MELA: Multilingual Evaluation of Linguistic Acceptability

Anonymous ACL submission

Abstract

In this work, we present the largest benchmark to date on linguistic acceptability: Multilingual Evaluation of Linguistic Acceptability—MELA, with 48K samples covering 10 languages from a diverse set of language families. We establish LLM baselines on this benchmark, and investigate cross-lingual transfer in acceptability judgements with XLM-R. In pursuit of multilingual interpretability, we analyze the weights of fine-tuned XLM-R to explore the possibility of identifying transfer difficulty between languages. Our results show that GPT-4 performs on par with fine-tuned XLM-R, while open-source instruction-finetuned multilingual models lags behind by a notable gap. Cross-lingual and multi-task learning experiments show that unlike semantic tasks, in-language training data is crucial in acceptability judgements. We also conduct edge probing to investigate the different syntax capacities between base XLM-R and MELA-finetuned XLM-R. Results of probing indicate that training on MELA improves the performance of XLM-R on sytax-related probing tasks. Our dataset will be made publicly available upon acceptance.

1 Introduction

The acceptability judgment task tests a language model's ability to distinguish syntactically acceptable sentences like (1a) from unacceptable ones like (1b) in a human language - for instance, the following example on island constraints in English (Ross, 1967).

(1) a. Whose book did you find?
b. *Whose did you find book?

As a core linguistic competence, it has been argued in the literautre of Chomskyan generative syntax that much if not all of such syntactic competence is innate (Chomsky, 1965). That is, human brains are born with such knowledge already wired in. If the "innate" hypothesis were to be true and linguistic competence were unique in humans, it would naturally follow that any language model—with no "innate" linguistic knowledge to begin with—cannot be taught to acquire certain key linguistic competence.

There have been many attempts in computational linguistics and cognitive science to investigate this hypothesis, directly or indirectly, using either a data-driven approach, where examples created by theoretical linguists in published textbooks are collected, e.g., CoLA—Corpus of Linguistic Acceptability (Warstadt et al., 2019), or a theory-driven approach, where minimal pairs targeting specific syntactic phenomena are generated semi-automatically via some template (Warstadt et al., 2020; Xiang et al., 2021; Hu et al., 2020a).

There have been growing interests recently to expand the data-driven paradigm into other languages. For instance, CoLA-style datasets have been proposed in Russian (Mikhailov et al., 2022), Italian (Trotta et al., 2021) and Chinese (Hu et al., 2023). However, to date there are no multilingual benchmarks in this area which can be used to systematically test such abilities of multilingual models.

On the other hand, recently introduced evaluation benchmarks for Large Language Models (LLMs) such as GPT-4 (OpenAI, 2023) have mostly focused on application-driven tasks such as world knowledge and commonsense reasoning (Hendrycks et al., 2021; Srivastava et al., 2022), math reasoning (Cobbe et al., 2021), and code generation (Zhang et al., 2023). Few works, however, have investigated these models from a purely linguistic aspect.

To address these gaps, we introduce MELA—Multilingual Evaluation of Linguistic Acceptability, the first large-scale multilingual acceptability benchmark with 48k examples covering 10 languages from a diverse set of language families.

Language	L. F.	label	Examples	W. O.	Script	Gender	Casing
English (en)	Germ	1	One more pseudo generalization and I'm giving up.	SVO	Latin	N.A.	N.A.
Chinese (zh)	Sino-Tbt	0	张三被李四打了自己。	SVO	Han	N.A.	N.A.
Italian (it)	Rom	1	Quest'uomo mi ha colpito.	SVO	Latin	2	N.A.
Russian (ru)	Slavic	0	Этим летом не никуда ездили.	SVO	Cyrillic	3	6
German (de)	Germ	1	Die Frau sagt, dass ihm nicht zu helfen ist.	SVO	Latin	3	4
French (fr)	Rom	1	Je lui ait couru après.	SVO	Latin	2	N.A.
Spanish (es)	Rom	1	María bailó.	SVO	Latin	2	N.A.
Japanese (ja)	Altaic	0	犬が道端で死んである。	SOV	Han, Hiragana, Katakana	N.A.	N.A.
Arabic (ar)	Semitic	1	قال عمر إن كل السيارات استقدموها من ألمانيا.	VSO	Arabic	2	3
Icelandic (is)	Germ	1	Útlendingar gengu oft þennan stíg.	SVO	Latin	3	4

