Controlled Evaluation of Syntactic Knowledge in Multilingual Language Models

Daria Kryvosheieva
MIT
daria_k@mit.edu

Roger Levy
MIT
rplevy@mit.edu

Abstract

Language models (LMs) are capable of acquiring elements of human-like syntactic knowledge. Targeted syntactic evaluation tests have been employed to measure how well they form generalizations about syntactic phenomena in high-resource languages such as English. However, we still lack a thorough understanding of LMs' capacity for syntactic generalizations in low-resource languages, which are responsible for much of the diversity of syntactic patterns worldwide. In this study, we develop targeted syntactic evaluation tests for three low-resource languages (Basque, Hindi, and Swahili) and use them to evaluate five families of open-access multilingual Transformer LMs. We find that some syntactic tasks prove relatively easy for LMs while others (agreement in sentences containing indirect objects in Basque, agreement across a prepositional phrase in Swahili) are challenging. We additionally uncover issues with publicly available Transformers, including a bias toward the habitual aspect in Hindi in multilingual BERT and underperformance compared to similar-sized models in $\text{XGLM}_{4.5\text{B}}$.

dariakryvosheieva/syntactic_generalization_multilingual

1 Introduction

There is a substantial body of work dedicated to evaluating the linguistic knowledge of language models. Popular evaluation methodologies include:

- probing, i.e., predicting linguistic properties from a network's internal activations (Giulianelli et al., 2018);
- classifying sentences as grammatically acceptable or unacceptable (Warstadt et al., 2019);
- targeted syntactic evaluation (TSE), a method based on comparing LM-assigned probabilities of minimally different sequences (Marvin and Linzen, 2018).

To date, most of the research investigating the linguistic knowledge of LMs has concentrated on high-resource languages such as English (Lin et al., 2019; Warstadt et al., 2020; Hu et al., 2020), German (Mueller et al., 2020; Zaczynska et al., 2020), Spanish (Pérez-Mayos et al., 2021; Bel et al., 2024), Italian (Trotta et al., 2021; Miaschi et al., 2022), Chinese (Wang et al., 2021; Xiang et al., 2021; Zheng and Liu, 2023), and Japanese (Futrell et al., 2018; Someya and Oseki, 2023; Someya et al., 2024). However, efforts have also been made to include less prominent languages. Notably, Torroba Hennigen et al. (2020) and Stanczak et al. (2022) probed masked LMs for morphosyntactic attributes of words across 36 and 43 languages, respectively. Acceptability benchmarks have been developed for North Germanic languages by Volodina et al. (2021), Jentoft and Samuel (2023), Nielsen (2023), and Zhang et al. (2024). TSE has been applied to Hebrew (Gulordava et al., 2018), Norwegian (Kobzeva et al., 2023), and Indonesian and Tamil (Leong et al., 2023).

We believe that evaluating LMs' linguistic knowledge across a diverse range of languages is crucial for developing a comprehensive picture of how they form linguistic generalizations. Assessing 'off-the-shelf' LMs in lower-resourced languages offers a further benefit of diagnosing limitations and challenges these models may face due to issues like insufficient training data, model biases, or difficulties in capturing particular linguistic features. We recognize TSE's advantage of focusing on one combination of linguistic phenomenon and sentence structure at a time, which enables a fine-grained analysis of how performance depends on the structure and complexity of input sentences. While many prior TSE studies considered LMs based on RNN and LSTM architectures (Linzen et al., 2016; Wilcox et al., 2021), we focus on modern LMs based on the Transformer architecture, which is the current state of the art. Therefore,

Proceedings of the First Workshop on Language Models for Low-Resource Languages (LoResLM 2025), pages 438–449
January 20, 2025.

we conduct three TSE case studies benchmarking publicly available Transformer LMs on distinctive morphosyntactic phenomena in low-resource languages—auxiliary verb agreement in Basque, split ergativity in Hindi, and noun class agreement in Swahili.

We find that LMs mostly do well on agreement in Basque, with errors linked to the presence of an indirect object in a sentence, and almost always succeed in selecting the correct aspectual form of the verb based on the presence or absence of the ergative clitic in Hindi, with the exception of multilingual BERT, which prefers the habitual aspect regardless of its grammaticality. However, LMs struggle to agree predicates with the noun class of their subjects in Swahili. Performance on our tests has a positive relationship with model size, but $\text{XGLM}_{4.5\text{B}}$ systematically underperforms similar-sized models for reasons possibly including the lack of low-resource upsampling and the 'curse of multilinguality'. Syntactically complex attractor phrases weaken performance in Swahili but not in Hindi.

2 Evaluation Paradigm

Following existing work on TSE, we organize our evaluation materials in *minimal pairs*—pairs of minimally different sentences such that one is grammatically acceptable and the other one is ungrammatical because it violates a specific linguistic rule. Below we provide an English example where sentences (A) and (B) differ by one property—the plurality of the copula. In sentence (B), the plural copula 'are' does not agree with the singular subject 'The teacher', rendering the sentence ungrammatical.

(A) The teacher is good.

(B) *The teacher are good.

In all our minimal pairs, the sentences differ by one word whose grammaticality can be determined from the left context. We call the left context the *condition* and the rest of the sentence the *target*: in the example above, 'The teacher' is the condition, 'is good.' is the grammatical target, and 'are good.' is the ungrammatical target. We use the `minicons` Python library (Misra, 2022) to compute the conditional log-probabilities of the grammatical and ungrammatical targets given the condition, expecting a model to assign a more positive log-probability to the grammatical target if it has learned the rule correctly. We group minimal pairs into *test suites*, each of which assesses models' knowledge of one phenomenon in sentences of a unified structure, and report a model's accuracy on a test suite as the proportion of minimal pairs for which it assigns higher log-probability to the grammatical target.

