MultiBLiMP 1.0:
A Massively Multilingual Benchmark of Linguistic Minimal Pairs

Jaap Jumelet[1], **Leonie Weissweiler**[2], **Joakim Nivre**[2], and **Arianna Bisazza**[1]
[1]University of Groningen, The Netherlands [2]Uppsala University, Sweden
j.w.d.jumelet@rug.nl,
{leonie.weissweiler,joakim.nivre}@lingfil.uu.se, a.bisazza@rug.nl

Abstract

We introduce MultiBLiMP 1.0, a massively multilingual benchmark of linguistic minimal pairs, covering 101 languages and 2 types of subject-verb agreement, containing more than 128,000 minimal pairs. Our minimal pairs are created using a fully automated pipeline, leveraging the large-scale linguistic resources of Universal Dependencies and UniMorph. MultiBLiMP 1.0 evaluates abilities of LLMs at an unprecedented multilingual scale, and highlights the shortcomings of the current state-of-the-art in modelling low-resource languages.[1]

1 Introduction

Large language models (LLMs) are often trained on highly multilingual corpora, which enable users to interact with them in a wide range of languages (Grattafiori et al., 2024; Üstün et al., 2024). Multilingual evaluation of LLMs, however, has mostly focused on their *functional linguistic competence* through tasks requiring world knowledge and language understanding (Singh et al., 2024, 2025), while little work has assessed their *formal linguistic competence*, i.e., the ‘‘knowledge of rules and statistical regularities of a language’’ (Mahowald et al., 2024). In multilingual contexts, the latter is commonly approximated intrinsically through perplexity (Chang et al., 2024b) or extrinsically through performance in generative tasks such as translation or summarization (Dang et al., 2024). These approaches do not truly disentangle formal from functional competence. Moreover, they are very coarse-grained and do not inform us on which specific constructions a model does (not) master.

Both shortcomings are addressed by the design of targeted syntactic evaluation benchmarks, typically structured as pairs of grammatical/ungrammatical sentences differing by a single syntactic aspect (Linzen et al., 2016; Warstadt et al., 2020), where a formally competent LM is expected to assign higher probability to the grammatical version. Such datasets, however, exist only for English and a few other, mostly high-resource languages (Gulordava et al., 2018; Mueller et al., 2020; Taktasheva et al., 2024).

To accelerate progress in this direction, we introduce **MultiBLiMP 1.0**, a massively multilingual benchmark of linguistic minimal pairs covering two types of subject-verb agreement (subject-finite-verb and subject-participle) for number, person, and gender; created automatically using two large-scale linguistic resources: Universal Dependencies (UD, Nivre et al., 2016, 2020; de Marneffe et al., 2021) and UniMorph (Batsuren et al., 2022). Multi-BLiMP is not only a *benchmark*, but also a *pipeline* for the automatic creation of highly multilingual benchmarks (Figure 1), which can scale to many more linguistic phenomena.

Taking subject-finite-verb[2] and subject-participle agreement as a use case, we present a first version of the benchmark including more than 128,000 minimal pairs across 101 languages. We use this to evaluate 42 LLMs, finding that linguistic competence is strongly driven by model size and language frequency in training data (Figure 7), is acquired during pre-training, and can deteriorate as a result of post-training.

Besides its practical relevance for LLM evaluation, MultiBLiMP enables the study of the learnability of specific linguistic constructions cross-lingually in a unified framework, at a much larger scale than currently possible (Mueller et al., 2020), with future applications in LM interpretability (Brinkmann et al., 2025) and quantitative typology (Levshina et al., 2023; Baylor et al., 2024).

[1]Code: github.com/jumelet/multiblimp.
Data: huggingface.co/datasets/jumelet/multiblimp.

[2]Referred to in the following as subject-verb agreement for simplicity.

Transactions of the Association for Computational Linguistics, vol. 14, pp. 193–216, 2026. https://doi.org/10.1162/TACL.a.600
Action Editor: Zeljko Agic. Submission batch: 4/2025; Revision batch: 7/2025; Published 1/2026.

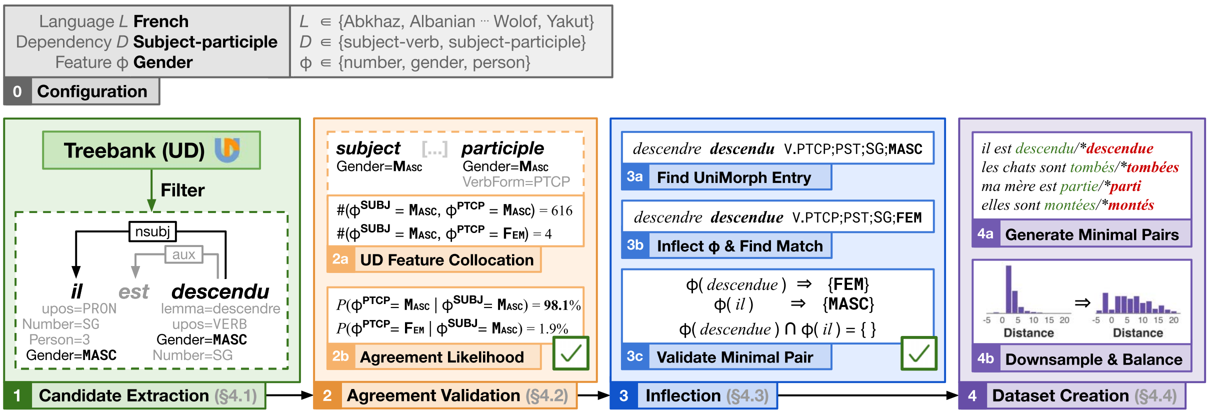


Figure 1: Pipeline of the minimal pair creation procedure of MultiBLiMP 1.0.

2 Background

2.1 Targeted Syntactic Evaluation

The syntactic abilities of language models have commonly been evaluated using syntactic minimal pairs: sentence pairs that are altered in a minimal way to create a specific grammaticality violation. By comparing a model's probability judgment of the grammatical sentence to that of the ungrammatical one, we can assess whether it has acquired a notion of the underlying phenomenon (we provide a technical description in §6.1).

This approach to test language models was introduced by Linzen et al. (2016), focusing on English subject-verb number agreement. Specifically, they extracted sentences from Wikipedia and flipped the number of the verb without detecting the subject, finding that models at that time struggled considerably with increased distance between the subject and the verb, as well as intervening nouns and other attractors. This work was extended by Marvin and Linzen (2018), who generated sentences with context-free grammars, enabling them to increase the difficulty of agreement across phrases and clauses. BLiMP (Warstadt et al., 2020) extended this approach to a wider range of linguistic phenomena, generating minimal pairs with templates for many linguistic phenomena by sampling from a vocabulary. Hu et al. (2020) and Gauthier et al. (2020) both present similar large-scale collections of syntactic evaluation for English.

Rather than relying on minimal pairs, Pratapa et al. (2021) develop a method to assess the morphosyntactic well-formedness of model outputs based on its dependency parse and a set of rules automatically extracted from parsed corpora. Their approach is complementary to ours as it focuses on the well-formedness of a machine-generate text as opposed to measuring the intrinsic syntactic abilities of a language model on controlled stimuli. However, the requirement of a well-performing parser in each language of interest considerably limits the applicability of their method to a wide set of languages.

Monolingual Non-English Benchmarks Minimal pair benchmarks have recently been developed for various other languages, using different methods. Some use existing annotated sentences from UD, some craft templates, and some others are built by translating an English dataset.

For Chinese, Xiang et al. (2021, CLiMP) generate minimal pairs with templates adapted from BLiMP, while Song et al. (2022, SLING) edit a constituency treebank with templates and then verify the results with human annotators. Liu et al. (2024, ZhoBLiMP) employ a mixture of adapting paradigms from English BLiMP, and extracting Chinese-specific paradigms from syntax journals and textbooks. Someya and Oseki (2023, JBLiMP) extract minimal pairs from journal articles on Japanese syntax. Suijkerbuijk et al. (2025, BLiMP-NL) create ten sentence pairs per paradigm by hand for Dutch, as well as 90 more sentences generated by ChatGPT and then hand-checked. Leong et al. (2023, LINDSEA) present minimal pairs for Tamil and Indonesian which are created using prompting with GPT-3.5 and GPT-4. Kryvosheieva and Levy (2025) create minimal pairs specifically for Swahili noun class agreement, Hindi split ergativity, and Basque verb agreement, each chosen for their complexity.

Closer to our approach, Taktasheva et al. (2024, RuBLiMP) create pairs from UD-parsed sentences using perturbation rules. Our approach, too, leverages UD to identify syntactic relations, which we extend multilingually by setting up a language-agnostic minimal pair creation procedure (§4).

Multilingual Two multilingual minimal pair benchmarks are known to us. The largest, CLAMS (Mueller et al., 2020), creates pairs for English, German, Hebrew, Italian, and Russian for subject-verb agreement by having native speakers translate the sentences from Marvin and Linzen (2018) where applicable. Gulordava et al. (2018) similarly cover English, Hebrew, Italian, and Russian for various types of agreement using sentences from UD treebanks, in which content words are swapped out with others of the same part-of-speech to test the models' reliance on semantics. Two other notable related efforts present a multilingual benchmark for linguistic acceptability—for which a model is fine-tuned to predict the grammaticality of a single sentence instead of judging a pair (Warstadt et al., 2019)—ScaLA, which is part of ScandEval (Nielsen, 2023), covering 5 Scandinavian languages, and MELA (Zhang et al., 2024), covering a diverse set of 10 languages.

We see MultiBLiMP 1.0 as complementary to all these efforts. While language-specific benchmarks can provide tremendous depth on the syntactic system of a language, it will prove challenging to scale such efforts to a wide variety of languages. Our approach, on the other hand, focuses on 'wide coverage' and can therefore be of use for both the evaluation of highly multilingual LLMs and for quantitative typological studies.

2.2 Representation Sharing in LLMs

Various works have evaluated the degree to which multilingual models make use of shared syntactic representations across languages. Early studies analyzed mBERT (Devlin et al., 2019) and smaller-scale LSTM models trained on a few languages, finding mixed evidence of sharing (Pires et al., 2019; Mueller et al., 2020; Dhar and Bisazza, 2021; Chi et al., 2020; Choenni and Shutova, 2022). Focusing on modern-scale LLMs, Wendler et al. (2024) find that models trained on English-dominated corpora use English to some degree as an internal 'pivot language' to solve tasks in other languages. Brinkmann et al. (2025) find significant overlap between the representations for number, gender, tense and other morphosyntactic features. Similarly, Stanczak et al. (2022) find overlap in the presentations of morphosyntactic features, particularly for grammatical number. Ferrando and Costa-jussà (2024) show the similar structures of circuits for subject-verb agreement in English and Spanish. The emergence of this kind of interpretability findings highlights the exciting opportunities presented by MultiBLiMP 1.0 for massively multilingual interpretability.