Table 1: Example sentences in the MELA training set, with information about the language family (L.F.), word order (W.O.), script, grammatical gender and casing for each language. Label "1" indicates the sentence is acceptable, "0" unacceptable. Data for the first four languages are from existing benchmarks while the rest are collected by us.

Data in four languages are from existing benchmarks mentioned above, and we complement them with newly collected data in six languages. Examples of MELA are demonstrated in Table 1. Following the CoLA tradition, all sentences in MELA are hand-written by linguists in respective languages, taken from textbooks, handbooks and journal articles in theoretical syntax, except for a small fraction of Russian sentences from Mikhailov et al. (2022).

We come up with three possible usages of MELA. In this work, we make a preliminary exploration in the following three directions:

Benchmarking We benchmark various multilingual LMs on MELA, including BLOOMZ (Scao et al., 2022; Muennighoff et al., 2023), mTk (Wang et al., 2022), mT0 (Muennighoff et al., 2023), Baichuan2-Chat (Yang et al., 2023), GPT-3.5 and GPT-4 (OpenAI, 2023).

Cross-lingual transfer We train XLM-R (Conneau et al., 2020) on different language combinations, finding in-language training data is crucial for acceptability judgements, in contrast to semantic tasks such as NLI (Conneau et al., 2018).

Syntax acquisition We probe the syntax capacity of MELA-finetuned XLM-Rs on syntax-related probing tasks, which indicates that XLM-R acquires syntax knowledge from the linguistic judgment task.

In the rest of this work, We first review relevant literature in §2, and then describe how we construct our benchmark MELA in §3. Next, we apply MELA as an evaluation benchmark for LLMS in §4. We investigate cross-lingual transfer in §5 and multi-task fine-tuning in §6. Finally, we probe the XLM-Rs trained on MELA for their syntax-related capacity in §7.

2 Related Work

2.1 Linguistic Acceptability

As we mentioned in §1, currently there exist four large-scale linguistic acceptability datasets: CoLA (Warstadt et al., 2019), ItaCoLA (Trotta et al., 2021), RuCoLA (Mikhailov et al., 2022), and CoLAC (Hu et al., 2023), all of which are annotated by expert linguists, while CoLAC also comes with an additional set of crowd labels.

Another line of work in linguistic acceptability is based on semi-automatic construction of example sentences, usually in minimal pairs. They compare the probabilities that language models assign to these sentences (Warstadt et al., 2020; Xiang et al., 2021), sometimes focusing on specific syntactic issues such as agreement (Varda and Marelli, 2023). A recent work also collects acceptability data in six Scandinavian languages (Nielsen, 2023), where the unacceptable examples are automatically generated by removing or swapping words in sentences from the Universal Dependency project.

In this work, we follow the CoLA style when building our benchmark, so that the unacceptable sentences are manually created by linguists to reflect certain syntactic constraints of the language in question. Compared with automatic methods, a wider coverage of syntactic phenomena is achieved in this way.

2.2 Multilingual Evaluation Benchmarks

XTREME (Hu et al., 2020b) and XGLUE (Liang et al., 2020) are two of the most popular multilingual evaluation benchmarks. Of the tasks therein, many are constructed by translating English samples entirely or partially into other languages, such as XNLI (Conneau et al., 2018), PAWS-X (Yang et al., 2019), and MLQA (Lewis et al., 2020).

Apart from these NLU benchmarks, the literature has also witnessed an abundance of multilingual generation benchmarks, ranging from summarization (Scialom et al., 2020; Ladhak et al., 2020) to translation (Fan et al., 2021; Goyal et al., 2022). After multitask instruction finetuning was found to unlock cross-task generalization ability in language models (Wei et al., 2022; Sanh et al., 2022), multilingual instruction datasets have also been proposed, represented by Supernatural Instruction (Wang et al., 2022) and xP3 (Muennighoff et al., 2023).