3 Test Suites

We consider three low-resource languages from different language families: Basque (isolate), Hindi (Indo-European), and Swahili (Niger-Congo). In each of them, we focus on one characteristic morphosyntactic phenomenon: in Basque, auxiliary verb agreement (§3.1); in Hindi, split ergativity (§3.2); in Swahili, noun class agreement (§3.3). We generate synthetic test suites (§3.4) and perform human validation to verify that the generated minimal pairs represent a genuine contrast in grammatical acceptability (§3.5).

3.1 Auxiliary verb agreement in Basque

The Basque verbal agreement system is more complex than that of most other languages because Basque verbs must agree with all of their arguments—not just the subject but also the direct and indirect object if present in the sentence. Verbs typically consist of a non-finite stem and an auxiliary that carries agreement morphology. For each possible set of arguments—Subject (S); Subject and Direct Object (S DO); Subject, Indirect Object, and Direct Object (S IO DO); Indirect Object and Subject (IO S)—we separately test the agreement of the auxiliary with each argument, resulting in a total of eight test suites. Example 1 below presents a minimal pair from the `basque-DO-S_DO_V_AUX` test suite, which tests the agreement of the auxiliary with the direct object in sentences of the form 'S DO V AUX':

(1.a) *Saltzaileak* salesman.ERG.SG *tomateak* tomato.ABS.PL *prestatu* prepare.PST.PFV *zituen.* PST.3SG>3PL
'The salesman prepared the tomatoes.'

(1.b) **Saltzaileak* salesman.ERG.SG *tomateak* tomato.ABS.PL *prestatu* prepare.PST.PFV *zuen.* PST.3SG>3SG
(ungrammatical)

In sentence (1.b), the auxiliary *zuen* correctly agrees with the singular subject. However, its direct object specification (singular) mismatches the actual number of the direct object (plural). We note that if the subject and the direct object in this example shared the same number, a model's preference for the correct auxiliary *zituen* would be compatible with an incorrect heuristic: associating the infix *-it-* with a plural subject rather than a plural direct object. To control for this, we generate minimal pairs ensuring that the number of the focused argument (here, the direct object) differs from the number of the other arguments.

3.2 Split ergativity in Hindi

Hindi exhibits ergative-absolutive alignment in the perfective aspect and nominative-accusative alignment otherwise. The subject receives the ergative clitic *ने* (*ne*) if and only if the sentence is perfective and transitive. Thus, given an input of the form 'S *ने* O' ('Subject *ne* Object'), models should prefer a perfective verb form over a non-perfective form (habitual or progressive). Conversely, given 'S O' without *ने*, non-perfective forms should be preferred.

We experiment with varying complexities of the direct object noun phrase, which stands between *ने* and the verb. The direct object structures include:

1. Noun (e.g., 'carrot')
2. Possessive pronoun + noun ('their carrot')
3. Possessive pronoun + noun_1 + genitive marker + noun_2 ('their friend's carrot')

For each of these direct object structures, we prepare a test suite with *ने* (ergative-absolutive), where we expect models to prefer a perfective verb, and one without *ने* (nominative-accusative), where we expect them to prefer a non-perfective verb. We select the habitual aspect as the alternative to the perfective in our minimal pairs to avoid possible unnatural combinations of the progressive aspect with stative verbs. Example 2 shows a minimal pair from the `hindi-S_PossPRN_O_V` test suite, which tests whether models prefer the habitual aspect over the perfective aspect in sentences with a 'possessive pronoun + noun' direct object that do not include the *ने* marker.

(2.a) *साँड़ इनकी गाजर खाता है।*
sā̃ṛ inkī gājar khātā hai
bull.M.SG their.F.SG carrot.F.SG eat.HAB.PRS.M.SG
'The bull eats their carrot.'

(2.b) **साँड़ इनकी गाजर खाया है।*
sā̃ṛ inkī gājar khāyā hai
bull.M.SG their.F.SG carrot.F.SG eat.PFV.PRS.M.SG
(ungrammatical)

3.3 Noun class agreement in Swahili

Swahili has a two-dimensional noun class system based on semantic meaning and number, comprising 18 classes in total. Every noun carries a prefix corresponding to its class, although in some cases the prefix may be zero. Typically, a verb must agree with the noun class of its subject, an adjective must agree with the class of the noun it modifies, and the preposition equivalent to English 'of' in possessive constructions ('X of Y') must agree with the class of the possessee (X).

We test the agreement of verbal and adjectival predicates with their subjects in sentences where the subject is modified by a possessor prepositional phrase, which stands between the subject and the predicate and thus serves as a potential distractor. We vary the complexity of the possessor:

1. Noun (e.g., 'scientists')
2. Noun + demonstrative ('these scientists')
3. Noun + demonstrative + adjective ('these old scientists')
4. Noun + demonstrative + adjective + relative verb ('these old scientists that jumped')

We independently vary whether the predicate is a verb or an adjective. To rule out cases where the selection of the correct prefix results from attending to the wrong noun, we ensure that the possessor's noun class is different from that of the subject. Example 3 below is taken from the `swahili-N_of_Poss_D_ni_A` test suite, which tests the agreement of an adjectival predicate with a subject modified by a 'noun + demonstrative' possessor.

(3.a) *Nyumba za wanasayansi hawa wazee ni*
ny-umba z-a w-anasayansi hawa wa-zee ni
10-house 10-of 2-scientist 2.this 2-old is
nyekundu.
ny-ekundu
10-red
'The houses of these old scientists are red.'

