map the embedding of "walk" to the POS tag "verb", by learning the mapping between word type and label (instead of the mapping between syntactic role and label). Hewitt and Liang (2019) propose *selectivity* as an alternative to raw accuracy. Selectivity is defined as the difference between linguistic task accuracy and the control accuracy. As a measure, it isolates the contribution of model-encoded linguistic knowledge to probe performance.

A control task is designed in two steps:

1. Define the random control behavior for each word type i.e. assign a label *randomly* to each word in the vocabulary.

2. Deterministically label the original task corpus based on control behaviour i.e. annotate each word with its assigned random label.

To define a control task for MasakhaPOS, we randomly map each unique word in the dataset to a random POS tag. Next, we use this synthetic mapping to re-annotate the train / validation / test set. As per Hewitt and Liang (2019), when creating the random mapping (the control behaviour) we sample POS tags according to their empirical distribution in the original MasakhaPOS dataset. Control tasks are designed to have both structure and randomness. Our MasakhaPOS contains structure in that a word type is always mapped to the same tag, but the assignment is random in that it is independent of the linguistic role of words.

3.3 NTC and NER Probe Baselines

Control tasks define word-level control behaviour, so they are not applicable to sentence-level tasks. To interpret NTC probes, we compare the performance of probes trained on PLMs to those trained on random contextual representations. Following Hewitt and Manning (2019), we use an untrained bi-LSTM, and mean-pooling word-level outputs to produce a single sentence representation. Probes trained on these outputs can leverage contextual information, but no linguistic knowledge.

NER is a word-level task, so controls tasks could plausibly be designed for MasakhaNER. However, the procedure is complicated by the fact that NER is actually a *span-level* task (named entities can span multiple words). It is not obvious how to extend the control behaviour design of Hewitt and Liang (2019) to multi-word spans. To contextualise NER results, we randomly re-initialise the architectures of our seven probed PLMs to serve as random baselines (Zhang and Bowman, 2018; Conneau et al., 2018). We estimate model-encoded knowledge by subtracting the F1 score of a random model from the F1 score of the corresponding PLM. For each layer, we refer to this as the probe *gain* over the random baseline.

4 Experimental Setup

4.1 Data

Both MasakhaPOS (Dione et al., 2023) and MasakhaNER (Adelani et al., 2022) cover 20 African languages. MasakhaNEWS (Adelani et al., 2023) covers 16 African languages and contains news articles annotated with one of seven topic labels (business, entertainment, health, politics, religion, sports, technology).

4.2 Languages

The six language in our study (Swahili, Igbo, Hausa, Luganda, isiXhosa, and Naija Pidgin) are included in all three Masakhane[1] datasets. We chose these languages specifically to cover several language families, a broad range of data availability, and varying levels of representation in existing PLMs (as shown in Table 1). As shown in Table 2, the languages cover four language families across East, West, and Southern Africa.

Language	Family	Region	mC4 tokens
Swahili	Bantu	East	1B
Igbo	Volta-Niger	West	90m
Hausa	Chadic	West	200m
Luganda	Bantu	East	0
isiXhosa	Bantu	Southern	60m
Naija Pidgin	Creole	West	0

Table 2: The languages used in our study. The number is tokens in the mC4 corpus (Xue et al., 2021) serves to give an indication of broader data availability.

4.3 PLMs

The seven PLMs selected for our study represent established approaches to developing PLMs for African languages. XLM-R-base and XLM-R-large (Conneau et al., 2020) employ massively multilingual pretraining, while AfroXLMR-base, AfroXLMR-large (Alabi et al., 2022), and Nguni-XLMR-large (Meyer et al., 2024) adapt XLM-R to a more limited set of African languages. AfroXLMR takes a broader adaptation approach than Nguni-XLMR, which focusses only on the four Nguni languages, a group of related languages which includes isiXhosa. AfriBERTa (Ogueji et al., 2021) represents memory-efficient pretraining – it is our smallest model both in terms of parameters and training data size. AfroLM (Dossou et al., 2022) represents sample-efficient pretraining, since its self-active learning maximizes available data by identifying beneficial training samples.

5 Results

We plot probing results for POS, NER, and NTC respectively in Figure 3, Figure 4, and Figure 5. We report and compare best-layer results for each language, model, and task in Table 3.