3 MELA: Multilingual Evaluation of Linguistic Acceptability

MELA consists of more than 48 thousand acceptability samples across 10 languages from a diverse group of language families. Specifically, it contains three Germanic languages: English, German and Icelandic, three Romance languages: Spanish, French and Italian, one Slavic language Russian, one Sino-Tibetan language Chinese, one Japonic language Japanese, and one Semitic language Arabic. Table 1 shows example sentences and properties of each language in MELA. For dataset statistics, see Table 2.

3.1 Data collection Procedure

High-resource languages. We use four existing datasets for four languages in MELA: CoLA (Warstadt et al., 2019) for English, ItaCoLA (Trotta et al., 2021) for Italian, RuCoLA (Mikhailov et al., 2022) for Russian, and CoLAC for Chinese (Hu et al., 2023), each having more than 6,000 data points. Since the out-of-domain samples of RuCoLA are produced by generative models, we additionally collected 1037 Russian samples from *The Syntax of Russian* (Bailyn, 2011a) (with the procedure described below) and add them 50-50 to the development and test sets of the Russian portion to keep a balance between validation-test discrepancy and generalization.

Low-resource languages. Apart from the four existing acceptability datasts, we also collected samples in 6 new languages, all annotated by theoretical syntacticians in their respective languages. These sentences are taken from five books/textbooks in the Cambridge Syntax Guides series, namely *The Syntax of German* (Bailyn, 2011b), *The Syntax of French* (Rowlett, 2007), *The Syntax of Spanish* (Zagona, 2001), *The Syntax of Arabic* (Aoun et al., 2009) and *The Syntax of Icelandic* (Thráinsson, 2007). Japanese data were collected from *Handbook of Japanese Syntax* (Shibatani et al., 2017).

Each book contains roughly one to three thousand example sentences with acceptability judgments made by linguists in respective languages. Graduate students majoring in linguistics in these languages were paid to extract all example sentences with their judgments in these books manually. Note that, following previous CoLA-style corpora, we only keep sentences labelled with * or ?? as our unacceptable sentences. All unmarked sentences are extracted as acceptable sentences.

Following previous acceptability datasets, we remove examples when the judgment is based on coindexing of pronouns, empty categories, prosody or semantic/pragmatic interpretation. We also complete the sentence if it composed of only a phrase, while keeping the judgment.

For Japanese, we remove examples of its dialects (N=99) and those about classical Japanese (N=13). For Arabic and Russian, as the original sentences are written in transliterations, we also convert them to their respective scripts manually.

The mean time for data collection for one language is about a month, with Icelandic taking about 3 months as there were more examples in the book.

As these books/textbooks and handbook are overviews of syntax of each language, we believe they cover a wide range of linguistic phenomena in these languages, and can therefore serve as a good resource to evaluate language models' *overall* ability to distinguish acceptable sentences from unacceptable ones.

ISO code	English en	Chinese zh	Russian ru	Italian it	German de	French fr	Spanish es	Japanese ja	Arabic ar	Icelandic is
Train	8551	6072	7869	7801	500	500	500	500	500	500
Dev	527	492	1483	946	402	521	321	693	313	1194
Test	516	931	2341	975	402	521	322	694	313	1194
acceptable%	70.3	66.4	73.2	84.5	75.5	85.0	73.1	80.0	74.7	75.1
len (char)	40.8	11.7	56.2	36.0	49.0	29.0	31.7	16.1	22.2	32.9
len (byte)	40.8	35.0	102.8	36.3	49.6	29.7	32.9	47.7	40.8	36.7
len (token)	10.5	9.5	15.2	9.7	11.5	8.1	8.7	11.1	7.9	9.7

Table 2: Statistics of MELA: train/dev/test splits, acceptable rate, and average sentence length by characters, bytes, and tokens (using the tokenizer of XLM-R (Conneau et al., 2020)).