(3.b) **Nyumba za wanasayansi hawa wazee ni*
ny-umba z-a w-anasayansi hawa wa-zee ni
10-house 10-of 2-scientist 2.this 2-old is
wekundu.
w-ekundu
2-red
(ungrammatical)

Here, the predicate adjective 'red' (*-ekundu*) can only refer to the subject 'houses' (*nuymba*), not the possessor 'scientists' (*wanasayansi*). Therefore, the correct noun class prefix for the adjective is the one corresponding to 'houses' (class 10), not 'scientists' (class 2).

3.4 Data generation

We adopt the approach from the BLiMP paper (Warstadt et al., 2020) to generate artificial sentences. For each language, we manually assemble a vocabulary of approximately 300 words, annotating them with relevant syntactic, morphological, and semantic properties. We then prepare generation scripts that randomly sample words from the vocabulary according to predefined templates specifying sentence structures and required word properties. Grouping words by semantic categories allows us to generate sentences that sound more or less plausible: for instance, we avoid sampling inanimate nouns as subjects of active verbs, or inedible nouns as objects of the verb 'to eat'. Inflections that follow highly regular patterns—like Basque case endings—are added to stems via rule-based algorithms. Inflected forms that are less regular—such as Hindi case forms—are listed as special entries in the vocabulary. Using this procedure, we generate 1,000 minimal pairs per test suite.

3.5 Human validation

We designed a human experiment on Prolific that mirrored the task given to LMs. For each language, we presented self-reported L1 speakers with a subset of minimal pairs we generated in that language and asked them to select the more grammatically acceptable sentence in every pair. The datasets presented to speakers consisted of five randomly sampled minimal pairs from every test suite associated with the language, resulting in a total of 40 pairs for Basque and Swahili and 30 pairs for Hindi. To ensure that the participants were legitimate speakers of the languages, we additionally included two control minimal pairs in each dataset. In the control pairs, the grammatical sentence was taken from a published text (Basque: *Euskaltzaindiaren Hiztegia* [Dictionary of the Royal Academy of the Basque Language]; Hindi: *Basic Hindi* by Rajiv Ranjan; Swahili: BBC Swahili news reports), and the ungrammatical sentence was created by altering the target word to a nonsensical form (Basque: rewriting the auxiliary backwards; Hindi: replacing the suffix on the aspectual participle with a nonsensical syllable; Swahili: replacing the class prefix on the verb with a nonsensical syllable). The order of the sentences was shuffled. All participants were paid for 20 minutes of work at a rate of $15.50/hour, but only submissions that selected the grammatical sentence in both control trials were considered for analysis. We stopped recruiting new participants once we received ten such submissions per language.

We used BLiMP's threshold for the inclusion of a test suite in the LM experiment: in at least four out of five minimal pairs, the majority of reviewers must have selected the intended-grammatical sentence. All test suites except two Swahili test suites passed the threshold. For the test suites that passed the threshold, we report human accuracy scores (the percentage of times that validators selected the grammatical sentence) together with LM accuracy scores in Figure 1.

4 Models

We evaluated five open-access multilingual Transformers from Hugging Face: three autoregressive models—mGPT, BLOOM, and XGLM—and two masked models—multilingual BERT (mBERT) and XLM-RoBERTa (XLM-R). Each of these models except mBERT is available in several versions differing by size; we evaluated all versions. For an overview of the models and their versions, see Appendix A.

5 Results

We present evaluation results in Figure 1 and provide an overview by language in §5.1. For models that come in several versions, we analyze the relationship between performance on our test suites

and the number of parameters (§5.2). For test suite classes where we varied the complexity of intervening phrases (split ergativity in Hindi and noun class agreement in Swahili), we analyze the relationship between performance and complexity of the intervening constituent (§5.3).

5.1 Overview by language

We find that LMs perform the best on split ergativity in Hindi (average accuracy score 0.873), followed by auxiliary verb agreement in Basque (0.741) and noun class agreement in Swahili (0.504). The top-performing model is mGPT_{13B} overall (average accuracy 0.815), $\text{mGPT}_{1.3B}$ on Basque (0.926), XLM-R_{XXL} on Hindi (0.931), and $\text{XGLM}_{7.5B}$ on Swahili (0.601).

Basque Multiple models, namely $\text{mGPT}_{1.3B}$, mGPT_{13B}, $\text{BLOOM}_{7.1B}$, BLOOM_{176B}, $\text{XGLM}_{1.7B}$, $\text{XGLM}_{2.9B}$, $\text{XGLM}_{7.5B}$, and XLM-R_{XXL}, perform significantly above chance ($p < 0.05$ using a one-sided binomial test) on all Basque test suites. At the same time, at least one model performs significantly below chance ($p < 0.05$) on the `basque-S-S_DO_V_AUX` test suite as well as all test suites containing indirect objects (including IO agreement and agreement with other arguments in the presence of an IO). Our observation that IOs confuse Transformer LMs is consistent with the observation of Ravfogel et al. (2018) that an LSTM-based classifier trained on a morphologically annotated Wikipedia corpus struggles to predict the number of dative arguments of Basque verbs. Ravfogel et al. hypothesized that the low recall scores they obtained on the dative argument plurality prediction task were caused by the relative rarity of dative nouns in the corpus their LSTM was trained on. Our examination of the Universal Dependencies (Nivre et al., 2020) treebank for Basque supports the hypothesis that dative nouns (IOs) are relatively infrequent in the Basque language: the treebank contains a total of 8,595 subject noun phrases, 7,508 direct objects, and 1,021 indirect objects, which means that indirect objects are approximately eight times less frequent than subjects and seven times less frequent than direct objects. We thus hold it plausible that the low frequency of IOs indeed hinders neural networks' ability to learn how they fit into sentences. We note that sentences containing IOs use completely different forms of auxiliaries from sentences without IOs because Basque auxiliaries follow different conjugation paradigms for each set of arguments. For this reason, LMs cannot use information from the more frequent sentences without IOs to infer the conjugations of auxiliaries used in sentences with IOs. Having to learn those conjugations from scratch in a low-frequency setting is what we presume reduces performance.