5.1 POS Tagging

We evaluate our POS probes based on selectivity, which is computed using the MasakhaPOS control task described in subsection 3.2. As shown in Figure 3, the PLMs exhibit positive selectivity across all layers for all languages, except in the case of Igbo. This aligns with previously reported PLM results for MasakhaPOS (Dione et al., 2023), where

[1]https://www.masakhane.io/

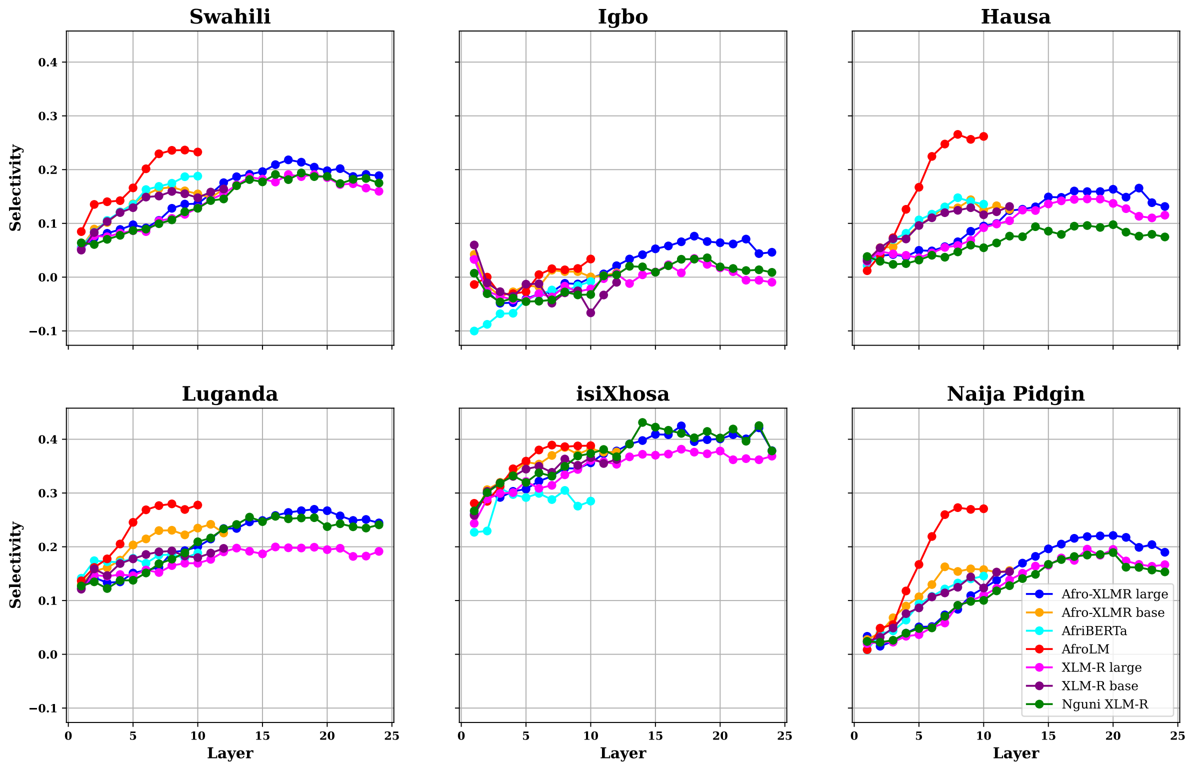

Figure 3: Probe selectivity for POS tagging (the difference between MasakhaPOS accuracy and control task accuracy), across all layers and 6 African languages.

POS tagging accuracies for Igbo were lower than all other languages. Igbo is from the Volta-Niger family, which is under-represented in the datasets of all seven models (as shown in Table 4 in the appendix). This limits the benefit of cross-lingual transfer for Igbo.

For all other languages, POS selectivity is consistently positive, indicating that syntactic roles are reliably being encoded in the hidden representations of the PLMs. A clear pattern emerges in the distribution of POS information across layers. Probe selectivity is low in early layers, increases steadily in middle layers, peaks and plateaus in deeper layers, and finally decreases slightly in final layers. This pattern aligns with existing literature showing that middle-to-last-layers encode syntactic features more effectively (Rogers et al., 2020).

AfroLM stands out as encoding a high amount of POS information, achieving the highest selectivity overall on four of the six languages. While the exact reason for this is unknown, it is possible that self-active learning is used to select training examples that improve the model's syntactic knowledge during pretraining. Among the deeper models, AfroXLMR-large exhibits the greatest internal syntactic knowledge overall, even achieving reasonable selectivity scores for Igbo in deeper layers. The difference in selectivity between AfroXLMR and XLM-R highlights the ability of multilingual adaptation to encode linguistic knowledge about specific languages. Similarly, Nguni-XLMR, exhibits the best probe performance for isiXhosa, one of its four target languages.