3.2 Resulting Corpus and Data Split

The resulting corpus contains more than 48k example sentences in 10 languages.

For Italian and Chinese, we use the original train/dev/test splits of ItaCoLA and CoLAC, and for CoLAC we use the crowd label following Hu et al. (2023) (see Appendix C for the alternative). For English and Russian, we keep the training splits of CoLA v.1.1 and RuCoLA, and use their in-domain development sets as our validation sets, and their out-of-domain development sets as our test sets.

For the six low-resource languages, we randomly sample 500 sentences from each of these languages to construct a training set, and divide the remaining sentences equally between validation and test sets.[1]

3.3 Comparison with Other Multilingual Benchmarks

We note that all samples in MELA are constructed individually in each language. While some early multilingual benchmarks opt to translate English sentences into other languages to obtain parallel samples (Conneau et al., 2018; Lewis et al., 2020), this approach does not suit our case. Firstly, as Clark et al. (2020) argue, translation introduces artifacts into multilingual benchmarks and often results in translationese. Secondly, the task of linguistic acceptability is highly language-dependent, and syntactic phenomena in one language most likely cannot be captured in another language through translation.

4 Evaluating LLMs with MELA

In this section, we report the performance of several LLMs, open-sourced or close-sourced, on MELA.

[1]We experimented with another split of these data and observe similar results in all experiments that follow.

4.1 Experimental Settings

For open-sourced models, we consider BLOOMZ (Scao et al., 2022; Muennighoff et al., 2023), two instruction finetuned variants of mT5 (Xue et al., 2021)—namely mTk (Wang et al., 2022) and mT0 (Muennighoff et al., 2023)—and Baichuan2-Chat (Yang et al., 2023). BLOOMZ is both pretrained and finetuned on 46 languages, which only covers 5 languages in MELA: English, Chinese, French, Spanish, and Arabic[2]. The pretraining corpus of mT5 includes all 10 languages in MELA, but mT0 is finetuned on the same instruction dataset as BLOOMZ. mTk's finetuning data, on the other hand, covers nine languages in MELA (the left out one is Icelandic) and includes the English CoLA dataset. For Baichuan2, the exact language distribution of pretraining and finetuning data is not disclosed. For close-sourced models, we consider GPT-3.5 and GPT-4 (OpenAI, 2023).

When evaluating mTk, we use 2-shot prompts following the format of its finetuning dataset. For other models, we consider both 0-shot and 2-shot evaluation. More details about the prompts used for evaluating these models are given in Appendix A.

4.2 Results

The results of LLMs' performance on MELA are given in Table 3. We make the following observations.

GPT-4 performs on par with supervised models. It performs only five points below XLM-R in the zero-shot setting, and only one point below it in two-shot setting. On German, French and Spanish even the zero-shot performance of GPT4

[2]Muennighoff et al. (2023) examine BLOOM's pretraining corpus ROOTS and estimate it to also contain a small amount of Russian, German, Italian, and Japanese.