Hindi Nearly all models perform significantly above chance on all Hindi test suites. The only exception is mBERT, which performs significantly below chance on the three ergative-absolutive test suites. Multilingual BERT displays a bias toward the habitual aspect, preferring it even in constructions where it is ungrammatical because the perfective aspect is expected. If the Universal Dependencies Hindi PUD treebank is representative of broader usage dynamics of aspectual forms in Hindi, this bias is not explained by the relative frequencies of perfective and habitual verbs: in the treebank, perfective forms are slightly more frequent than habitual forms both overall (641 vs. 515) and specifically in 'subject-object-verb' clauses (284 vs. 190).

Swahili LM scores on Swahili test suites concentrate within 0.2 from the random guessing baseline of 0.5, with the exception of three scores above 0.7, obtained by mGPT_{13B}, $\text{XGLM}_{2.9B}$, and $\text{XGLM}_{7.5B}$ on the `swahili-N_of_Poss_V` test suite. Only one model ($\text{XGLM}_{7.5B}$) performs significantly above chance on every Swahili test suite, while three models (BLOOM_{560M}, $\text{BLOOM}_{1.7B}$, $\text{BLOOM}_{7.1B}$) never perform significantly above chance on Swahili test suites. The simplest agreement task we consider, where the subject and a verbal predicate are separated by a simple 'preposition + noun' prepositional phrase, proves to be the easiest for LMs: 10 out of 18 LMs achieve their highest Swahili score on this test suite, this is the only Swahili test suite on which some LM scores exceed 0.7, and it also has the highest number of models performing significantly above chance among Swahili test suites. However, performance plummets once demonstratives and adjectives are added to the intervening phrase (see §5.3 for details).

5.2 Number of parameters and performance

For Transformer LMs, capability is known to improve with size. For example, Kaplan et al. (2020) found a power-law relationship between number of parameters and crossentropy loss on the test set, and Tay et al. (2023) found a positive linear