We include the raw probe accuracies for POS tagging in Figure 6 in the appendix. The accuracies are quite high, comparing favourably to previously reported PLM performance for MasakhaPOS (Dione et al., 2023).

5.2 NER

To contextualise our NER probes, we compute the per-layer difference between the F1 scores of probes trained on PLMs and their re-initialised counterparts (described in subsection 3.3). As shown in Figure 4, probes trained on PLMs consistently exhibit performance gains over random baselines across all layers and languages. The general trends observed for NER probes are similar to those of POS probes. AfroXLMR achieves the highest probe gains across different languages, while Nguni-XLMR does particularly well for isiXhosa. As in POS tagging, probe performance peaks in

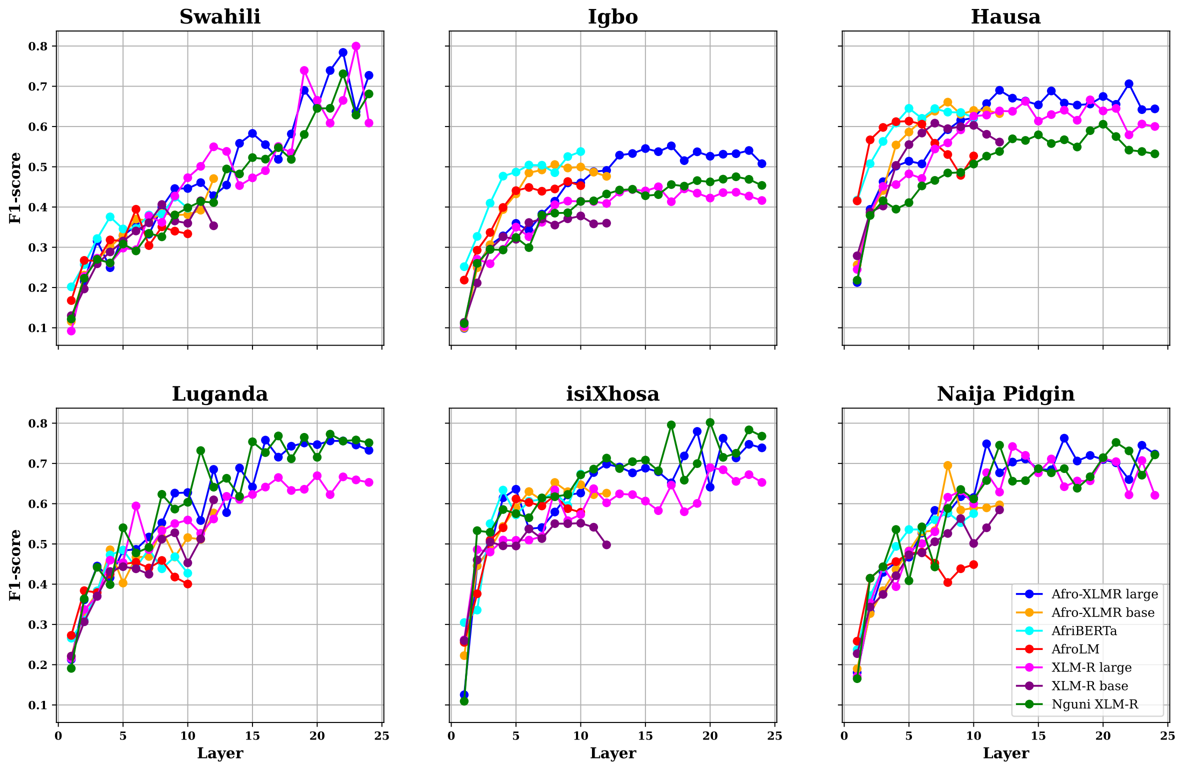

Figure 4: Probe performance gains for NER tagging (F1 improvements over randomly re-initialised PLM architectures), across all layers and 6 African languages.

middle-to-later layers.