3 Linguistic Resources

We briefly describe here the key linguistic resources that we use to create MultiBLiMP: Universal Dependencies and UniMorph.

3.1 Universal Dependencies

Universal Dependencies (Nivre et al., 2016, 2020; de Marneffe et al., 2021) is a multilingual treebank collection containing 296 treebanks for 168 languages. Besides syntactic dependency relations, UD also contains rich morphological feature annotations that we leverage in our pipeline. An example of a UD-annotated sentence can be found in Figure 1, step 1. We use UD version 2.15 for the 1.0 release of MultiBLiMP (Zeman et al., 2024).

We exclude treebanks of spoken language, historical varieties, as well as certain genres, e.g., technical manuals. Generally, we only exclude treebanks for low-resource languages where absolutely necessary, while for high-resource languages, we also exclude very small treebanks that will be of little additional benefit, but may introduce annotation inconsistencies. A full list of the excluded treebanks with reasons for exclusion can be found in Appendix A.1. We then consider the union of all remaining treebanks for each language.

3.2 UniMorph

UniMorph (UM, `v4.0`, Batsuren et al., 2022) is a multilingual collection of morphological feature annotations, which define word-level features such as NUMBER, MOOD, or GENDER for nouns, verbs, and adjectives. It covers 183 languages in total, 81 of which are also covered by UD. Similarly to UD, it uses a universal set of features (Sylak-Glassman et al., 2015), making it highly suitable for our goal of defining a language-independent inflection mechanism to create minimal pairs. UM entries contain the lemma and morphological features of a word form, making it possible to efficiently disambiguate and create inflections. For example, the English word *saw* is represented with the following four entries, expressing its various meanings:

lemma	form	features
see	saw	V;PST
saw	saw	N;SG
saw	saw	V;NFIN;IMP+SBJV
saw	saw	V;PRS;3;IND;PL

Note that features can be left implicit: The first entry of past tense *see* does not specify PERSON or NUMBER, indicating that this form covers all values of those features. We preprocess UM features to be compatible with UD feature values (`SG` → *Sing*, `PRS` → *Pres*, etc.), and transliterate languages that use different scripts from those in UD.

UD Features To broaden the number of languages for which we can create inflections, we leverage the morphological feature annotations that are present in UD, extending the approach of UDLexicons (Sagot, 2018)—which covers 38 languages—to the 142 languages in UD that contain annotations. For each language, we extract the ⟨*lemma*, *form*, *features*⟩ triplets for each token in the treebank. To ensure the quality of these triplets, we incorporate their frequency to filter out potential annotation errors. We make the assumption that a ⟨*lemma*, *features*⟩ tuple should map onto a single *form*; if the tuple yields multiple forms, it is likely that a feature has been annotated erroneously.[3] In such cases, we discard all entries occurring three times less than the most frequent entry. This procedure results in 4.2M unique triplets for 142 languages, 61 of which are not covered by UM.

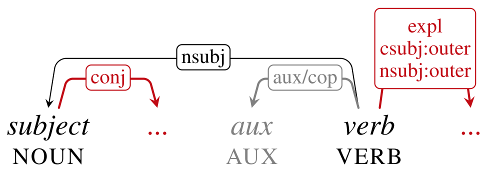
4 Pipeline

Our pipeline for creating minimal pairs consists of four stages. First, we **extract** suitable candidates for a phenomenon using dependency parse trees (§4.1). This allows us to determine the key items of interest for a phenomenon, like the subject and verb in subject-verb agreement. Next, we **validate** whether the agreement phenomenon is present in that language based on UD collocation statistics (§4.2). We then **inflect** one of the words with respect to a particular *feature* (e.g., NUMBER) to create an agreement violation (§4.3). By replacing the item with this inflected form, we create the ungrammatical counterpart for the minimal pair. Finally, we create a minimal pair dataset for each language, **balancing** the morphological features and number of pairs (§4.4). We now explain this procedure in more detail; a graphical overview is provided in Figure 1.

4.1 Candidate Extraction

The 167 UD languages form the initial set of candidate languages we consider for our benchmark. After having excluded some treebanks as described in §3.1, we also set various sentence-level constraints. Specifically, we discard sentences containing `reparandum` dependencies or a `Style`, `Foreign`, or `Typo` feature, which typically signal sentences that are malformed in some way.

Filters To extract candidates for a particular linguistic phenomenon, we define filters based on dependency edges, part-of-speech tags, and morphological features. In this paper, we restrict our attention to various forms of subject-verb and subject-participle agreement. The filters for this phenomenon are defined as follows:



That is, we look for `nsubj` dependency edges between a NOUN and a VERB and collect any auxiliary verb connected to it. If the VERB or one of its auxiliaries has the feature VERBFORM=FINITE, it is considered as a candidate for subject-verb agreement. For finding subject-participle agreement relations, we place an additional constraint on the morphological features of the verb of VERBFORM=PART. We drop subjects containing conjunctions, since these may have conflicting feature values. Furthermore, we drop verbs that have an outgoing `expl`, `csubj:outer`, or

[3]For example, ⟨*walk*, TENSE=PRES; PERSON=3; MOOD=IND; NUMBER=SING⟩ should map only to *walks*; if it would also yield *walk* it indicates an annotation error. Since such annotation errors are infrequent, we select the most frequent item instead.

`nsubj:outer` dependency, since these indicate clausal subject constructions that may invalidate our minimal pair formation methodology.

4.2 Agreement Validation

Determining whether a particular phenomenon exists in a given language, and how it can be expressed as a minimal pair, is challenging. We address this with a data-driven procedure that computes the probability of a language containing a phenomenon, based on UD and UM annotations.[4] This acts as a final filter on the minimal pairs, ensuring that our perturbed sentences are actually ungrammatical in a given language.[5]

Agreement The agreement phenomena we focus on can be turned into a minimal pair by changing the agreement feature on a word and re-inflecting it:

The boys walk ⇒ **The boys walks*
N;PL V;**PL** N;PL V;**SG**

Evaluating agreement in this way requires that the inflected form is ungrammatical in this context. For example, if we were to inflect a past tense verb for NUMBER in English, the inflected form would remain the same. Another example is Turkish, where number is marked *optionally* on the verb:

Oğlanlar yürüyorlar ⇒ *Oğlanlar yürüyor*
N;PL V;**PL** N;PL V;**SG**

Both these constructions are correct, and the construction with the singular verb is even more frequent. A model provided with this minimal pair would likely assign a higher probability to the sentence with the singular verb and be penalized for it, resulting in an incorrect assessment of the model's grammatical abilities.

Requirements Based on these observations, we define two requirements to determine if a language contains agreement for some feature ϕ (Baker, 2008):

R1: ϕ must be specified for both elements of the agreement relation, either lexically inherent (*la fille est tombée*; *the girl fell*) or morphologically realized (*les garçons sont tombés*; *the boys fell*). This way, for example, we rule out subject-verb gender agreement in English, since gender is not specified on verbs. We easily validated this using UniMorph and UD by checking whether both subjects and verbs are annotated for the feature.

R2: Re-inflecting one of the elements with a different value for ϕ for one of the elements must result in a grammatical violation. This requirement is more challenging to test.

Collocations We approach Requirement 2 from a collocational perspective—a similar approach was used by Chaudhary et al. (2020) for grammatical rule detection from treebanks. In a language having strict subject-verb agreement for ϕ, we expect very high co-occurrence between nouns and verbs of the same feature value, and low co-occurrence between nouns and verbs of contrasting values. This can be expressed as a conditional probability:

$$P_{agr}(\phi^v = x | \phi^n = x, \mathrm{WO}) \approx 1 \qquad (1)$$

where ϕ^v and ϕ^n denote the feature of verb and noun, x a specific feature value (e.g., SG or PL), and WO the subject-verb word order (SV/VS). If this probability is close to 1, it implies that the feature of the noun co-occurs highly with that of the verb, and the probability of the verb having a contrasting feature value is close to 0.

Agreement can depend on various extra factors, such as word order. For example, person agreement in Dutch depends on this: Second person singular verbs in VS order lose their second person ending and coincide with first person (*jij doet* vs. *doe jij*). A different example is French, for which subject-participle agreement depends on the auxiliary: Participles inflect for NUMBER and GENDER with an *être* auxiliary, but not with an *avoir* auxiliary (*elles sont montées* vs. *elles ont monté*). Arabic has flexible subject-verb order and does not inflect for plural in VS order, but only has agreement in SV order for human subjects. In our pipeline we therefore condition Equation 1 on a binary word order feature, but leave a more fine-grained exploration of agreement conditions open for future work, for instance by extending the data-driven agreement extraction approach of

[4] An alternative to approaching this from a data-driven perspective would be to use typological databases like Grambank (Skirgård et al., 2023) and WALS (Dryer and Haspelmath, 2013), but their discrete nature and framing agreement in terms of being *possible* and not *compulsory* (Baylor et al., 2023) make them unsuitable for our goals.

[5] In addition to this data-driven procedure, we also conduct extensive spot checks to ensure our procedure yields valid minimal pairs.

ϕ^n	WO	ϕ^v	$P_{agr}(\phi^v \mid \phi^n, \mathrm{WO})$ Dutch	Turkish
PL	SV	PL	**0.989**	0.161
		SG	0.011	0.839
PL	VS	PL	**0.990**	0.000
		SG	0.010	1.000
SG	SV	PL	0.012	0.025
		SG	**0.988**	**0.975**
SG	VS	PL	0.008	0.018
		SG	**0.992**	**0.982**

Table 1: Dutch and Turkish agreement probabilities for subject-verb number agreement. Significant agreement is denoted in **boldface**. WO stands for word order, either subject verb (SV) or verb subject (VS).

Chaudhary et al. (2020) with word order and lexical features.

Agreement Conditions We compute agreement probabilities for each language, verb type (finite main verb, finite auxiliary or participle), feature value, and word order. We refer to this configuration as an **agreement condition**. Probabilities for each agreement condition are computed by counting feature co-occurrences based on the morphological annotations in the UD tree, using the dependency relations we extract from UD (§4.1). In case a feature is not annotated, we look it up in UniMorph and, if present, take the value from there. We only count co-occurrences for cases where a feature is defined for both the subject and the verb (e.g., NUMBER is not specified for past tense verbs in English). Based on the co-occurrence counts of the feature values, we then compute the conditional probability distribution of Equation 1.

Note that it is possible this way for a language to have strict agreement in only one direction. We show an example in Table 1, providing subject-verb number agreement probabilities for two languages: Dutch, which has strict number agreement for all conditions, and Turkish, where plurality is optionally marked on the verb, but singular verbs are commonly used for both singular and plural subjects. Dutch number agreement is marked as significant for both singular and plural, whereas Turkish agreement only demonstrates agreement for singular items.