model	size	examples	en	zh	it	ru	de	fr	es	ja	ar	is	avg
Supervised													
XLM-R	550M	-	70.65	55.20	**53.97**	**50.04**	37.60	22.46	45.89	44.90	**30.59**	**35.39**	**44.67**
Open-sourced, instruction-finetuned													
BLOOMZ^0	7.1B	-	-2.28	9.99	-1.34	-1.60	-0.90	-3.20	-0.91	-1.86	7.50	3.46	0.88
BLOOMZ^2	7.1B	in-lang.	7.74	17.63	4.87	-0.25	-0.14	-0.34	7.05	3.81	-1.85	-1.92	3.66
BLOOMZ^2	7.1B	en	7.74	12.53	2.66	-0.47	3.86	-1.52	6.30	2.31	-2.36	-3.92	2.71
mT0^0	13B	-	7.32	20.13	10.83	1.95	10.28	0.39	9.32	13.71	0.04	4.95	7.89
mTk^2	13B	in-lang.	39.13	32.18	18.26	11.83	9.91	13.09	24.42	22.45	12.72	15.54	19.95
mTk^2	13B	en	39.13	31.48	12.12	14.92	16.46	12.81	15.77	15.17	6.34	11.21	17.54
Baichuan2-Chat^0	13B	-	13.46	15.78	7.07	13.29	5.77	3.34	16.43	13.85	5.76	-0.98	9.38
Baichuan2-Chat^2	13B	in-lang.	27.26	25.89	13.14	7.23	6.78	6.68	16.43	17.87	3.04	0.94	12.52
Baichuan2-Chat^2	13B	en	27.26	14.88	7.44	1.97	2.76	9.54	13.77	10.19	3.04	-1.05	8.98
Close-sourced													
GPT-3.5^0	-	-	37.16	30.34	29.43	17.88	29.51	25.59	49.23	31.71	10.24	5.97	26.71
GPT-3.5^2	-	in-lang.	67.00	45.64	38.46	24.47	27.29	23.63	**59.76**	38.71	18.42	14.60	35.80
GPT-3.5^2	-	en	67.00	15.22	13.99	8.11	13.44	13.86	38.60	16.61	5.28	3.76	19.59
GPT-4^0	-	-	69.31	50.75	35.57	37.87	**43.03**	32.45	51.52	45.87	16.44	9.88	39.27
GPT-4^2	-	in-lang.	**72.29**	**55.57**	51.40	38.31	36.54	**35.57**	56.16	**49.36**	17.09	22.60	43.49
GPT-4^2	-	en	**72.29**	45.49	14.57	-0.94	23.48	12.97	43.66	39.01	3.04	6.62	26.02

Table 3: Validation performance of large language models, in comparison with XLM-R finetuned on MELA training set (all 10 languages). Superscripts denote the number of in-context examples. The 2-shot performance of mT0 is below random guess (i.e. smaller than 0) and not presented here See Table 7 and 8 for the complete results.

is noticeably higher than XLM-R. On Arabic and Icelandic, however, it lags behind even in the two-shot setting, suggesting that GPT-4 may be weaker at understanding these languages.

In few-shot evaluation, using only English examples hurts performance. As indicated by the results of GPT-3.5, GPT-4 and Baichuan-2, prompting with two English examples leads to even lower performance than 0-shot evaluation. In contrast, prompting with English instructions and in-language examples boosts performance. This suggests that these LLMs fail to transfer the concept of linguistic acceptability acquired from the in-context examples across languages.

Instruction finetuning on acceptability judgements helps cross-lingual transfer. Of the open-source instruction-finetuned models, mTk performs much better than other models, as its finetuning dataset includes English CoLA. However, mTk also performs much better in non-English examples, and its performance gap between prompting with in-language and English examples is much smaller compared with Baichuan or GPT, suggesting that finetuning on acceptability judgements may unlock the ability of cross-lingual generalization in this task.

5 Cross-lingual Transfer of Linguistic Acceptability

In this section, we investigate cross-lingual transfer in linguistic acceptability by finetuning XLM-RoBERTa (Conneau et al., 2020), which is a multilingual version of RoBERTa (Liu et al., 2019) pretrained on 2.5TB CommonCrawl corpus covering one hundred languages.

5.1 Experimental Settings

To observe the transfer of acceptability judgements across languages, we train the model on one language, and evaluate it on all ten languages' development sets. Further training details can be found in Appendix B. We report the median MCC of seven runs for all results to mitigate inter-run variance.

5.2 Results

The main results of cross-lingual transfer in acceptability judgements are presented in Table 4. Here we make several key observations.