We also observe evidence of cross-lingual knowledge representation. Luganda is not included in the pretraining data of either AfroXLMR or Nguni-XLMR but both exhibit high probe performance gains for Luganda than AfroLM, which is pretrained on Luganda. Luganda is of the Bantu language family, which is better represented than other families in the pretraining data of our PLMs (as shown in Table 4 in the appendix). This shows that the PLMs are encoding linguistic similarities between different languages. This cross-lingual representation learning is the mechanism behind the impressive zero-shot performance of PLMs previously reported on MasakhaNER (Adelani et al., 2022).

5.3 News Topic Classification (NTC)

To contextualise our NTC probes, we compare the classification accuracies of probes trained on PLMs and probes trained on random, contextual baselines (described in subsection 3.3). Figure 5 plots probe accuracies alongside random baseline performance. As for POS and NER, multilingual adaptation enhances sentence-level representations for target languages. Beyond this, NTC probing results diverge from the trends reported for POS and NER.

Probe accuracy remains relatively consistent across layers, which aligns with previous work showing that sentence-level semantic information is spread across layers (Tenney et al., 2019a). The one exception to this is Luganda, which exhibits high variance across layers and is the only language for which some PLM layers fall below random probe performance. We are unable to explain this behaviour. It is possible that the data scarcity of Luganda (see Table 2) is a contributing factor and that, unlike for syntactic knowledge, cross-lingual semantic knowledge does not transfer as effectively.

5.4 Analysing Trends Across Tasks

Table 3 lists results for the top-performing layer for each PLM, across all languages and tasks. For each PLM and language, it also shows to what extent the language is represented by the model: (1) not included at all, (2) included in pretraining but not adaptation, or (3) included in pretraining or adaptation. The table reveals trends that hold across all three tasks.

Multilingual adaptation is known to be a reliable method to improve downstream performance for low-resource languages. Our results show that this

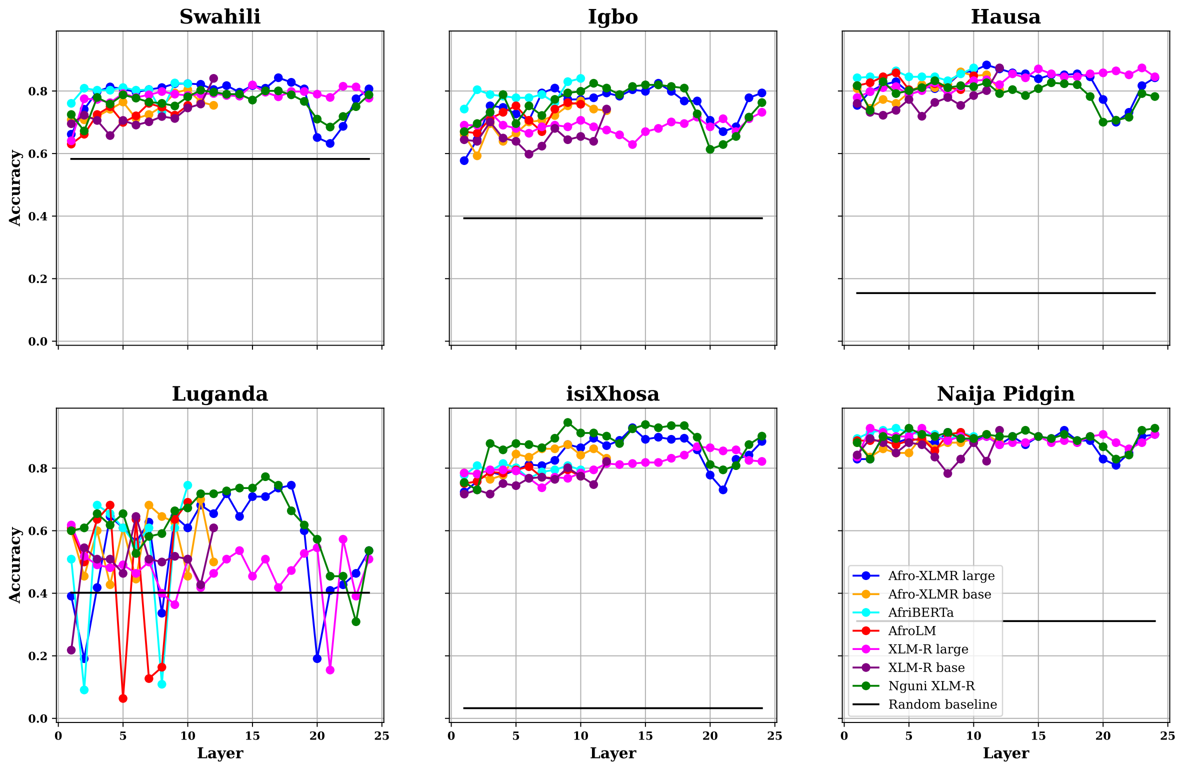

Figure 5: Probe accuracy for news topic classification (visualised in comparison to a random contextual baseline) across all layers and 6 African languages.