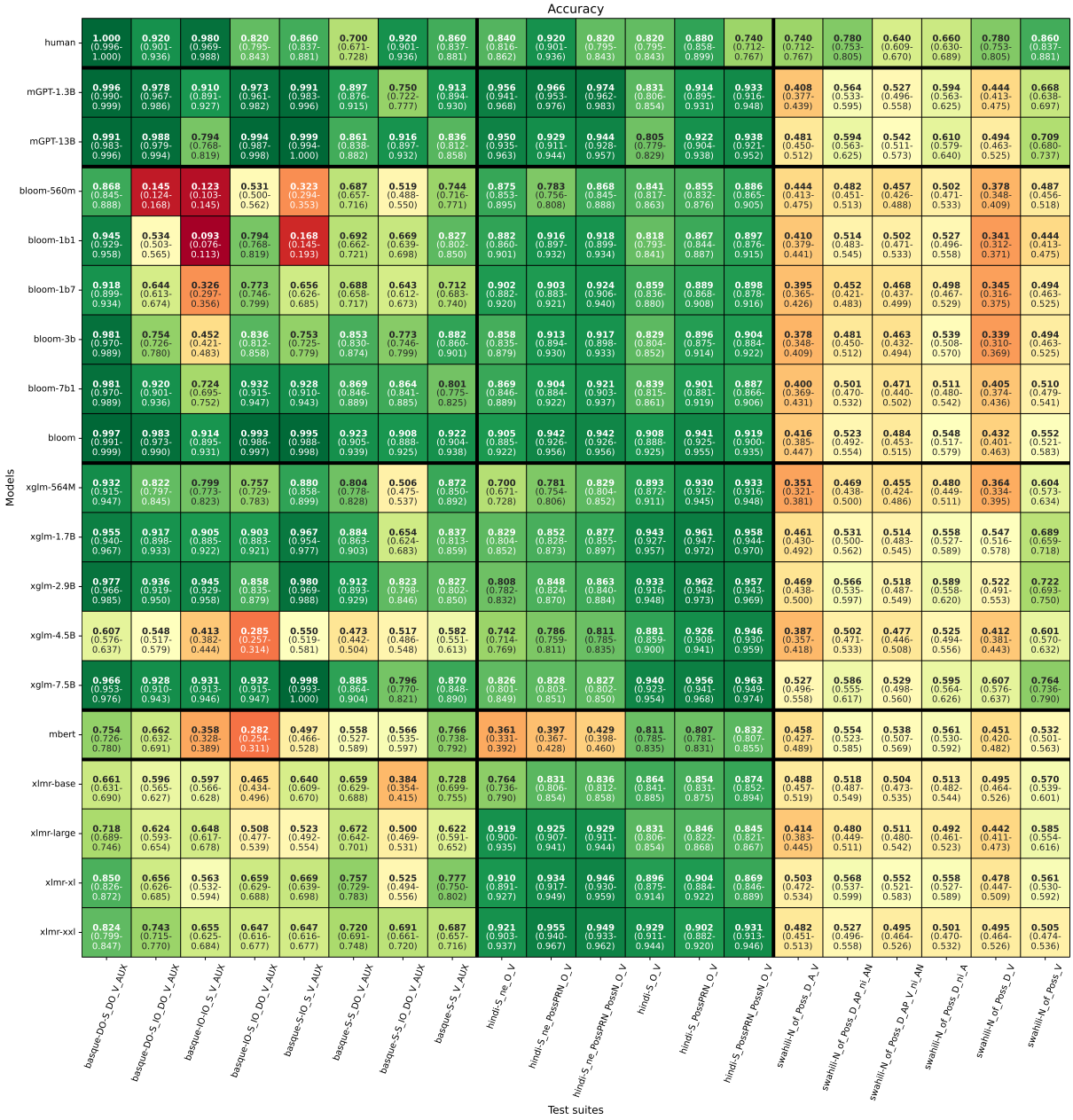

Figure 1: Accuracy scores of the models (vertical axis) on our test suites (horizontal axis). In each cell, the bolded value denotes the fraction of minimal pairs in which the model selected the grammatical target, while values in parentheses denote the left and right 95% confidence intervals. The expectation for random guessing is 0.5.

relationship between number of parameters and an aggregate of GLUE, SuperGLUE, and SQuAD scores. At the same time, Warstadt et al. (2020) argued that GPT-2 model size has no significant effect on BLiMP accuracy.

The fact that mGPT, BLOOM, XGLM, and XLM-R are each available in multiple versions differing by size but having the same architecture and trained on the same corpus provides us with ground for a controlled analysis of the relationship between model size and performance on our test suites. We plot accuracy as a function of parameter count for each test suite and model family and present the plots in Figure 2. In the XGLM family, we exclude $\text{XGLM}_{4.5\text{B}}$ because it was trained on a different variant of the corpus covering more languages. We use linear regression to obtain slopes of best fit lines, which we report in Table 1.[1] We observe that the majority of slopes are positive, including 12 out of 20 slopes in the case of mGPT, all slopes in the case of BLOOM, all slopes except the one corresponding to the `hindi-`

[1]For mGPT, the best fit line is simply the line connecting the accuracy scores of $\text{mGPT}_{1.3\text{B}}$ and $\text{mGPT}_{13\text{B}}$.

S_ne_PossPRN_PossN_O_V test suite in the case of XGLM, and all slopes except three corresponding to Swahili test suites in the case of XLM-R. Furthermore, we find that the average slope over the four model families is positive for all test suites except basque-S-S_V_AUX. Contrary to the finding in the BLiMP paper, this suggests a positive relationship between model size and accuracy.

$\text{XGLM}_{4.5\text{B}}$ shows the poorest performance among XGLM versions on 10 out of 18 test suites (including 7 out of 8 Basque test suites) and the second-poorest performance after the smallest version ($\text{XGLM}_{564\text{M}}$) on the remaining 8 test suites. Additionally, it is outperformed on most test suites by similar-sized non-XGLM models ($\text{mGPT}_{1.3\text{B}}$, $\text{BLOOM}_{1.7\text{B}}$, $\text{BLOOM}_{3\text{B}}$, $\text{BLOOM}_{7.1\text{B}}$, XLM-R_{XL}). Available information about the model's training procedure and dataset is limited, apart from the fact that it was trained on all 134 languages featured in the CC100 XL corpus (Lin et al., 2022), by contrast to other XGLM models, which were trained on a 30-language subset sampled from the same corpus with an upsampling of lower-resourced languages. Therefore, the reasons behind the model's underperformance remain unclear, but we conjecture that the underperformance could be attributed to the lack of low-resource upsampling—in particular, this would explain the low performance on Basque, since the CC100 XL corpus contains much less Basque data (0.35 GiB) than the corpora used to train BLOOM (2.2 GiB) and XLM-R (2.0 GiB)—and the 'curse of multilinguality' (Conneau et al., 2020), the phenomenon that training a small model on many languages leads to performance degradation if the number of languages exceeds a certain threshold. We note that $\text{XGLM}_{4.5\text{B}}$ supports the largest number of languages among all models we consider.

5.3 Robustness to intervening content

Prior studies have yielded different results on the stability of Transformer LMs to intervening constituents: Wang et al. (2021) found that increasing complexity of intervening material causes performance to degrade, Hu et al. (2020) found no significant performance degradation, and the BLiMP paper found that some Transformers are more prone to degradation than others.

Figure 3 shows accuracy as a function of the complexity of the intervening constituent for Hindi and Swahili test suites. It is visually apparent that the general trend is upward (i.e., no degradation