Significance Since our procedure depends on the amount of co-occurrence counts we can extract from UD for a language, we need to determine whether the collocation probability is statistically significant. For this, we conduct a binomial test to assess whether Equation 1 was significantly greater than the baseline $p_0 = 0.9$. To rule out agreement, we conduct a binomial test in the opposite direction. If neither test is significant, we mark agreement as being *uncertain*. For significance we use $\alpha = 0.1$ with Bonferroni correction. We only proceed to the minimal pair creation step for the agreement conditions that have certain and uncertain agreement. The conditions with no agreement are discarded.

4.3 Inflection

To form the ungrammatical counterpart of a sentence extracted from UD, we inflect a specific word to a different feature that makes the sentence ungrammatical. For example, to create a minimal pair for subject-verb number agreement for *The boy walks*, we re-inflect the verb *walk* with NUMBER=PLUR to create **The boy walk*. While we focus here on GENDER, PERSON, and NUMBER, this can work for any agreement feature annotated on the verb.

Our inflection procedure has two stages. We start by defining the feature to inflect for, with a candidate word from the filtered treebank of § 4.1. The first stage is to find the corresponding lemma and morphological features of this word in the UD and UM databases, based on its form and the features that were annotated in the UD tree. This may yield multiple candidate lemmas and features, depending on the level of detail of the UD features. In the second stage, we then find all matching rows containing *contrasting values* of the feature that we inflect for. The inflected forms of these matching rows are the inflection candidates that we use to create the minimal pairs. By considering all opposite values we may create multiple pairs from a single sentence, e.g., *He is [..]* $\rightarrow$ *He are/am [..]*.

4.4 Dataset Balancing

To level data imbalances, we limit the number of minimal pairs per agreement condition to 100 items.

We obtain these 100 items using a weighted downsampling procedure that also balances sentence-level features. The features we incorporate in this procedure are: *subject form*, *verb*

Number	SG		PL		DU		Total		
	SV	VS	SV	VS	SV	VS	SV	VS	BOTH
S-Verb	88 (74)	82 (63)	73 (58)	58 (43)	8 (3)	3 (2)	89 (79)	84 (65)	90 (80)
S-Participle	35 (33)	28 (21)	23 (18)	19 (13)	3 (0)	1 (0)	35 (33)	30 (22)	35 (33)
Gender	MASC		NEUT		FEM		Total		
	SV	VS	SV	VS	SV	VS	SV	VS	BOTH
S-Verb	31 (16)	27 (16)	13 (7)	13 (6)	30 (13)	25 (11)	34 (19)	30 (16)	34 (22)
S-Participle	26 (25)	25 (20)	11 (8)	11 (8)	20 (17)	20 (12)	26 (25)	26 (21)	26 (25)
Person	1		2		3		Total		
	SV	VS	SV	VS	SV	VS	SV	VS	BOTH
S-Verb	71 (42)	55 (25)	63 (32)	38 (15)	83 (62)	69 (38)	86 (66)	74 (44)	87 (68)
S-Participle	2 (1)	0 (0)	1 (1)	0 (0)	2 (2)	1 (0)	3 (2)	1 (0)	3 (2)
Total							100 (85)	90 (66)	101 (87)

Table 2: The number of unique languages that yielded at least 10 minimal pairs, for each agreement condition. The number between brackets denotes the number of languages for which the binomial test of §4.2 was significant. SV and VS denote the subject-verb word order.

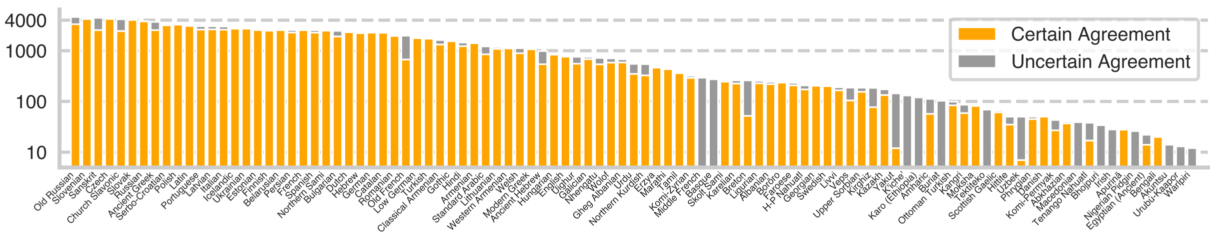


Figure 2: Number of minimal pairs per language in MultiBLiMP, split out for certain and uncertain agreement cases using the agreement detection procedure of §4.2. Note the log-scale on the *y*-axis.

form, *subject-verb distance*, *all attractors congruent*, and *all attractors incongruent*. The last two features indicate whether the intervening material between subject and verb contained items that had a congruent or incongruent feature with respect to the agreement feature that we re-inflect (e.g., *The* ***keys*** *to the* <u>*cabinet*</u> ***are***). We define the sampling probability Q of a minimal pair as the inverse of the joint probability of its features $P(x)$, assuming feature independence:

$$P(x) = \prod_i P(x_i) \qquad Q(x) = \frac{P(x)^{-1}}{\sum_i P(x_i)^{-1}}$$

The feature probability $P(x_i)$ is expressed as the relative frequency of the feature. By sampling from Q (without replacements), we can balance all features simultaneously. This results in a dataset that is as balanced as possible for multiple features, ensuring both lexical diversity and an even spread of easier and harder sentences.

5 MultiBLiMP 1.0

We run our minimal pair creation pipeline for two agreement phenomena and three features: subject-verb and subject-participle agreement for NUMBER, PERSON, and GENDER. Within each condition we create sub-conditions based on the inflected feature (SG → PL, 2 → 3, etc.), and the order of subject and verb. Our procedure results in **128,321 minimal pairs** across **101 languages**. The total number of minimal pairs we obtain before balancing is 1.4 million. We provide an aggregated overview of the number of languages per condition in Table 2, and a sample of pairs in Appendix A.2. Subject-verb number agreement is the most common agreement type, with 90 languages, whereas we find subject-participle person agreement for only 3 languages. We present the number of minimal pairs per language in Figure 2. We also provide a detailed breakdown of the language families that are currently covered

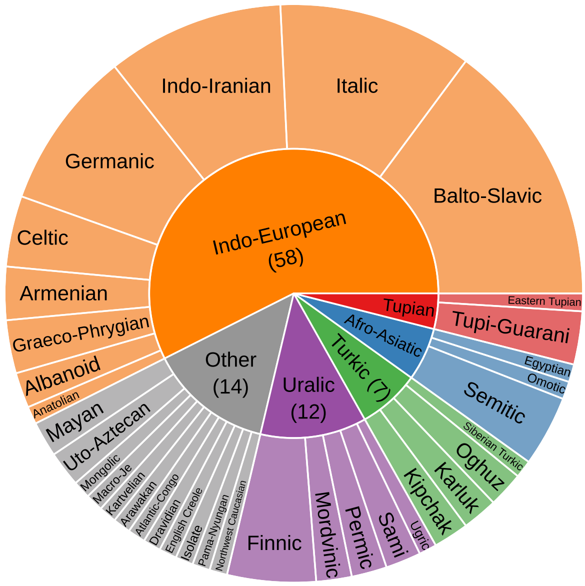


Figure 3: Distribution of language families present in MultiBLiMP 1.0. See Appendix B.1 for a detailed version of this figure, including the individual languages.

in MultiBLiMP 1.0 in Figure 3. There is a strong Indo-European (IE) bias present in our minimal pairs, caused both by the over-representation of such languages in UD, as well as our choice to focus on these types of agreement, which tend to be more present in IE languages. In future iterations of MultiBLiMP, we intend to broaden our coverage by focusing on a more diverse set of phenomena.

6 LLM Evaluation Setup

6.1 Metrics

Following prior work (Marvin and Linzen, 2018) we measure the grammaticality of the predictions of a model based on the probability it assigns to a grammatical sentence versus an ungrammatical one.[6] We choose to evaluate on the sentence-level, and not at the position of the key items, as has been done previously (Linzen et al., 2016; Jumelet and Hupkes, 2018), as this yields the same metric for both subject-verb word orders. We use two metrics in our experiments. The *accuracy score* is based on the number of items for which the model assigns a higher probability to the grammatical sentence (s^+):

$$Acc(\mathcal{M};\mathcal{D}) = \frac{1}{|\mathcal{D}|}\sum_{s\in\mathcal{D}} \mathbb{1}\left[P_{\mathcal{M}}(s^+) > P_{\mathcal{M}}(s^-)\right]$$

for model $\mathcal{M}$ and minimal pair dataset $\mathcal{D}$. We also measure the *certainty* of a model's judgment as the log probability difference for the minimal pair:

$$\Delta(\mathcal{M};\mathcal{D}) = \frac{1}{|\mathcal{D}|}\sum_{s\in\mathcal{D}} \ln P_{\mathcal{M}}(s^+) - \ln P_{\mathcal{M}}(s^-)$$

6.2 Language Models

We evaluate the following 42 LLMs, accessed through HuggingFace (Wolf et al., 2020). For all models, we evaluate both their `base` version (the model after *pre-training*), and `chat` version (the model after *post-training*).

- **Llama3** (Grattafiori et al., 2024) in sizes 1B, 3B, 8B, and 70B. We also evaluate **Tülu3** (Lambert et al., 2025), which is post-trained on the Llama3 8B base model.
- **Aya-expanse** (Dang et al., 2024), in sizes 8B and 32B, which have been post-trained on a set of 23 languages, 18 of which are in MultiBLiMP 1.0.
- **Gemma3** (Kamath et al., 2025) in sizes 4B, 12B, and 27B, which were pre-trained on a balanced multilingual distribution of 140 languages.
- **Qwen3** (Yang et al., 2025) in sizes 0.6B, 1.7B, 4B, 8B, and 14B.
- **OLMo2** (OLMo et al., 2025) in sizes 1B, 7B, 14B, and 32B.
- **EuroLLM** (Martins et al., 2024), for sizes 1.7B and 9B, which are pre-trained on 35 (mostly) European languages, 32 of which are present in MultiBLiMP 1.0.

Monolingual Models The aforementioned LLMs are all multilingual, i.e., trained on a mixture of different languages. To place their performance into perspective, we also evaluate the

[6] The best way to evaluate instruction-tuned LLMs on minimal pairs and grammaticality remains a topic of debate. 'Meta-linguistic' evaluation assesses grammaticality judgments not based on direct probability comparisons, but instead prompts the model to pick the grammatical sentence. While Dentella et al. (2023) found GPT3's abilities lacking using this approach, Hu et al. (2024) show that direct evaluation may yield better results, and Song et al. (2025) show that meta-linguistic evaluation depends strongly on model size, and outperforms direct evaluation for larger models.