is being achieved, in part, by enhancing internal representations of target languages. AfroXLMR-large and Nguni-XLMR-large have the most instances of top-performing layers (six cases each). AfroXLMR-large performs best across Swahili, Igbo, and Hausa, all three of which belong to different language families. The multilingual adaptation of AfroXLMR is highly effective at enhancing linguistic feature encoding across typologically diverse languages. Nguni-XLMR-large performs best for isiXhosa and Luganda (which is also of the Bantu language family). The more focussed, linguistically oriented adaptation of Nguni-XLMR effectively enhances linguistic knowledge for a more limited set of related languages.

A clear pattern in Table 3 is the fact that all top-performing layers (except two) occur in PLMs that represent probed languages in their final pretraining stage (either during adaptation or in pretraining without subsequent adaptation). Best-layer performances (**boldface** in the table) almost always co-occur with maximal language representation (★★). The only exception to this is Luganda, for which Nguni-XLMR-large achieves two best-layer results. While we have previously discussed evidence of zero-shot cross-lingual representation learning, it is clear that including languages in pretraining is essential for encoding language-specific syntactic and semantic knowledge.

6 Conclusion

This paper presents a systematic analysis of the linguistic knowledge encoded in PLMs for African languages. Through extensive probing experiments across seven PLMs and six typologically diverse African languages, we highlight trends in how PLMs represent syntactic and semantic information. To contextualise our results we design a control task for POS tagging and employ randomly initialised baselines to compare against NER and NTC probing results. We show that multilingual adaptation reliably enhances hidden representations for target languages. While token-level linguistic knowledge is primarily encoded in middle and deeper layers, sentence-level semantic information is distributed across layers. We find evidence that cross-lingual learning enhances representations for low-resource languages, such as Luganda, but cannot be relied on to do so for under-represented languages, such as Igbo. We hope this work inspires further research at the intersection of interpretability and NLP for African language.

		Swahili	Igbo	Hausa	Luganda	isiXhosa	Naija
XLM-R-base	Language coverage	★★	☆☆	★★	☆☆	★★	☆☆
	POS selectivity	16.39	5.98	13.15	19.23	36.32	15.39
	NER gain	41.54	37.79	60.84	60.95	55.16	58.46
	NTC accuracy	84.03	74.23	87.38	64.55	82.15	92.11
XLM-R-large	Language coverage	★★	☆☆	★★	☆☆	★★	☆☆
	POS selectivity	19.09	3.29	14.54	19.97	38.15	19.56
	NER gain	**79.98**	45.03	66.63	66.95	68.97	74.20
	NTC accuracy	81.93	73.20	87.38	61.82	86.87	92.76
AfroXLMR-base	Language coverage	★★	★★	★★	☆☆	★★	★★
	POS selectivity	16.73	4.22	14.40	24.18	38.48	16.28
	NER gain	47.07	50.55	66.06	57.7	65.27	69.52
	NTC accuracy	80.46	76.80	86.12	70.00	87.54	90.79
AfroXLMR-large	Language coverage	★★	★★	★★	☆☆	★★	★★
	POS selectivity	21.82	**7.62**	16.56	26.97	42.49	22.03
	NER gain	78.41	**55.20**	**70.63**	75.79	77.96	**76.27**
	NTC accuracy	**84.24**	82.47	**88.33**	74.55	92.93	90.79
Nguni-XLMR-large	Language coverage	★☆	★☆	★☆	☆☆	★★	☆☆
	POS selectivity	19.09	3.6	9.76	25.67	**43.13**	18.92
	NER gain	73.10	47.52	60.57	**77.28**	**80.18**	58.46
	NTC accuracy	80.25	82.47	83.28	**77.27**	**94.61**	**92.76**
AfriBERTa	Language coverage	★★	★★	★★	☆☆	☆☆	★★
	POS selectivity	18.79	-0.82	14.77	18.85	30.87	14.55
	NER gain	42.45	53.78	64.52	48.75	67.36	57.80
	NTC accuracy	82.56	**84.02**	87.38	74.55	81.48	92.76
AfroLM	Language coverage	★★	★★	★★	★★	★★	★★
	POS selectivity	**23.64**	3.37	**26.55**	**27.98**	38.92	**27.28**
	NER gain	39.45	46.33	61.21	45.90	62.07	47.97
	NTC accuracy	76.05	76.29	85.80	69.09	80.47	92.76

Table 3: Best-layer performance for each probing task, with best task performance overall indicated in **boldface**. We show this alongside model language coverage to indicate how language inclusion improves probe performance. ☆☆ indicates no language included in pretraining or adaptation, ★☆ shows the language was included in the base model but not in the adapted model, while ★★ shows the model was either pretrained or adapted for the language.