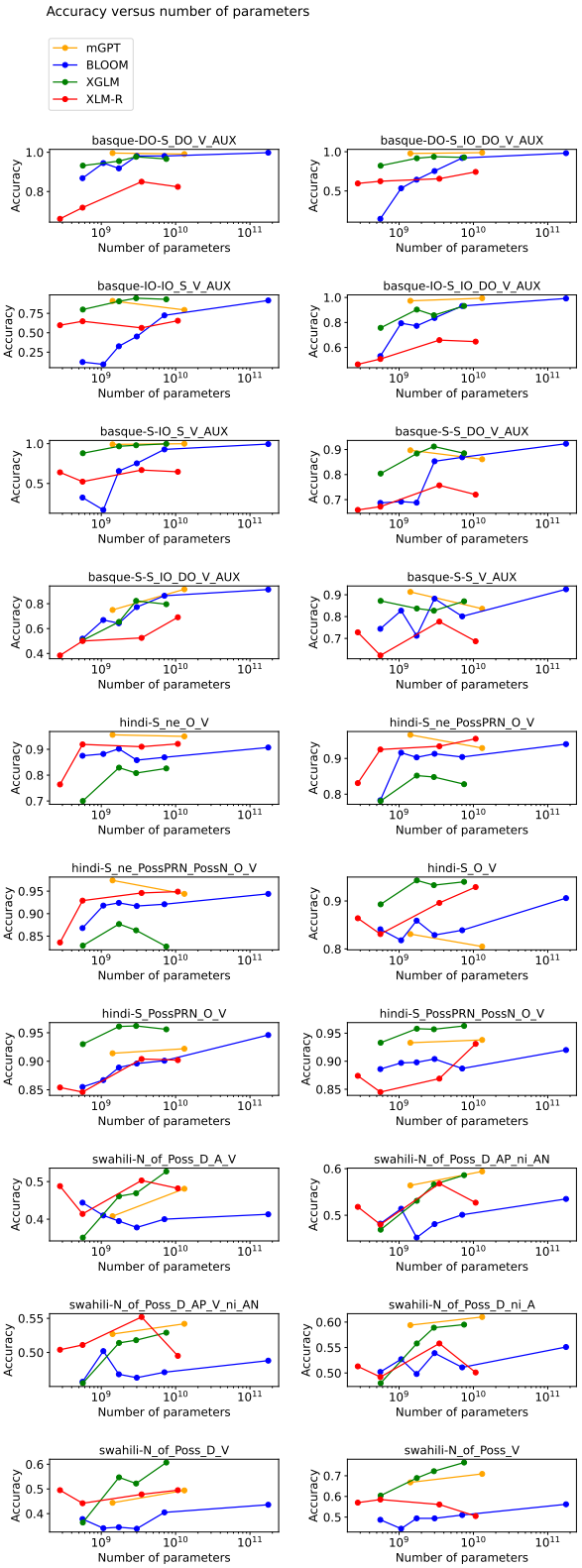

Figure 2: Accuracy as a function of parameter count for each model family and test suite.

at all) for Hindi but downward for Swahili. In Swahili, a particularly sharp drop in accuracy (minus 138.889 correct selections on average) results from the insertion of a demonstrative between the

Test suite	mGPT	BLOOM	XGLM	XLM-R	Average
basque-DO-S_DO_V_AUX	-0.043	0.035	0.371	1.270	0.408
basque-DO-S_IO_DO_V_AUX	0.086	0.231	1.057	1.294	0.667
basque-IO-IO_S_V_AUX	-0.992	0.339	1.377	0.353	0.269
basque-IO-S_IO_DO_V_AUX	0.180	0.131	1.878	1.518	0.927
basque-S-IO_S_V_AUX	0.068	0.258	1.298	0.586	0.553
basque-S-S_DO_V_AUX	-0.308	0.098	0.750	0.502	0.261
basque-S-S_IO_DO_V_AUX	1.420	0.128	3.475	2.437	1.865
basque-S-S_V_AUX	-0.659	0.075	0.192	0.057	-0.084
hindi-S_ne_O_V	-0.051	0.016	1.196	0.785	0.486
hindi-S_ne_PossPRN_O_V	-0.316	0.034	0.302	0.753	0.193
hindi-S_ne_PossPRN_PossN_O_V	-0.257	0.019	-0.313	0.647	0.024
hindi-S_O_V	-0.222	0.041	0.436	0.775	0.258
hindi-S_PossPRN_O_V	0.068	0.035	0.215	0.488	0.202
hindi-S_PossPRN_PossN_O_V	0.043	0.014	0.316	0.693	0.267
swahili-N_of_Poss_D_A_V	0.624	0.006	2.068	0.270	0.742
swahili-N_of_Poss_D_AP_ni_AN	0.257	0.022	1.417	0.241	0.484
swahili-N_of_Poss_D_AP_V_ni_AN	0.128	0.007	0.791	-0.139	0.197
swahili-N_of_Poss_D_ni_A	0.137	0.019	1.291	-0.041	0.352
swahili-N_of_Poss_D_V	0.428	0.041	2.696	0.246	0.853
swahili-N_of_Poss_V	0.351	0.039	1.949	-0.703	0.409

Table 1: Slopes of best fit lines representing the relationship between accuracy (in percentage) and parameter count (in billions) for each model family on each test suite, given to three decimal places.

possessor and a verbal predicate. For both 'preposition + noun + demonstrative' and 'preposition + noun + demonstrative + adjective' PPs, accuracy is lower when the PP stands before a verbal predicate than before an adjectival predicate.

6 Conclusion

We assessed the ability of open-access multilingual Transformer LMs to form syntactic generalizations across three low-resource languages—Basque, Hindi, and Swahili. We found that models mostly performed well on Basque auxiliary agreement, albeit with challenges in sentences containing indirect objects, likely due to their relatively low frequency in training corpora. In Hindi, all LMs demonstrated a solid grasp of split ergativity except multilingual BERT, which failed to select the perfective aspect as the only grammatical aspect in sentences containing the ergative clitic. Noun class agreement in Swahili posed the greatest challenge, with models often performing near random guessing.

We hope that our work will motivate further investigations into LMs' linguistic knowledge in low-resource languages and will help LM developers identify and address areas for improvement, ultimately guiding the design of better LMs for low-resource languages and enabling fair access to high-quality NLP technologies for their speakers.