			Subject-Verb			Subject-Participle			Resources			Language Subset					
Model	Size	Version	N	P	G	N	P	G	Low	Mid	High	GF	Aya	EU	Eng	All	#best
Llama3	`8B`	`base`	84.2	87.8	88.0	89.4	94.0	88.0	77.2	90.3	96.2	89.4	95.2	92.7	99.4	86.9	0
	`70B`	`base`	**87.4**	**91.2**	91.1	92.1	**97.5**	91.2	**81.1**	93.9	97.8	92.6	97.1	95.5	99.0	**90.2**	2
	`70B`	`chat`	86.5	90.3	90.3	91.7	97.2	90.6	80.3	93.1	96.9	91.9	96.2	94.6	98.3	89.3	0
Aya	`32B`	`chat`	82.9	87.8	88.1	87.4	95.1	87.0	75.7	89.4	97.7	89.0	97.3	92.8	98.4	86.4	1
Gemma3	`27B`	`base`	87.2	91.1	**91.3**	**92.8**	96.6	**91.7**	78.3	**96.3**	**98.0**	93.2	**97.4**	**97.1**	98.6	**90.2**	3
	`27B`	`chat`	82.7	87.0	86.3	88.7	95.8	86.7	73.1	92.3	94.4	88.9	93.7	93.3	96.0	85.8	0
OLMo2	`32B`	`base`	79.8	85.0	80.2	85.7	88.2	80.9	74.4	84.6	92.5	85.1	90.8	87.8	**99.5**	82.7	0
	`32B`	`chat`	78.2	83.9	79.1	84.2	87.1	80.0	72.5	83.6	91.9	84.0	90.0	86.9	99.1	81.5	0
Qwen3	`14B`	`chat`	82.2	86.4	86.4	87.8	91.5	86.6	74.1	89.9	94.8	88.2	93.8	92.2	98.3	85.3	0
EuroLLM	`9B`	`base`	82.7	86.5	87.6	89.1	71.5	89.4	72.6	92.0	95.7	88.9	94.9	96.7	99.4	85.8	0
Goldfish	`125M`		92.4	95.3	92.2	95.2	98.2	90.9	88.0	95.6	95.9	**93.8**	95.2	95.8	96.4	93.8	14

Table 3: Average accuracies per LLM, split out for different phenomena and language subsets. N stands for Number, P for Person, G for Gender. GF is the language subset of Goldfish languages; Aya the subset of Aya languages; EU the subset of EuroLLM languages; Eng the performance on English. The best performing model per category is denoted in **boldface**, but we exclude Goldfish models from this ranking as they are only evaluated on a subset of languages. #best denotes the number of languages for which this model was significantly better than the others.[7]

Goldfish suite (Chang et al., 2024b), a collection of monolingual models each trained on the same amount of data. This allows us to control for language frequency differences, which is not possible for the pre-trained LLMs. In our experiments we consider the models trained on 1GB of data, or the *full* models for languages with less than 1GB available data. All models have 500M parameters. Goldfish models exist for 70 out of the 101 languages in MultiBLiMP.

6.3 Training Corpus Language Distribution

In our experiments we connect model performance to the language distribution of the training data. Since the training corpora of most LLMs are not publicly available, we estimate this distribution based on the language frequencies of Kargaran et al. (2024), which were computed on a 3.9T token split of the Common Crawl corpus. Common Crawl provides a good reflection of the language distribution of the web-scraped data that is at the core of many LLM training corpora.

7 Experimental Results

7.1 General Performance

We report the average accuracy of the largest model for each model family in Table 3. We consider results separately for subject-verb and subject-participle agreement, and then split further by feature between Number, Person, and Gender. We also report results for language subgroups split based on the Common Crawl language frequencies: low-resource for the least frequent 60% languages, mid-resource for the languages in the 60–90% frequency range, and high-resource for the most frequent 10% languages. Finally, we report results for the language subsets of the Goldfish, Aya, and EuroLLM models; performance on English and the overall average. We also denote the number of languages for which a model was significantly better than all other models.[7]

Llama3-70B-base and Gemma3-27B-base perform best overall, both scoring 90.2% on average. These two models also score best on most of the categories, with the Llama model scoring better on low-resource languages, and Gemma better on mid- and high-resource languages. For all models, performance increases with language frequency, demonstrating much of the variance in linguistic ability can be explained by data disparity. Gemma3-27B-base outperforms the Aya and EuroLLM models on their specific language subsets, validating the claim of Kamath et al. (2025) that its better balance of languages leads

[7]We conduct a McNemar test between the best and second-best model based on the item-level binary grammaticality judgments, with $p < 0.1$ as significance threshold.

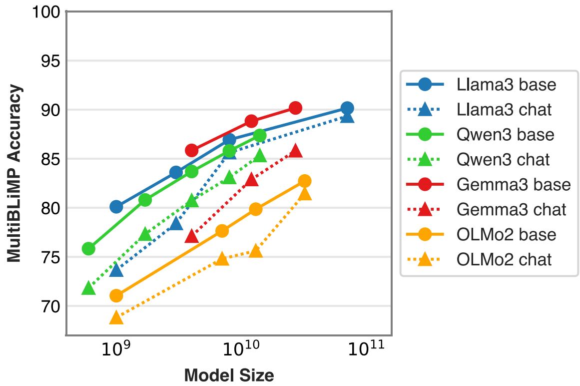


Figure 4: The impact of model size (in number of parameters on a log scale) against overall MultiBLiMP accuracy for the Llama3, Qwen3, Gemma3, and OLMo2 model families.

to strong multilingual performance. Interestingly, both these models are still being outperformed by the Goldfish model series. The Goldfish models are significantly better on 14 out of 101 languages, whereas Llama3-70B and Gemma3-27B are best for 2 and 3 languages, respectively. We present the best model per language in Appendix B.2, and a breakdown of per-language performance in Appendix B.3.

7.2 Impact of Model Properties

We investigate how model properties drive performance, focusing on **model size** and **post-training**. For this analysis we make use of the full set of all 42 LLMs described in §6.2. We fitted a linear regression model to predict the average accuracy of the LLMs, using model size as a continuous variable, model family as a categorical fixed effect, and a binary indicator for post-training status (post-trained vs. base).

The fitted model yields an R^2 of 0.936. Model size has a significantly positive impact on performance ($\beta = 4.30$; $t = 15.88$). Post-training, however, has significantly *negative* impact on performance ($\beta = -3.29$; $t = -6.47$). This finding is in line with Chirkova and Nikoulina (2024), who report that instruction tuning (a stage of post-training) can have a negative effect on multilingual fluency in LLMs.[8] We illustrate this in Figure 4 for the four largest model families: Performance increases with size, and post-trained models consistently underperform their base version.

[8]We experimented with different *chat templates* for the `chat` version of the Gemma3 models, but did not observe any improvement. We leave more elaborate 'prompt engineering' open for future work.

7.3 Cross-Model Comparisons

While the previous experiment demonstrates that model size and post-training have an impact on performance, it remains unclear how individual languages are impacted by this. To investigate this, we plot the language-level performance of various models against each other in Figure 5.

We consider the impact of **language-specific post-** and **pre-training** in Figure 5a) and 5b), comparing the `chat` version of Llama3-8B to Aya-8B and the `base` version of Llama3-8B to EuroLLM-9B. We have highlighted the specific languages on which Aya has been post-trained, and EuroLLM has been pre-trained. Languages outside this selection do not appear to benefit from cross-lingual transfer, as both Aya and EuroLLM perform worse than Llama3-8B in the languages they were *not* trained on, and McNemar testing shows Llama3-8B is significantly better than Aya for 43 languages, and better than EuroLLM for 31 languages. This suggests that language-specific *pre-training* has a stronger effect on linguistic ability than post-training: EuroLLM significantly outperforms Llama3-8B on 19 of its 32 languages (59%), whereas Aya only outperforms it on 8 out of its 18 languages (44%). Further research in a setting with better control over the training data is needed to identify more precisely the data properties that drive multilingual grammar acquisition.

Next, we consider the impact of **model size** in Figure 5c, by comparing the 8B model to the 70B model. The average performance goes up from 86.9% to 90.2% for this model, driven by gradual improvements: no language improves by more than 10%, but almost all languages improve to some degree. We also conduct a McNemar test between the two models, with $p < 0.05$, resulting in Llama3-70B being significantly better for 48 languages, and none the other way around.

Finally, we look at the impact of **monolingual** vs. multilingual language modelling in Figure 5d. Goldfish models significantly outperform Llama3-8B for 39 out of 70 languages, whereas Llama performs better only for 7 languages. This warrants a more detailed investigation in future work, to assess if this difference is solely driven by differences in data frequency and data quality, or by the multilingual LLM

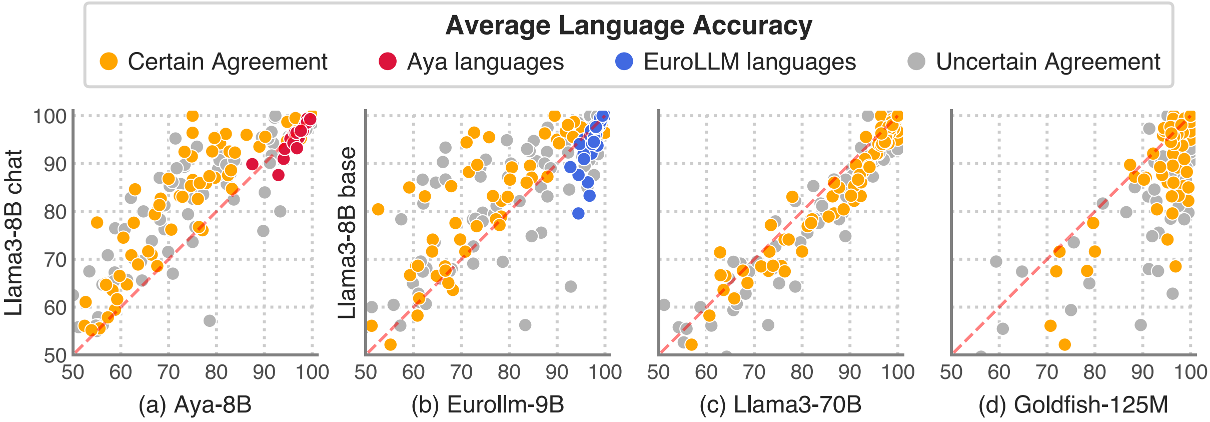


Figure 5: Average accuracy score per language, plotted for Llama3-8B `chat` (a) and `base` (b–d) against various other models. The Aya and EuroLLM languages are a subset of the Certain Agreement class, and are highlighted for visualization purposes.

having to handle all these languages in a shared representation space (Chang et al., 2024a).

7.4 What Factors Drive Performance?

We set up a linear modelling experiment in which we predict the Δ scores (§6.1) of a few models based on various factors that we hypothesize to be driving model judgments. We include the following fixed effects in our regression model: the six **agreement types** as categorical variables; **subject-verb distance**; **perplexity** of the grammatical sentence; **sentence length**; **subword delta** expressing the difference in the number of subwords for the grammatical and ungrammatical key item; two binary features denoting if the intervening material between subject and verb contained **congruent** or **incongruent attractors**; and the **language frequency** of §6.3. To interpret this model, we plot the β coefficients for the standardized factors, fitted for the two best-performing LLMs (Llama3-70B and Gemma3-27B), and the Goldfish models.