Limitations

As discussed in Section 2, designing control tasks for NER proved challenging. While control tasks are primarily designed for word-level tasks, NER presents complications because named entities often span multiple words. This makes it difficult to apply the typical control task framework in a meaningful way. Instead, we relied on random baselines, which, although commonly used, are known to have certain limitations (Belinkov, 2022; Hewitt and Liang, 2019).

In this study, we used the first subword as input for the classifier to align tokens with their hidden representations. However, even the choice of subword pooling strategy can make a difference in probing performance (Ács et al., 2021). Other pooling strategies, such as last subword, mean pooling, or attention over subwords, could provide different insights, especially for morphologically rich languages with high subword tokenization rates. Future work should systematically compare the effects of different subword pooling strategies across various syntactic and semantic tasks for African languages.

Acknowledgements

Wisdom Aduah was a Google DeepMind scholar at AIMS South Africa, in the AI for Science Masters program when he conducted this research.

References

Judit Ács, Ákos Kádár, and András Kornai. 2021. Subword pooling makes a difference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2021*. Association for Computational Linguistics.

David Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Alabi, Shamsuddeen Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Mboning Tchiaze Elvis, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo Lerato Mokono, Ignatius Ezeani, Chiamaka Chukwuneke, Mofetoluwa Oluwaseun Adeyemi, Gilles Quentin Hacheme, Idris Abdulmumin, Odunayo Ogundepo, Oreen Yousuf, Tatiana Moteu, and Dietrich Klakow. 2022. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure F. P. Dossou, Akintunde Oladipo, Doreen Nixdorf, Chris Chinenye Emezue, Sana Al-azzawi, Blessing Sibanda, Davis David, Lolwethu Ndolela, Jonathan Mukiibi, Tunde Ajayi, Tatiana Moteu, Brian Odhiambo, Abraham Owodunni, Nnaemeka Obiefuna, Muhidin Mohamed, Shamsuddeen Hassan Muhammad, Teshome Mulugeta Ababu, Saheed Abdullahi Salahudeen, Mesay Gemeda Yigezu, Tajuddeen Gwadabe, Idris Abdulmumin, Mahlet Taye, Oluwabusayo Awoyomi, Iyanuoluwa Shode, Tolulope Adelani, Habiba Abdulganiyu, Abdul-Hakeem Omotayo, Adetola Adeeko, Abeeb Afolabi, Anuoluwapo Aremu, Olanrewaju Samuel, Clemencia Siro, Wangari Kimotho, Onyekachi Ogbu, Chinedu Mbonu, Chiamaka Chukwuneke, Samuel Fanijo, Jessica Ojo, Oyinkansola Awosan, Tadesse Kebede, Toadoum Sari Sakayo, Pamela Nyatsine, Freedmore Sidume, Oreen Yousuf, Mardiyyah Oduwole, Kanda Tshinu, Ussen Kimanuka, Thina Diko, Siyanda Nxakama, Sinodos Nigusse, Abdulmejid Johar, Shafie Mohamed, Fuad Mire Hassan, Moges Ahmed Mehamed, Evrard Ngabire, Jules Jules, Ivan Ssenkungu, and Pontus Stenetorp. 2023. MasakhaNEWS: News topic classification for African languages. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 144–159, Nusa Dua, Bali. Association for Computational Linguistics.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Guillaume Alain and Yoshua Bengio. 2017. Understanding intermediate layers using linear classifier probes.

Arnav Arora, Lucie-aimée Kaffee, and Isabelle Augenstein. 2023. Probing pre-trained language models for cross-cultural differences in values. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 114–130, Dubrovnik, Croatia. Association for Computational Linguistics.