7 Limitations

First, the present study focuses on one syntactic

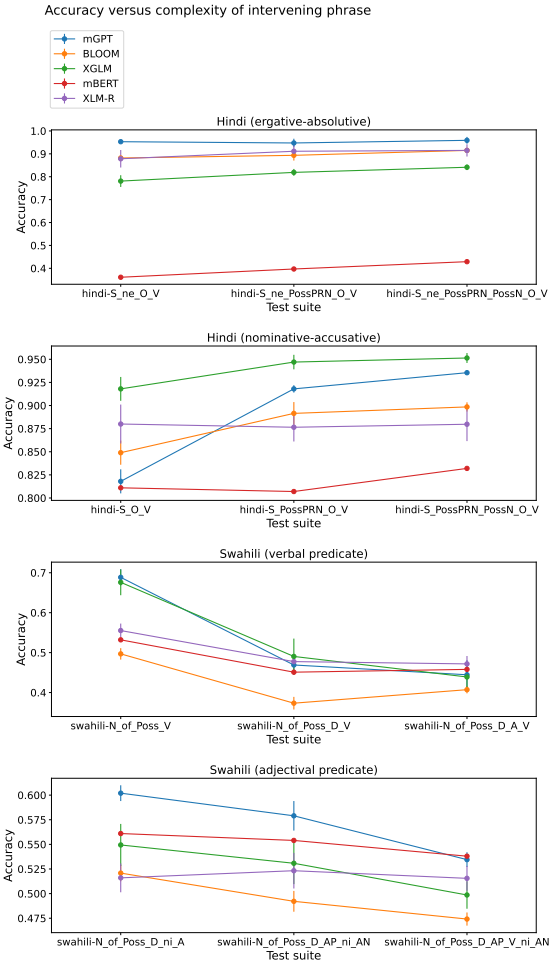

Figure 3: Accuracy as a function of the complexity of the intervening constituent for Hindi and Swahili test suites. For models available in multiple versions, we show mean accuracy over versions; error bars denote 95% confidence intervals on the standard error of the mean.

phenomenon per language, which limits the generalizability of the findings. Future work could provide a more comprehensive analysis by considering multiple phenomena in each language and comparing performance on the same phenomenon across languages.

Second, since demographic data are self-reported on Prolific, we cannot be certain that the participants of our human validation study were true proficient speakers of the languages we investigated. Our method for verifying proficiency, the inclusion of two control minimal pairs in each dataset presented to human reviewers, sometimes allows non-proficient participants to pass: a participant who makes every selection by guessing selects the grammatical option in both control pairs with a 25% probability.

Third, we were unable to find exact figures for the training dataset sizes of mGPT, XGLM (post-upsampling), and mBERT on Basque, Hindi, and Swahili, which limited our ability to analyze the relationship between performance and training dataset size.

References

Nuria Bel, Marta Punsola, and Valle Ruíz-Fernández. 2024. EsCoLA: Spanish corpus of linguistic acceptability. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6268–6277, Torino, Italia. ELRA and ICCL.

BigScience Workshop. 2023. Bloom: A 176b-parameter open-access multilingual language model. *Preprint*, arXiv:2211.05100.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Richard Futrell, Ethan Wilcox, Takashi Morita, and Roger Levy. 2018. Rnns as psycholinguistic subjects: Syntactic state and grammatical dependency. *Preprint*, arXiv:1809.01329.

Mario Giulianelli, Jack Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema. 2018. Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 240–248, Brussels, Belgium. Association for Computational Linguistics.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. Association for Computational Linguistics.

Matias Jentoft and David Samuel. 2023. NoCoLA: The Norwegian corpus of linguistic acceptability. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 610–617, Tórshavn, Faroe Islands. University of Tartu Library.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *Preprint*, arXiv:2001.08361.

Anastasia Kobzeva, Suhas Arehalli, Tal Linzen, and Dave Kush. 2023. Neural networks can learn patterns of island-insensitivity in Norwegian. In *Proceedings of the Society for Computation in Linguistics 2023*, pages 175–185, Amherst, MA. Association for Computational Linguistics.

Wei Qi Leong, Jian Gang Ngui, Yosephine Susanto, Hamsawardhini Rengarajan, Kengatharaiyer Sarveswaran, and William Chandra Tjhi. 2023. Bhasa: A holistic southeast asian linguistic and cultural evaluation suite for large language models. *Preprint*, arXiv:2309.06085.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with

multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open sesame: Getting inside BERT's linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253, Florence, Italy. Association for Computational Linguistics.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Alessio Miaschi, Gabriele Sarti, Dominique Brunato, Felice Dell'Orletta, and Giulia Venturi. 2022. Probing linguistic knowledge in italian neural language models across language varieties. *Italian Journal of Computational Linguistics*, 8.

Kanishka Misra. 2022. minicons: Enabling flexible behavioral and representational analyses of transformer language models. *Preprint*, arXiv:2203.13112.

Aaron Mueller, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, and Tal Linzen. 2020. Cross-linguistic syntactic evaluation of word prediction models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5523–5539, Online. Association for Computational Linguistics.

Dan Nielsen. 2023. ScandEval: A benchmark for Scandinavian natural language processing. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 185–201, Tórshavn, Faroe Islands. University of Tartu Library.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Laura Pérez-Mayos, Alba Táboas García, Simon Mille, and Leo Wanner. 2021. Assessing the syntactic capabilities of transformer-based multilingual language models. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3799–3812, Online. Association for Computational Linguistics.

Shauli Ravfogel, Yoav Goldberg, and Francis Tyers. 2018. Can LSTM learn to capture agreement? the case of Basque. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 98–107, Brussels, Belgium. Association for Computational Linguistics.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Anastasia Kozlova, Vladislav Mikhailov, and Tatiana Shavrina. 2024. mGPT: Few-Shot Learners Go Multilingual. *Transactions of the Association for Computational Linguistics*, 12:58–79.

Taiga Someya and Yohei Oseki. 2023. JBLiMP: Japanese benchmark of linguistic minimal pairs. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1581–1594, Dubrovnik, Croatia. Association for Computational Linguistics.