Results As shown in Figure 6, the agreement type coefficients show that person agreement yields high Δ scores (i.e., more confidence in detecting the grammatical sentence). This can be due to the fact that first and second person subjects tend to be closed-class pronouns, whereas gender and number agreement requires acquisition of open-class subject features. Surprisingly, subject-verb distance does *not* have a negative impact on Δ: One might expect agreement to be more challenging for long-distance dependencies. An increase in sentence perplexity strongly *decreases* Δ: The more surprised a model is by a sentence, the less likely it is to make the right grammaticality judgment.

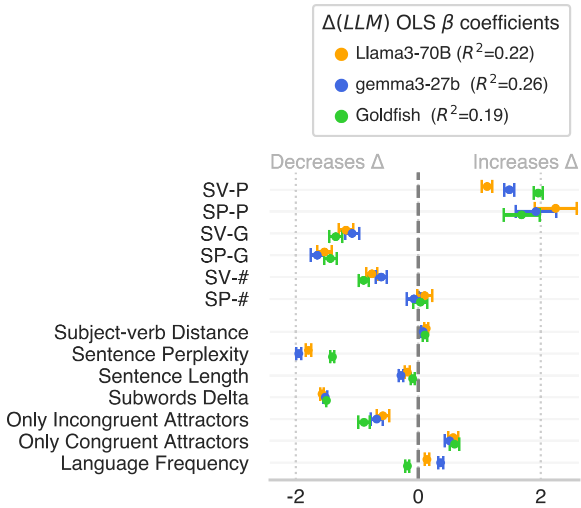


Figure 6: The β coefficient estimates of OLS models fitted on Δ scores of Llama3-70B, Gemma3-27B, and Goldfish.

Another important factor with a negative effect is the *subword delta*: If the correct form of the verb is split into *more* subwords than the incorrect form, the model is more likely to make a wrong judgment. As we expected, incongruent attractors result in a decrease in Δ, while congruent attractors boost it. Finally, language frequency has a positive effect on the performance of the LLMs, whereas for Goldfish (which controls for frequency) this is not the case. This is the only factor for which the Goldfish coefficient is opposite to the LLMs, for all other factors it is

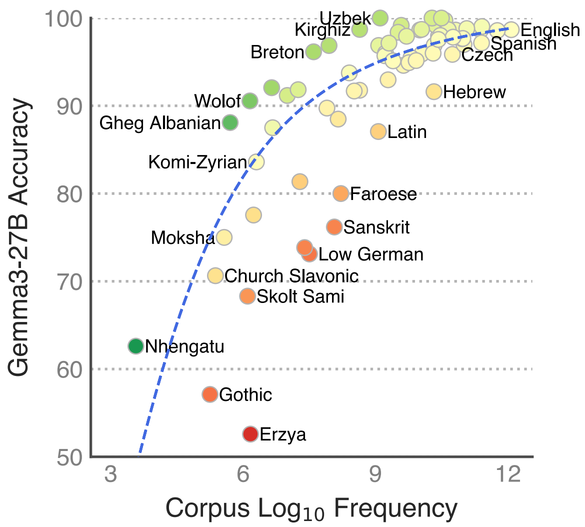


Figure 7: Gemma3-27B accuracy per language on MultiBLiMP 1.0, plotted against language frequency in Common Crawl. Accuracy is measured based on the model assigning a higher probability to a grammatical sentence over a minimally different but ungrammatical sentence. Languages are coloured by their positive or negative deviation from the general trend of accuracy increasing with corpus frequency, highlighting languages that over- or underperform relative to the amount of resources available for them.

similar. A more detailed breakdown of impact of language frequency on Gemma3-27B is shown in Figure 7, where we highlight the languages that over- and underperform with respect to a fitted logistic curve. An interesting question for future work would be to investigate the extent to which the variance in performance is driven by typological features.

8 Discussion and Conclusion

Takeaways for Model Training Based on the results of evaluating 42 LMs on MultiBLiMP 1.0 and analyzing the results, we formulate several recommendations for the training of future massively multilingual models. Based on the comparison between Aya and EuroLLM in Figure 5, we surmise that to boost specific focus languages, fine-tuning is not as effective as pre-training. We observed that the Goldfish models significantly outperform LLMs as large as 70B parameters on 14 out of 101 languages, many of them low-resource. This suggests that formal linguistic competence can greatly suffer when the training data of a particular language only forms a tiny part of the general mix—especially if that language is not closely related to the bulk of this mix. Thus, communities of languages underserved by current NLP technology may be best helped not by integrating their languages into LLM projects, but by more targeted regional LM initiatives. Our finding that the accuracy on MultiBLiMP 1.0 depended strongly on the difference in number of subwords between the correct and the incorrect inflection is in keeping with previous work (Rust et al., 2021, i.a.) which suggested that performance differences between languages in multilingual LLMs are strongly driven by the space allocated to the language in the tokenizer. Taken together, these results call for new tokenization strategies to mitigate this issue (Mielke et al., 2021).

The Potential of Multilingual Annotated Resources in the Age of LLMs The creation of MultiBLiMP 1.0 was only possible due to the existence of UD and UniMorph, both large resources created by a large number of annotators over many years. While these resources have become increasingly marginalised since the introduction of LLMs, we see this work as a prime example of their continued usefulness. The wealth of linguistic knowledge that is captured by these resources will continue to prove invaluable for informed linguistic evaluation of LLMs (Opitz et al., 2024).

Future Work We see two major opportunities for future work. First, we plan to expand MultiBLiMP 1.0 to more constructions beyond subject-verb agreement, investigating more diverse phenomena not attested in English and broadening the current language set. Second, this benchmark presents a unique opportunity for computational typology, due to the diverse set of languages included and the number of phenomena covered. We hope that MultiBLiMP will enable learnability studies across typologically diverse languages, bringing new insights into their linguistic structure, but also into the work that is needed to put them onto equal footing in language modelling.

Limitations

Our minimal pair creation procedure is limited by the size, diversity, and annotation quality of the resources which we rely on, namely, UD and UniMorph. Extending our procedure to more complex phenomena may also be challenging, since it requires one to define grammaticality violations in terms of morphological inflections.

We are further limited by the sheer number of languages in our benchmark, which makes manual evaluation cost-prohibitive, except for spot checks for a subset of languages. We see MultiBLiMP as a continuous effort, for which we have already taken feedback from language experts in our network and adjusted individual languages accordingly.

Acknowledgments

We would like to thank Amir Kargaran for providing the language frequencies in CommonCrawl. We further thank Ahmet Üstün, Charlotte Pouw, Jirui Qi, Omer Goldman, Kanishka Misra, and Will Merrill, as well as the GroNLP and Edinburgh ILCC groups for helpful feedback. We thank the TACL reviewers for their valuable feedback. Jaap Jumelet and Arianna Bisazza were supported by NWO grant `VI.Vidi.221C.009`. Leonie Weissweiler was supported by a postdoctoral fellowship from the German Research Foundation (DFG, `WE 7627/1-1`). This work used the Dutch national e-infrastructure with the support of the SURF Cooperative using grant no. `EINF-13403`.

References

M. C. Baker. 2008. *The Syntax of Agreement and Concord*. Cambridge Studies in Linguistics. Cambridge University Press. https://doi.org/10.1017/CBO9780511619830

Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, Maria Ryskina, Sabrina Mielke, Elena Budianskaya, Charbel El-Khaissi, Tiago Pimentel, Michael Gasser, William Abbott Lane, Mohit Raj, Matt Coler, Jaime Rafael Montoya Samame, Delio Siticonatzi Camaiteri, Esaú Zumaeta Rojas, Didier López Francis, Arturo Oncevay, Juan López Bautista, Gema Celeste Silva Villegas, Lucas Torroba Hennigen, Adam Ek, David Guriel, Peter Dirix, Jean-Philippe Bernardy, Andrey Scherbakov, Aziyana Bayyr-ool, Antonios Anastasopoulos, Roberto Zariquiey, Karina Sheifer, Sofya Ganieva, Hilaria Cruz, Ritván Karahóga, Stella Markantonatou, George Pavlidis, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Candy Angulo, Jatayu Baxi, Andrew Krizhanovsky, Natalia Krizhanovskaya, Elizabeth Salesky, Clara Vania, Sardana Ivanova, Jennifer White, Rowan Hall Maudslay, Josef Valvoda, Ran Zmigrod, Paula Czarnowska, Irene Nikkarinen, Aelita Salchak, Brijesh Bhatt, Christopher Straughn, Zoey Liu, Jonathan North Washington, Yuval Pinter, Duygu Ataman, Marcin Wolinski, Totok Suhardijanto, Anna Yablonskaya, Niklas Stoehr, Hossep Dolatian, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Aryaman Arora, Richard J. Hatcher, Ritesh Kumar, Jeremiah Young, Daria Rodionova, Anastasia Yemelina, Taras Andrushko, Igor Marchenko, Polina Mashkovtseva, Alexandra Serova, Emily Prud'hommeaux, Maria Nepomniashchaya, Fausto Giunchiglia, Eleanor Chodroff, Mans Hulden, Miikka Silfverberg, Arya D. McCarthy, David Yarowsky, Ryan Cotterell, Reut Tsarfaty, and Ekaterina Vylomova. 2022. UniMorph 4.0: Universal Morphology. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 840–855, Marseille, France. European Language Resources Association.

Emi Baylor, Esther Ploeger, and Johannes Bjerva. 2023. The past, present, and future of typological databases in NLP. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1163–1169, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.82

Emi Baylor, Esther Ploeger, and Johannes Bjerva. 2024. Multilingual gradient word-order typology from Universal Dependencies. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 42–49, St. Julian's, Malta. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.eacl-short.6

Jannik Brinkmann, Chris Wendler, Christian Bartelt, and Aaron Mueller. 2025. Large language models share representations of latent grammatical concepts across typologically diverse languages. In *Proceedings of the 2025 Conference of the Nations of the Americas*

Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 6131–6150, Albuquerque, New Mexico. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.naacl-long.312

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Ben Bergen. 2024a. When is multilinguality a curse? Language modeling for 250 high- and low-resource languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4074–4096, Miami, Florida, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.236

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K. Bergen. 2024b. Goldfish: Monolingual language models for 350 languages.

Aditi Chaudhary, Antonios Anastasopoulos, Adithya Pratapa, David R. Mortensen, Zaid Sheikh, Yulia Tsvetkov, and Graham Neubig. 2020. Automatic extraction of rules governing morphological agreement. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5212–5236, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.422

Ethan A. Chi, John Hewitt, and Christopher D. Manning. 2020. Finding universal grammatical relations in multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.493

Nadezhda Chirkova and Vassilina Nikoulina. 2024. Zero-shot cross-lingual transfer in instruction tuning of large language models. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 695–708, Tokyo, Japan. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.inlg-main.53

Rochelle Choenni and Ekaterina Shutova. 2022. Investigating language relationships in multilingual sentence encoders through the lens of linguistic typology. *Computational Linguistics*, 48(3):635–672. https://doi.org/10.1162/coli_a_00444

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan N. Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *CoRR*, abs/2412.04261.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308. https://doi.org/10.1162/coli_a_00402

Vittoria Dentella, Fritz Günther, and Evelina Leivada. 2023. Systematic testing of three language models reveals low language accuracy, absence of response stability, and a yes-response bias. *Proceedings of the National Academy of Sciences*, 120(51):e2309583120. https://doi.org/10.1073/pnas.2309583120, PubMed: 38091290

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.