David Arps, Laura Kallmeyer, Younes Samih, and Hassan Sajjad. 2024. Multilingual nonce dependency treebanks: Understanding how language models represent and process syntactic structure. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7822–7844, Mexico City, Mexico. Association for Computational Linguistics.

Yonatan Belinkov. 2022. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, 48(1):207–219.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada. Association for Computational Linguistics.

Grzegorz Chrupała, Bertrand Higy, and Afra Alishahi. 2020. Analyzing analytical methods: The case of phonology in neural models of spoken language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4146–4156, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard 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, Online. Association for Computational Linguistics.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Cheikh M. Bamba Dione, David Ifeoluwa Adelani, Peter Nabende, Jesujoba Alabi, Thapelo Sindane, Happy Buzaaba, Shamsuddeen Hassan Muhammad, Chris Chinenye Emezue, Perez Ogayo, Anuoluwapo Aremu, Catherine Gitau, Derguene Mbaye, Jonathan

Mukiibi, Blessing Sibanda, Bonaventure F. P. Dossou, Andiswa Bukula, Rooweither Mabuya, Allahsera Auguste Tapo, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Fatoumata Ouoba Kabore, Amelia Taylor, Godson Kalipe, Tebogo Macucwa, Vukosi Marivate, Tajuddeen Gwadabe, Mboning Tchiaze Elvis, Ikechukwu Onyenwe, Gratien Atindogbe, Tolulope Adelani, Idris Akinade, Olanrewaju Samuel, Marien Nahimana, Th'eog'ene Musabeyezu, Emile Niyomutabazi, Ester Chimhenga, Kudzai Gotosa, Patrick Mizha, Apelete Agbolo, Seydou Traore, Chinedu Uchechukwu, Aliyu Yusuf, Muhammad Abdullahi, and Dietrich Klakow. 2023. MasakhaPOS: Part-of-speech tagging for typologically diverse African languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10883–10900, Toronto, Canada. Association for Computational Linguistics.

Bonaventure F. P. Dossou, Atnafu Lambebo Tonja, Oreen Yousuf, Salomey Osei, Abigail Oppong, Iyanuoluwa Shode, Oluwabusayo Olufunke Awoyomi, and Chris Emezue. 2022. AfroLM: A self-active learning-based multilingual pretrained language model for 23 African languages. In *Proceedings of The Third Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, pages 52–64, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Muhammad Usman Hadi, Qasem Al Tashi, Abbas Shah, Rizwan Qureshi, Amgad Muneer, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, et al. 2024. Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea Preprints*.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. 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 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Jue Hou, Anisia Katinskaia, Lari Kotilainen, Sathianpong Trangcasanchai, Anh-Duc Vu, and Roman Yangarber. 2024. What do transformers know about government? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17459–17472, Torino, Italia. ELRA and ICCL.

Anisia Katinskaia and Roman Yangarber. 2024. Probing the category of verbal aspect in transformer language models. In *Findings of the Association for Computational Linguistics*, pages 3347–3366, United States. The Association for Computational Linguistics. Findings of the Association for Computational Linguistics : NAACL 2024 ; Conference date: 16-06-2024 Through 21-06-2024.

Daoyang Li, Mingyu Jin, Qingcheng Zeng, Haiyan Zhao, and Mengnan Du. 2024. Exploring multilingual probing in large language models: A cross-language analysis. *Preprint*, arXiv:2409.14459.

Haoyan Luo and Lucia Specia. 2024. From understanding to utilization: A survey on explainability for large language models. *Preprint*, arXiv:2401.12874.

Francois Meyer, Haiyue Song, Abhisek Chakrabarty, Jan Buys, Raj Dabre, and Hideki Tanaka. 2024. NGLUEni: Benchmarking and adapting pretrained language models for nguni languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12247–12258, Torino, Italia. ELRA and ICCL.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622, Online. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019b. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings* *of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Kelly Zhang and Samuel Bowman. 2018. Language modeling teaches you more than translation does: Lessons learned through auxiliary syntactic task analysis. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 359–361, Brussels, Belgium. Association for Computational Linguistics.