Taiga Someya, Yushi Sugimoto, and Yohei Oseki. 2024. JCoLA: Japanese corpus of linguistic acceptability. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9477–9488, Torino, Italia. ELRA and ICCL.

Karolina Stanczak, Edoardo Ponti, Lucas Torroba Hennigen, Ryan Cotterell, and Isabelle Augenstein. 2022. Same neurons, different languages: Probing morphosyntax in multilingual pre-trained models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1589–1598, Seattle, United States. Association for Computational Linguistics.

Yi Tay, Mostafa Dehghani, Samira Abnar, Hyung Chung, William Fedus, Jinfeng Rao, Sharan Narang, Vinh Tran, Dani Yogatama, and Donald Metzler. 2023. Scaling laws vs model architectures: How does inductive bias influence scaling? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12342–12364, Singapore. Association for Computational Linguistics.

Lucas Torroba Hennigen, Adina Williams, and Ryan Cotterell. 2020. Intrinsic probing through dimension selection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 197–216, Online. Association for Computational Linguistics.

Daniela Trotta, Raffaele Guarasci, Elisa Leonardelli, and Sara Tonelli. 2021. Monolingual and cross-lingual acceptability judgments with the Italian CoLA corpus. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2929–2940, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Elena Volodina, Yousuf Ali Mohammed, and Julia Klezl. 2021. DaLAJ – a dataset for linguistic acceptability judgments for Swedish. In *Proceedings of the 10th Workshop on NLP for Computer Assisted Language Learning*, pages 28–37, Online. LiU Electronic Press.

Yiwen Wang, Jennifer Hu, Roger Levy, and Peng Qian. 2021. Controlled evaluation of grammatical knowledge in Mandarin Chinese language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5604–5620, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Ethan Wilcox, Pranali Vani, and Roger Levy. 2021. A targeted assessment of incremental processing in neural language models and humans. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 939–952, Online. Association for Computational Linguistics.

Beilei Xiang, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann. 2021. CLiMP: A benchmark for Chinese language model evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2784–2790, Online. Association for Computational Linguistics.

Karolina Zaczynska, Nils Feldhus, Robert Schwarzenberg, Aleksandra Gabryszak, and Sebastian Möller. 2020. Evaluating german transformer language models with syntactic agreement tests. In *Swiss Text Analytics Conference and Conference on Natural Language Processing*.

Ziyin Zhang, Yikang Liu, Weifang Huang, Junyu Mao, Rui Wang, and Hai Hu. 2024. MELA: Multilingual evaluation of linguistic acceptability. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2658–2674, Bangkok, Thailand. Association for Computational Linguistics.

Jianyu Zheng and Ying Liu. 2023. What does chinese bert learn about syntactic knowledge? *PeerJ Computer Science*.

A Models Evaluated

A.1 mGPT

mGPT is an autoregressive model based on GPT-3 architecture introduced in Shliazhko et al. (2024). It supports 61 languages and is trained on a combination of Wikipedia and Colossal Clean Crawled Corpus (C4). mGPT is available in two versions: $\text{mGPT}_{1.3B}$ (1,417,596,928 parameters) and mGPT_{13B} (13,108,070,400 parameters).

A.2 BLOOM

BLOOM is an autoregressive model developed over the course of a year-long open research workshop involving more than a thousand researchers (BigScience Workshop, 2023). The model supports 46 natural languages and 13 programming languages and was trained on the ROOTS corpus, a composite collection of 498 Hugging Face datasets compiled by BigScience. BLOOM is available in six versions, ranging from 560 million to 176 billion parameters (see Table 2).

Version	Number of parameters
BLOOM_{560M}	559,214,592
$\text{BLOOM}_{1.1B}$	1,065,314,304
$\text{BLOOM}_{1.7B}$	1,722,408,960
BLOOM_{3B}	3,002,557,440
$\text{BLOOM}_{7.1B}$	7,069,016,064
BLOOM_{176B}	176,247,271,424

Table 2: BLOOM versions.

A.3 XGLM

XGLM is an autoregressive model developed by Meta (Lin et al., 2022), trained on a corpus covering 68 Common Crawl snapshots. The model is available in five versions (see Table 3). $\text{XGLM}_{4.5B}$ supports 134 languages, while other versions support 30 languages.

Version	Number of parameters
XGLM_{564M}	564,463,616
$\text{XGLM}_{1.7B}$	1,732,907,008
$\text{XGLM}_{2.9B}$	2,941,505,536
$\text{XGLM}_{4.5B}$	4,552,511,488
$\text{XGLM}_{7.5B}$	7,492,771,840

Table 3: XGLM versions.

A.4 Multilingual BERT

Multilingual BERT (mBERT) is the multilingual variant of BERT, an encoder-only Transformer developed by Google for the masked language modeling objective (Devlin et al., 2019). The model contains 177,974,523 parameters, was trained on Wikipedia, and supports 104 languages with the largest Wikipedias at the time of training.

A.5 XLM-RoBERTa

XLM-RoBERTa (XLM-R) is a masked model developed by Facebook with the goal of improving upon previous state-of-the-art models such as mBERT (Conneau et al., 2020). The model was trained on a cleaned version of the Common Crawl corpus that encompasses 100 languages, including 93 natural languages, the constructed language Esperanto, five romanizations of South Asian languages typically written in non-Latin scripts, and a variant of Burmese written in the non-Unicode-compliant Zawgyi font. XLM-RoBERTa is available in four versions, as outlined in Table 4.

Version	Number of parameters
XLM-R_{Base}	278,295,186
XLM-R_{Large}	560,142,482
XLM-R_{XL}	3,482,741,760
XLM-R_{XXL}	10,712,994,816

Table 4: XLM-RoBERTa versions.