Prajit Dhar and Arianna Bisazza. 2021. Understanding cross-lingual syntactic transfer in multilingual recurrent neural networks. In

Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa), pages 74–85, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online (v2020.4)*. Zenodo.

Javier Ferrando and Marta R. Costa-jussà. 2024. On the similarity of circuits across languages: A case study on the subject-verb agreement task. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10115–10125, Miami, Florida, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.591

Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 2020. SyntaxGym: An online platform for targeted evaluation of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-demos.10

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, et al. 2024. The LlaMa 3 herd of models.

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. https://doi.org/10.18653/v1/N18-1108

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. https://doi.org/10.18653/v1/2020.acl-main.158

Jennifer Hu, Kyle Mahowald, Gary Lupyan, Anna Ivanova, and Roger Levy. 2024. Language models align with human judgments on key grammatical constructions. *Proceedings of the National Academy of Sciences*, 121(36):e2400917121. https://doi.org/10.1073/pnas.2400917121, PubMed: 39186652

Jaap Jumelet and Dieuwke Hupkes. 2018. Do language models understand anything? On the ability of LSTMs to understand negative polarity items. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 222–231, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-5424

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucinska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, et al. 2025. Gemma 3 technical report.

Amir Hossein Kargaran, François Yvon, and Hinrich Schuetze. 2024. GlotCC: An open broad-coverage commoncrawl corpus and pipeline for minority languages. In *Advances in Neural Information Processing Systems*, volume 37, pages 16983–17005. Curran Associates, Inc. https://doi.org/10.52202/079017-0540

Daria Kryvosheieva and Roger Levy. 2025. Controlled evaluation of syntactic knowledge in multilingual language models. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 402–413, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. 2025. Tulu 3: Pushing frontiers in open language model post-training.

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. *CoRR*, abs/2309.06085.

Natalia Levshina, Savithry Namboodiripad, Marc Allassonnière-Tang, Mathew Kramer, Luigi Talamo, Annemarie Verkerk, Sasha Wilmoth, Gabriela Garrido Rodriguez, Timothy Michael Gupton, Evan Kidd, Zoey Liu, Chiara Naccarato, Rachel Nordlinger, Anastasia Panova, and Natalia Stoynova. 2023. Why we need a gradient approach to word order. *Linguistics*, 61(4):825–883. https://doi.org/10.1515/ling-2021-0098

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. https://doi.org/10.1162/tacl_a_00115

Yikang Liu, Yeting Shen, Hongao Zhu, Lilong Xu, Zhiheng Qian, Siyuan Song, Kejia Zhang, Jialong Tang, Pei Zhang, Baosong Yang, Rui Wang, and Hai Hu. 2024. ZhoBLiMP: A systematic assessment of language models with linguistic minimal pairs in chinese. *CoRR*, abs/2411.06096.

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28(6):517–540. https://doi.org/10.1016/j.tics.2024.01.011, PubMed: 38508911

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2024. EuroLLM: Multilingual language models for europe. https://doi.org/10.1016/j.procs.2025.02.260

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. https://doi.org/10.18653/v1/D18-1151

Sabrina J. Mielke, Zaid Alyafeai, Elizabeth Salesky, Colin Raffel, Manan Dey, Matthias Gallé, Arun Raja, Chenglei Si, Wilson Y. Lee, Benoît Sagot, and Samson Tan. 2021. Between words and characters: A brief history of open-vocabulary modeling and tokenization in NLP. *CoRR*, abs/2112.10508.

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. https://doi.org/10.18653/v1/2020.acl-main.490

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, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

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.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Allyson Ettinger, Michal Guerquin, David Heineman, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Jake Poznanski, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2025. 2 olmo 2 furious.

Juri Opitz, Shira Wein, and Nathan Schneider. 2024. Natural language processing RELIES on linguistics. *CoRR*, abs/2405.05966.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1493

Adithya Pratapa, Antonios Anastasopoulos, Shruti Rijhwani, Aditi Chaudhary, David R. Mortensen, Graham Neubig, and Yulia Tsvetkov. 2021. Evaluating the morphosyntactic well-formedness of generated texts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7131–7150, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.570

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? On the monolingual performance of multilingual language models.

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 3118–3135, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.243

Benoît Sagot. 2018. A multilingual collection of CoNLL-U-compatible morphological lexicons. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermis, and Sara Hooker. 2025. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.919

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Chien, Sebastian Ruder, Surya Guthikonda, Emad Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567, Bangkok, Thailand. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.620

Hedvig Skirgård, Hannah J. Haynie, Damián E. Blasi, Harald Hammarström, Jeremy Collins, Jay J. Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Sam Passmore, Angela Chira, Luke Maurits, Russell Dinnage, Michael Dunn, Ger Reesink, Ruth Singer, Claire Bowern, Patience Epps, Jane Hill, Outi Vesakoski, Martine Robbeets, Noor Karolin Abbas, Daniel Auer, Nancy A. Bakker, Giulia Barbos, Robert D. Borges, Swintha Danielsen, Luise Dorenbusch, Ella Dorn, John Elliott, Giada Falcone, Jana Fischer, Yustinus Ghanggo Ate, Hannah Gibson, Hans-Philipp Göbel, Jemima A. Goodall, Victoria Gruner, Andrew Harvey, Rebekah Hayes, Leonard Heer, Roberto E. Herrera Miranda, Nataliia Hübler, Biu Huntington-Rainey, Jessica K. Ivani, Marilen Johns, Erika Just, Eri Kashima, Carolina Kipf, Janina V. Klingenberg, Nikita König, Aikaterina Koti, Richard G. A. Kowalik, Olga Krasnoukhova, Nora L. M. Lindvall, Mandy Lorenzen, Hannah Lutzenberger, Tônia R. A. Martins, Celia Mata German, Suzanne van der Meer, Jaime Montoya Samamé, Michael Müller, Saliha Muradoğlu, Kelsey Neely, Johanna Nickel, Miina Norvik, Cheryl Akinyi Oluoch, Jesse Peacock, India O. C. Pearey, Naomi Peck, Stephanie Petit, Sören Pieper, Mariana Poblete, Daniel Prestipino, Linda Raabe, Amna Raja, Janis Reimringer, Sydney C. Rey, Julia Rizaew, Eloisa Ruppert, Kim K. Salmon, Jill Sammet, Rhiannon Schembri, Lars Schlabbach, Frederick W. P. Schmidt, Amalia Skilton, Wikaliler Daniel Smith, Hilário de Sousa, Kristin Sverredal, Daniel Valle, Javier Vera, Judith Voß, Tim Witte, Henry Wu, Stephanie Yam, Jingting Ye, Maisie Yong, Tessa Yuditha, Roberto Zariquiey, Robert Forkel, Nicholas Evans, Stephen C. Levinson, Martin Haspelmath, Simon J. Greenhill, Quentin D. Atkinson, and Russell D. Gray. 2023. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances*, 9. https://doi.org/10.1126/sciadv.adg6175, PubMed: 37075104

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. https://doi.org/10.18653/v1/2023.findings-eacl.117

Siyuan Song, Jennifer Hu, and Kyle Mahowald. 2025. Language models fail to introspect about their knowledge of language. *CoRR*, abs/2503.07513.

Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, and Mohit Iyyer. 2022. SLING: Sino linguistic evaluation of large language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4606–4634, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.305

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. https://doi.org/10.18653/v1/2022.naacl-main.114

Michelle Suijkerbuijk, Zoë Prins, Marianne de Heer Kloots, Willem Zuidema, and Stefan L. Frank. 2025. BLiMP-NL: A corpus of dutch minimal pairs and acceptability judgments for language model evaluation. *Computational Linguistics*, pages 1–35. https://doi.org/10.1162/coli_a_00559

John Sylak-Glassman, Christo Kirov, David Yarowsky, and Roger Que. 2015. A language-independent feature schema for inflectional morphology. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 674–680, Beijing, China. Association for Computational Linguistics. https://doi.org/10.3115/v1/P15-2111

Ekaterina Taktasheva, Maxim Bazhukov, Kirill Koncha, Alena Fenogenova, Ekaterina Artemova, and Vladislav Mikhailov. 2024. RuBLiMP: Russian benchmark of linguistic minimal pairs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9268–9299, Miami, Florida, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.522

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.845

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. https://doi.org/10.1162/tacl_a_00321

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641. https://doi.org/10.1162/tacl_a_00290

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do Llamas work in English? On the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.820

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony

Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.6

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. https://doi.org/10.1175/JCLI-D-20-0301.1

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Arofat Akhundjanova, Furkan Akkurt, Gabrielė Aleksandravičiūtė, Ika Alfina, Avner Algom, Khalid Alnajjar, Chiara Alzetta, Erik Andersen, Matthew Andrews, et al. 2024. Universal dependencies 2.15. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

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. https://doi.org/10.18653/v1/2024.acl-long.146

A Additional Details

A.1 Excluded Treebanks and Criteria

We exclude a number of treebanks for various reasons. For low-resource languages, treebanks are only excluded where absolutely necessary, for high-resource languages, treebanks may be excluded to make the resulting genres and annotation schemata more uniform. An overview of the treebanks we removed is provided in Table 4. We remove diacritics from Latin, Slovenian, and Western Farsi, and cantillations from Biblical Hebrew to ensure consistency between their UD and UM data sources. We transliterate Uyghur from Arabic to Latin, Sanskrit from Latin to Devanagari, and Tatar from Latin to Cyrillic using the official guidelines `https://suzlek.antat.ru/about/TAAT2019/8.pdf`. We skip all UD languages that do not have an ISO-639-3 code: code-switched Frisian Dutch, Turkish German, and Telugu English, as well as Cappadocian, Maghrebi Arabic French, Pomak.

A.2 Minimal Pair Sample

Figure 8.

Language	**Corpus**	**Characteristics**
Vietnamese	TueCL	Spoken
Hebrew	IAHLTknesset	Spoken
Hebrew	IAHLT	Incompatible with HTB
Latvian	Cairo	Only 20 sentences
German	LIT	Historical
German	HDT	Annotation quality
Czech	Poetry	Genre
Russian	Poetry	Genre
Slovenian	SST	Spoken
Faroese	FarPaHC	Historical
Icelandic	IcePaHC	Historical
English	GUM	Spoken (among others)
English	Pronouns	Niche
English	ESLSPok	Spoken
English	CTeTex	Technical genre
Sanskrit	UFAL	Comparatively small
French	ParisStories	Spoken
French	Rhapsodie	Spoken
Galician	CTG	Low star rating
Italian	Old	Historical
Italian	Valico	Learner corpus
Romanian	Nonstandard	Genre
Romanian	ArT	Dialect
Spanish	COSER	Spoken
Class. Chinese	TueCL	Comparatively small

Table 4: Excluded treebanks with reasons.