A Data Information

	Bantu (%)	Volta-Niger (%)	Afro-Asiatic (%)	Others (%)
XLM-R	33.3	0.0	6.3	60.4
AfroXLMR	28.0	5.8	7.4	58.8
AfriBERTa	0.0	7.4	16.0	76.6
AfroLM	32.8	9.7	18.4	39.1

Table 4: Distribution of African datasets by language family

B Figures

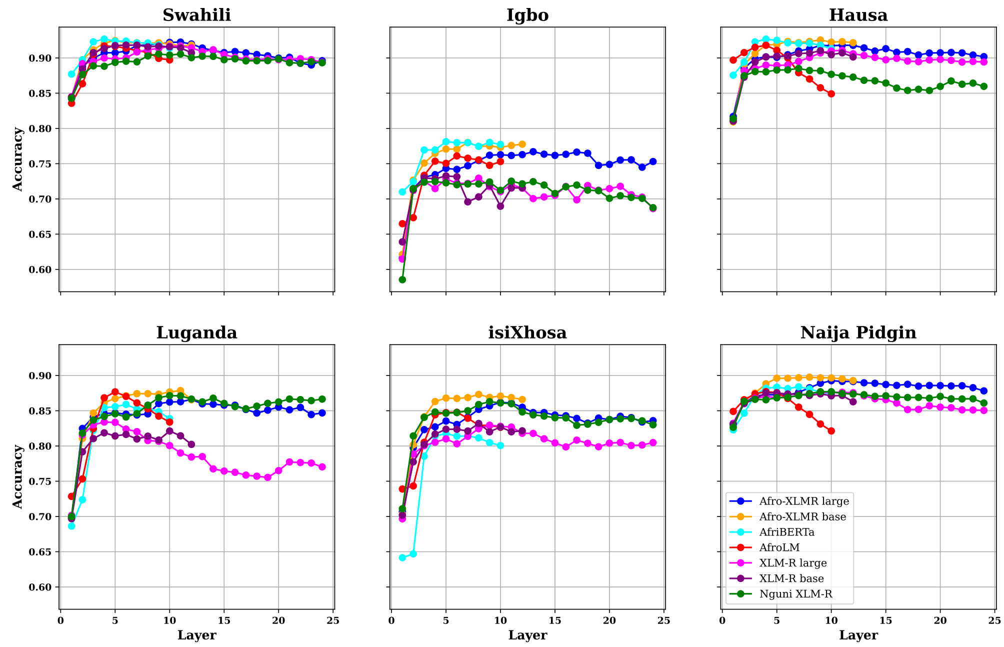

Figure 6: Raw accuracies for POS tagging.

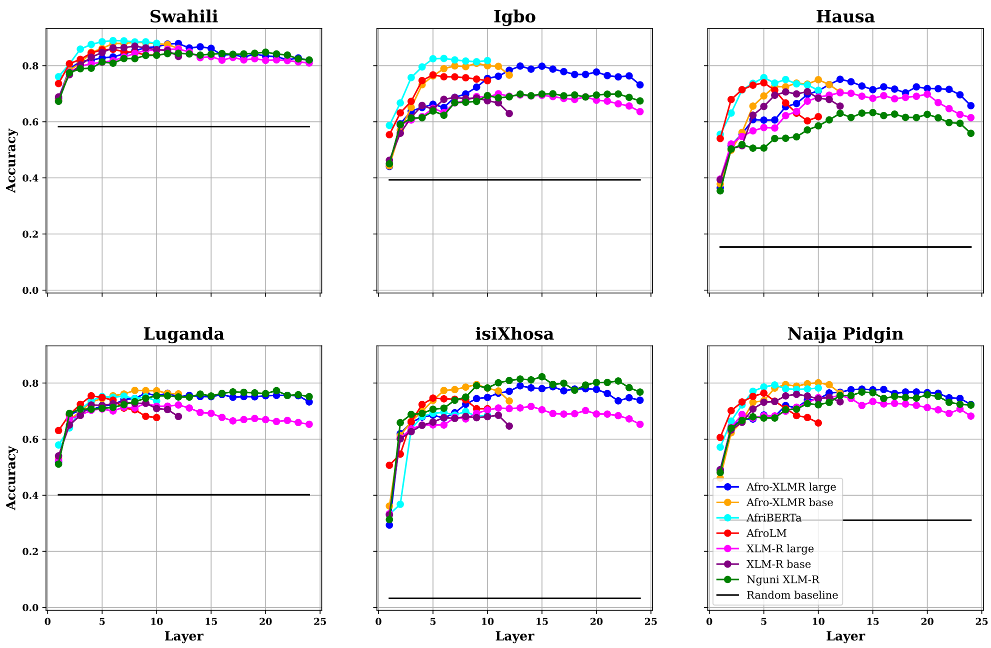

Figure 7: Raw F1-scores for Named Entity Recognition.