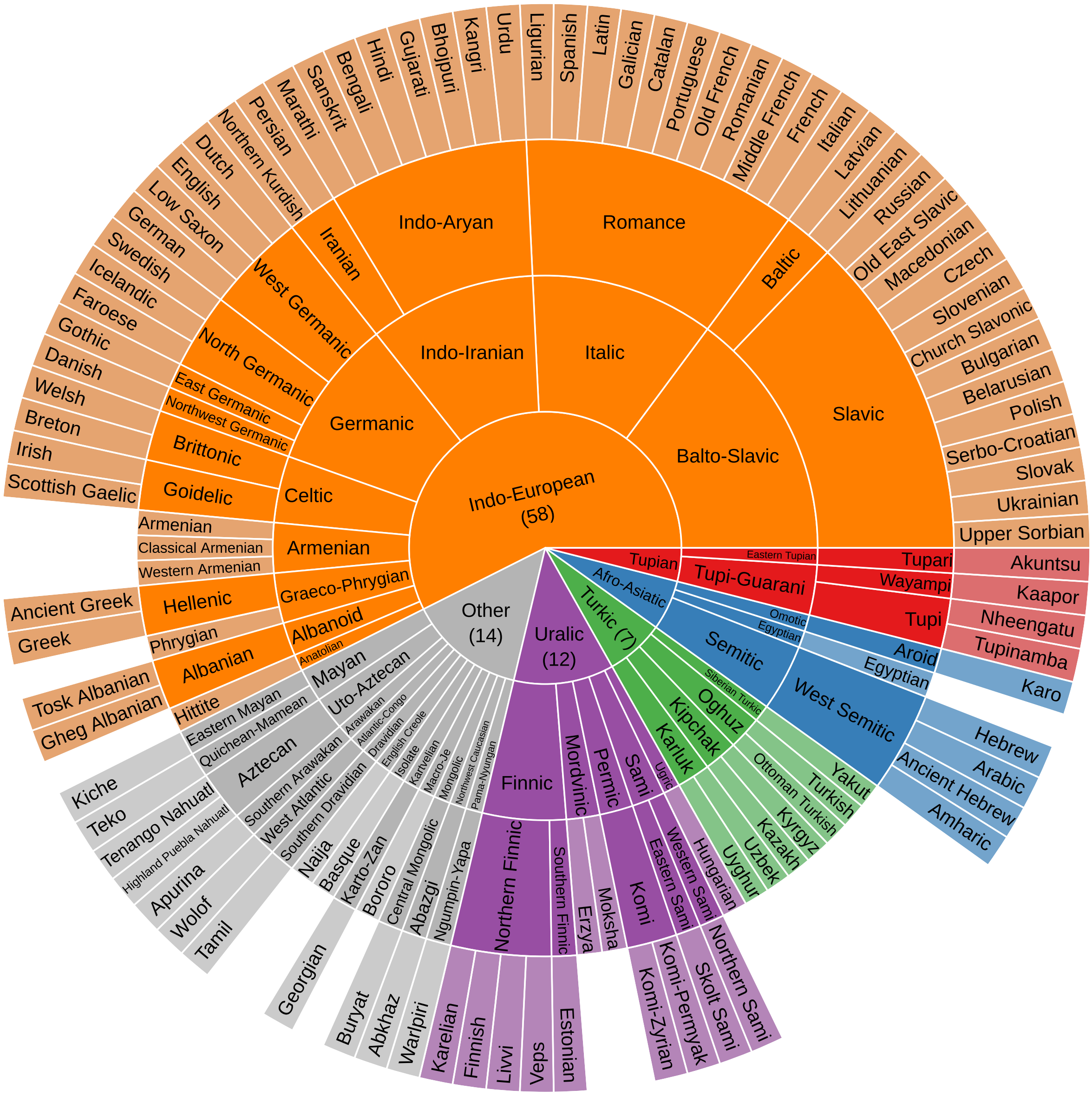
Nhengatu	SV-P	Awá kurí ti uruyari, [**tauyuká**/***peyuká**] kurí arupí aé.
Gheg Albanian	SV-P	dhe ata e [**pan**/***pam**] se ky isht ërrxue .
Wolof	SV-N	Njiitam Séydu Nuura Njaay woyof [**na**/***nañu**]!
Low German	SV-N	De jungens [**lachen**/***lach**] luudhals: »Ney, dat büst du!«
Faroese	SV-N	Í 2008 [**var**/***vóru**] ASFALT tó avlýst.
Latin	SV-G	sese propediem cum magno exercitu ad urbem [**accessurum**/***accessuram**].
Breton	SV-N	Ne [**lennan**/***lennomp**]-me ket al lizher.
Kirghiz	SV-P	Балдар ыйлакташып, кечирим сурашып, экинчи кайталабайбыз деп убада [**беришет**/***бердик**].
Hebrew	SP-N	"מי ש יוצא הרבה מביא חזרה לכלוך רב", [**אומרת**/***אומרות**] מימרה טורקית.
Spanish	SP-N	Ninguno de los dos escritores ha [**colaborado**/***colaborados**] en los guiones.
Moksha	SV-N	Весть очижить карта попсь алашаса кудрядс [**ётась**/***ётазь**].
Skolt Sami	SV-P	Son [**vuâlgg**/***vuâlgam**], tõid neävveez kuâđđ.
French	SP-G	L'argent qui devait financer le film n'est jamais [**arrivé**/***arrivée**].

Figure 8: Sample of sentences from MultiBLiMP 1.0. The full dataset can be found at `huggingface.co/datasets/jumelet/multiblimp`.

B Complementary Results

B.1 Language Distribution

Distribution of the number of language families and languages present in MultiBLiMP 1.0:



B.2 Significantly Best Model per Language

Model	Language	Accuracy
Goldfish	Albanian	99.2 ($p = 0.041$)
	Buriat	91.3 ($p = 0.002$)
	Catalan	97.7 ($p = 0.004$)
	Erzya	73.7 ($p = 0.000$)
	Faroese	99.6 ($p = 0.006$)
	Galician	98.3 ($p = 0.001$)
	Icelandic	98.4 ($p = 0.000$)
	Ligurian	94.5 ($p = 0.000$)
	Marathi	95.2 ($p = 0.001$)
	Northern Kurdish	94.7 ($p = 0.033$)
	Northern Sami	96.9 ($p = 0.000$)
	Welsh	99.4 ($p = 0.000$)
	Wolof	97.0 ($p = 0.000$)
	Yakut	95.8 ($p = 0.029$)
Llama3-70B	Church Slavonic	68.1 ($p = 0.009$)
	Gothic	75.4 ($p = 0.000$)
	Komi-Zyrian	77.2 ($p = 0.070$)
	Latin	92.1 ($p = 0.026$)
	Old French	81.8 ($p = 0.001$)
	Old Russian	80.0 ($p = 0.024$)
	Sanskrit	81.4 ($p = 0.000$)
aya-32b	French	99.5 ($p = 0.063$)
gemma3-27b	Bulgarian	99.2 ($p = 0.000$)
	Gheg Albanian	80.4 ($p = 0.050$)
	Polish	98.8 ($p = 0.002$)

B.3 Language-specific Results

Table 5.

ISO	Language	n	Llama3			Aya	Gemma3		OLMo2		Qwen3	EuroLLM	Goldfish
			8B	70B	70B-it	32B-it	27B	27B-it	32B	32B-it	14B-it	9B	125M
abk	Abkhazian	40	65.0	70.0	60.0	67.5	**77.5**	70.0	70.0	67.5	55.0	47.5	**77.5**
aqz	Akuntsu	14	35.7	42.9	35.7	28.6	35.7	14.3	35.7	28.6	35.7	**57.1**	—
sqi	Albanian	243	86.0	88.5	90.9	83.1	96.7	92.2	80.7	81.9	85.2	61.3	**99.2**
amh	Amharic	112	**100.0**	98.2	97.3	99.1	96.4	91.1	96.4	93.8	94.6	94.6	96.4
grc	Ancient Greek	3695	86.8	90.8	**91.3**	87.7	87.7	79.9	91.2	90.3	84.7	87.7	88.1
hbo	Ancient Hebrew	983	86.9	91.8	90.2	90.9	**92.9**	83.1	90.0	85.7	83.7	80.7	—
apu	Apurinã	28	**96.4**	**96.4**	**96.4**	92.9	85.7	92.9	**96.4**	89.3	**96.4**	**96.4**	—
hye	Armenian	1415	96.5	**98.4**	97.3	93.7	97.9	93.9	86.9	84.9	94.8	72.7	**98.4**
eus	Basque	273	94.1	95.2	96.3	91.2	97.1	89.4	90.1	90.5	95.6	91.6	**98.9**
bel	Belarusian	2570	89.2	93.1	91.6	84.6	96.9	92.3	74.7	74.4	85.2	80.5	**97.3**
ben	Bengali	21	90.5	95.2	95.2	95.2	95.2	85.7	81.0	85.7	95.2	85.7	**100.0**
bho	Bhojpuri	34	85.3	82.4	85.3	82.4	82.4	79.4	55.9	67.6	73.5	67.6	**88.2**
bor	Borôro	241	66.0	64.7	64.3	**67.6**	61.0	61.8	63.1	60.2	63.1	65.6	—
bre	Breton	260	94.6	95.0	96.2	94.2	97.3	93.5	75.0	64.6	80.0	86.2	**99.2**
bul	Bulgarian	2458	93.6	96.0	95.0	85.5	**99.2**	95.8	89.7	87.8	94.8	97.6	97.2
bua	Buriat	103	68.0	73.8	73.8	69.9	77.7	70.9	67.0	66.0	67.0	68.0	**91.3**
cat	Catalan	2284	94.8	96.1	95.8	95.1	96.4	92.6	90.9	90.2	93.9	95.5	**97.7**
chu	Church Slavonic	4166	63.7	**68.1**	67.5	64.2	66.3	62.3	64.0	61.1	62.3	63.0	—
xcl	Classical Armenian	1623	70.0	75.4	74.7	69.6	**76.9**	70.5	66.4	65.1	72.0	64.1	—
ces	Czech	4256	92.1	95.7	94.8	97.0	96.2	91.5	83.7	82.8	90.4	**97.2**	92.2
dan	Danish	50	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	90.0	88.0	**100.0**	**100.0**	**100.0**
nld	Dutch	2331	96.1	97.7	95.6	98.0	97.6	89.8	90.8	90.5	91.0	**98.2**	97.3
egy	Egyptian (Ancient)	22	**50.0**	45.5	**50.0**	**50.0**	**50.0**	45.5	45.5	40.9	**50.0**	40.9	—
eng	English	770	99.4	99.0	98.3	98.4	98.6	96.0	**99.5**	99.1	98.3	99.4	96.4
myv	Erzya	464	52.2	56.9	53.4	54.1	52.6	46.3	57.1	55.8	54.1	55.2	**73.7**
est	Estonian	2575	79.6	86.7	86.0	74.9	95.4	89.3	71.3	70.6	79.5	94.4	**96.2**
fao	Faroese	232	79.7	88.4	85.8	75.9	94.8	89.7	83.2	78.4	75.0	74.6	**99.6**
fin	Finnish	2570	91.4	94.5	93.8	86.0	**96.4**	93.9	91.5	91.4	87.1	95.2	96.2
fra	French	2548	98.7	99.0	98.0	**99.5**	99.1	96.4	98.6	97.8	97.5	99.1	98.5
glg	Galician	753	89.9	92.3	91.4	89.6	94.7	88.8	84.3	83.9	88.7	95.2	**98.3**
kat	Georgian	204	97.5	96.6	94.6	95.6	94.6	94.1	87.3	87.3	**98.0**	95.1	96.6
deu	German	2298	98.0	**99.0**	97.5	98.9	98.8	94.1	97.3	97.0	96.0	98.8	98.0
aln	Gheg Albanian	677	72.8	76.4	75.5	76.8	**80.4**	74.0	73.1	70.3	69.3	66.9	—
got	Gothic	1579	68.3	**75.4**	69.2	63.4	55.9	56.0	63.7	62.5	58.6	60.4	—
guj	Gujarati	7	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
heb	Hebrew	2330	88.1	90.0	90.4	**90.9**	90.7	84.1	80.1	78.1	81.9	71.2	85.9
azz	H-P Nahuatl	207	69.1	67.1	67.6	70.5	65.7	66.2	65.2	67.1	68.6	**72.5**	—
hin	Hindi	1447	99.3	99.3	99.2	**99.6**	99.2	98.5	98.3	98.3	98.4	99.0	98.7
hit	Hittite	50	62.0	**66.0**	58.0	52.0	46.0	46.0	58.0	62.0	58.0	36.0	—
hun	Hungarian	845	97.3	**99.2**	98.6	90.4	99.1	94.7	88.0	86.4	94.6	98.3	97.5
isl	Icelandic	2801	85.3	91.0	90.7	81.7	96.0	90.6	79.7	79.3	81.9	67.8	**98.4**
gle	Irish	28	82.1	78.6	89.3	82.1	64.3	57.1	82.1	85.7	78.6	82.1	**92.9**
ita	Italian	2999	96.2	**97.7**	97.0	97.3	97.0	92.5	91.8	90.7	95.4	97.1	94.6
quc	K'iche'	131	69.5	65.6	70.2	71.0	61.8	62.6	65.6	67.2	67.9	**78.6**	—
xnr	Kangri	86	77.9	74.4	75.6	**79.1**	73.3	75.6	74.4	67.4	72.1	74.4	—
krl	Karelian	260	65.8	68.1	**69.6**	61.2	68.5	68.8	56.2	53.5	53.8	67.7	—
kxh	Karo (Ethiopia)	120	45.0	45.0	46.7	40.0	**50.0**	45.0	40.0	40.8	45.8	47.5	—
kaz	Kazakh	173	87.3	94.8	93.6	82.1	94.8	93.1	80.9	77.5	93.6	74.0	**97.1**
kir	Kirghiz	185	95.1	96.8	95.1	91.9	96.8	92.4	88.6	90.3	93.5	88.1	**98.4**
koi	Komi-Permyak	43	**60.5**	51.2	48.8	53.5	**60.5**	**60.5**	44.2	41.9	48.8	55.8	—
kpv	Komi-Zyrian	320	68.4	**77.2**	75.6	66.9	69.4	60.0	54.4	55.0	52.2	51.6	70.9
lat	Latin	3149	87.0	**92.1**	91.6	85.1	87.6	81.6	84.7	83.8	81.4	84.7	90.7
lav	Latvian	3032	82.5	89.2	87.5	75.6	95.1	90.6	75.5	73.9	88.7	96.6	**96.8**
lij	Ligurian	254	72.8	77.2	80.3	80.3	81.5	79.9	59.1	56.7	79.9	70.9	**94.5**
lit	Lithuanian	1180	93.6	96.6	96.1	92.3	98.1	95.5	85.1	83.2	96.3	**98.5**	97.9
olo	Livvi	190	79.5	**86.3**	82.1	67.4	83.2	70.5	67.4	60.0	66.8	76.8	—
nds	Low German	1774	71.6	**73.2**	73.0	72.2	71.1	67.4	65.1	62.7	68.8	70.8	72.6
mkd	Macedonian	39	94.9	**100.0**	97.4	92.3	97.4	92.3	82.1	79.5	89.7	79.5	**100.0**
mar	Marathi	460	81.7	85.4	90.2	76.5	90.2	88.3	83.0	85.9	86.3	60.4	**95.2**
frm	Middle French	294	96.9	**98.3**	97.6	94.2	95.6	91.2	96.9	93.9	94.9	96.9	—
ell	Modern Greek	1096	97.9	99.3	98.4	99.2	**99.5**	97.2	96.9	95.2	97.9	99.3	98.8
mdf	Moksha	82	56.1	43.9	52.4	56.1	56.1	58.5	50.0	50.0	51.2	51.2	**70.7**
yrl	Nhengatu	720	59.4	**63.1**	62.9	62.1	55.1	56.9	61.0	58.9	60.8	61.3	—
pcm	Nigerian Pidgin	26	92.3	96.2	96.2	**100.0**	**100.0**	92.3	96.2	**100.0**	96.2	92.3	**100.0**
kmr	Northern Kurdish	544	85.7	91.2	91.0	71.3	90.6	84.7	65.8	67.5	75.9	61.8	**94.7**
sme	Northern Sami	2536	68.2	71.7	72.2	67.5	72.5	69.4	66.0	63.7	66.6	66.2	**96.9**
fro	Old French	1976	78.3	**81.8**	81.0	76.3	73.7	69.2	76.7	74.6	73.1	72.4	—
orv	Old Russian	4615	76.9	**80.0**	78.4	77.6	78.6	72.5	72.2	69.6	72.7	75.4	—
ota	Ottoman Turkish	99	96.0	97.0	97.0	91.9	96.0	91.9	89.9	89.9	97.0	**99.0**	—
fas	Persian	2553	94.8	97.0	97.4	96.8	**97.5**	95.1	90.3	89.7	93.1	76.1	96.4
xpg	Phrygian	50	**92.0**	90.0	90.0	**92.0**	80.0	80.0	84.0	72.0	86.0	84.0	—
pol	Polish	3272	93.9	96.7	95.9	97.6	**98.8**	95.1	85.9	85.1	93.0	97.3	96.3
por	Portuguese	3048	96.9	**98.0**	97.8	97.7	97.2	93.5	93.5	93.4	96.7	97.1	94.4
ron	Romanian	2056	95.4	97.4	96.5	97.4	**97.7**	94.7	92.0	91.4	94.7	**97.7**	96.5
rus	Russian	3832	97.4	**98.6**	97.3	98.2	98.5	96.6	94.1	93.3	96.2	98.2	94.5
san	Sanskrit	4442	76.2	**81.4**	80.8	68.7	74.2	66.4	78.4	74.8	74.7	67.4	78.5
gla	Scottish Gaelic	66	92.4	**98.5**	93.9	90.9	97.0	95.5	**98.5**	93.9	93.9	87.9	97.0
hbs	Serbo-Croatian	3286	94.0	96.0	95.2	89.7	**98.3**	95.5	88.0	87.2	91.8	94.6	—
sms	Skolt Sami	263	78.3	74.9	74.9	78.3	64.3	64.6	73.0	71.1	**78.7**	77.9	—
slk	Slovak	4145	85.5	91.8	90.4	90.2	95.7	92.0	77.2	76.9	87.9	**96.3**	95.2
slv	Slovenian	4483	86.4	91.3	90.9	85.4	**94.2**	88.7	82.1	79.3	87.2	93.9	93.6
spa	Spanish	2541	97.8	98.2	97.7	98.0	97.7	95.0	96.0	95.2	97.0	**98.3**	96.1
arb	Standard Arabic	1215	91.0	94.5	92.7	**96.0**	95.4	90.8	87.3	86.4	91.9	93.6	95.2
swe	Swedish	201	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	99.5	99.5	99.0	99.0	99.5	**100.0**
tam	Tamil	382	97.9	**98.7**	98.4	98.2	97.9	97.6	96.9	96.9	97.1	90.8	98.2
ttc	Tektiteko	69	42.0	39.1	37.7	**44.9**	43.5	43.5	**44.9**	42.0	**44.9**	43.5	—
tpn	Tupinambá	9	0.0	**11.1**	0.0	0.0	**11.1**	**11.1**	**11.1**	**11.1**	0.0	0.0	—
tur	Turkish	1742	89.3	94.1	93.1	93.9	**97.2**	94.6	83.2	82.6	88.7	92.8	93.6
uig	Uighur	758	75.1	79.4	78.6	76.4	**80.9**	77.6	73.6	71.1	77.8	69.3	80.7
ukr	Ukrainian	2744	94.5	97.3	95.4	**98.1**	97.8	94.4	87.1	85.5	94.2	97.7	95.9
hsb	Upper Sorbian	186	71.5	79.6	79.6	75.3	**80.6**	78.5	66.1	64.0	60.2	65.1	**80.6**
urd	Urdu	550	96.7	**97.5**	97.1	93.3	**97.5**	96.7	96.5	96.2	97.1	87.6	96.4
urb	Urubú-Kaapor	13	46.2	46.2	46.2	46.2	38.5	38.5	38.5	46.2	46.2	**53.8**	—
uzb	Uzbek	50	94.0	**100.0**	**100.0**	94.0	**100.0**	**100.0**	88.0	86.0	94.0	86.0	**100.0**
vep	Veps	187	69.5	**70.6**	65.8	58.3	62.0	61.0	51.3	51.3	52.4	61.5	59.4
wbp	Warlpiri	12	25.0	8.3	8.3	16.7	25.0	16.7	16.7	25.0	25.0	**41.7**	—
cym	Welsh	1120	90.2	94.0	93.5	91.8	90.3	88.1	90.0	89.8	88.3	75.3	**99.4**
hyw	Western Armenian	1153	94.1	**96.7**	94.5	88.0	96.6	89.9	80.4	80.6	91.0	70.3	—
wol	Wolof	705	85.0	86.1	85.8	83.4	85.0	69.2	82.0	83.8	84.7	82.6	**97.0**
sah	Yakut	144	78.5	82.6	79.9	76.4	88.2	73.6	77.8	75.7	69.4	72.9	**95.8**
nhi	Tenango Nahuatl	38	52.6	**55.3**	47.4	47.4	44.7	47.4	**55.3**	50.0	47.4	36.8	—

Table 5: MultiBLiMP accuracy scores are split by language and LM. The best performing model per language is marked in **boldface**.