The first is that **the ability to perform judgement of linguistic acceptability can be transferred non-trivially across languages**, as indicated by the last column of Table 4[3]. The second

[3]The evaluation metric used for acceptability judgements, namely MCC, is designed such that random guessing would

↓train (size) / eval→	en	zh	it	ru	de	fr	es	ja	ar	is	avg
en (8551)	**71.66**	47.41	28.23	31.91	24.85	18.96	**32.21**	**34.50**	21.50	24.47	**33.57**
zh (6072)	45.72	**52.71**	23.18	22.80	21.31	17.61	29.01	31.48	22.16	20.57	28.65
it (7801)	39.13	34.86	**53.75**	17.02	17.23	21.23	22.46	20.10	19.87	17.92	26.36
ru (7869)	50.29	39.77	24.26	**47.22**	20.47	14.11	28.62	32.48	20.11	24.49	30.18
de (500)	35.87	37.97	15.44	18.38	**36.13**	16.45	22.06	22.68	12.27	21.67	23.89
fr (500)	18.57	21.16	6.52	9.19	9.85	**29.73**	14.28	13.32	11.63	12.74	14.70
es (500)	35.48	38.76	17.71	16.01	11.43	11.38	26.75	24.48	19.14	13.46	21.46
ja (500)	22.67	20.32	10.20	12.40	13.82	10.44	10.81	33.62	8.85	11.21	15.43
ar (500)	9.26	13.34	6.52	3.12	11.95	10.44	8.82	5.90	**37.42**	7.61	11.44
is (500)	27.40	23.16	9.82	11.60	7.58	18.72	18.45	12.46	7.50	**25.12**	16.18
avg. high-resource	51.70	43.69	32.35	29.74	20.96	17.98	28.07	29.64	20.91	21.86	29.69
avg. low-resource	24.88	25.79	11.04	11.78	15.13	16.19	16.86	18.74	16.14	15.30	17.18
avg. w.o. in-lang.	31.60	30.75	15.76	15.83	15.39	15.48	20.75	21.93	15.89	17.13	-

Table 4: Cross-lingual transfer results of finetuned XLM-R. The top four training languages are high-resource languages in MELA (whose training samples vary from 6000 to 8500). The middle six are low-resource languages in MELA (all of which have 500 training samples). All results are the median MCC of seven runs. "Avg. high-resource" refers to the average of the first four rows, while "avg. low-resource" is the average of the next six rows. To illustrate the effects of in-language training, figures in the last row are the average MCC on each language's validation set of 9 rows, except the one where the model is trained in-language.

is that **in-language training significantly boosts XLM-R's performance**. Comparing the figures on the diagonal with the last row, this is most prominent for the four high-resource languages. For example, when evaluating on English, training on English leads to 71.66 MCC, compared with an average of 31.60 when training on other nine languages. For low-resource languages, the gap is smaller, but still notable (e.g. for Icelandic the comparison is 25.12 against 17.13). However, we note that for Spanish and Japanese, the highest performance is not obtained when training in-language, but training on English. This leads to our third observation—**the number of training samples matters**. As indicated by the antepenultimate and penultimate lines of Table 4, when training on high-resource languages, XLM-R obtains an average of 29.69 MCC, compared with 17.18 when training on low-resource languages.

6 Multi-task Fine-tuning with Linguistic Acceptability

In previous two sections, we investigated the transfer of linguistic acceptability with both LLMs and supervised XLM-R, and found that in-language training data or in-context examples play a key role in linguistic acceptability. To further assess the importance of in-language data, we experiment with training on multiple languages, i.e. multitask finetuning (MFT) on acceptability judgement[4]. To compensate the impact of training set size, we first downsample data in all languages to the same amount, and then finetune XLM-R on different combinations of languages. Training details are provided in Appendix B.

6.1 Experimental Setting

We downsample sentences in each language to the same number, and train XLM-R in three settings: 1) in-language finetuning; 2) all-language multitask finetuning, where the model is trained on a mixture of data containing an equal number of sentences from ten languages; and 3) all-but-in-language multitask finetuning, where the model is trained on a mixture of data containing an equal number of sentences from nine languages, except the one being evaluated on. Additional experiments on bilingual training are provided in Appendix D.1.

6.2 Results

The results on ten languages' validation sets are plotted in Figure 1. When trained and evaluated on the same language, the model's performance scales smoothly with the number of training sam-

result in 0 performance, regardless of class imbalance.

[4]Following Hu et al. (2023), we regard linguistic acceptability in each language as a related but different task, since the negative samples in MELA are constructed by (manually) injecting language-specific grammar errors into sentences.

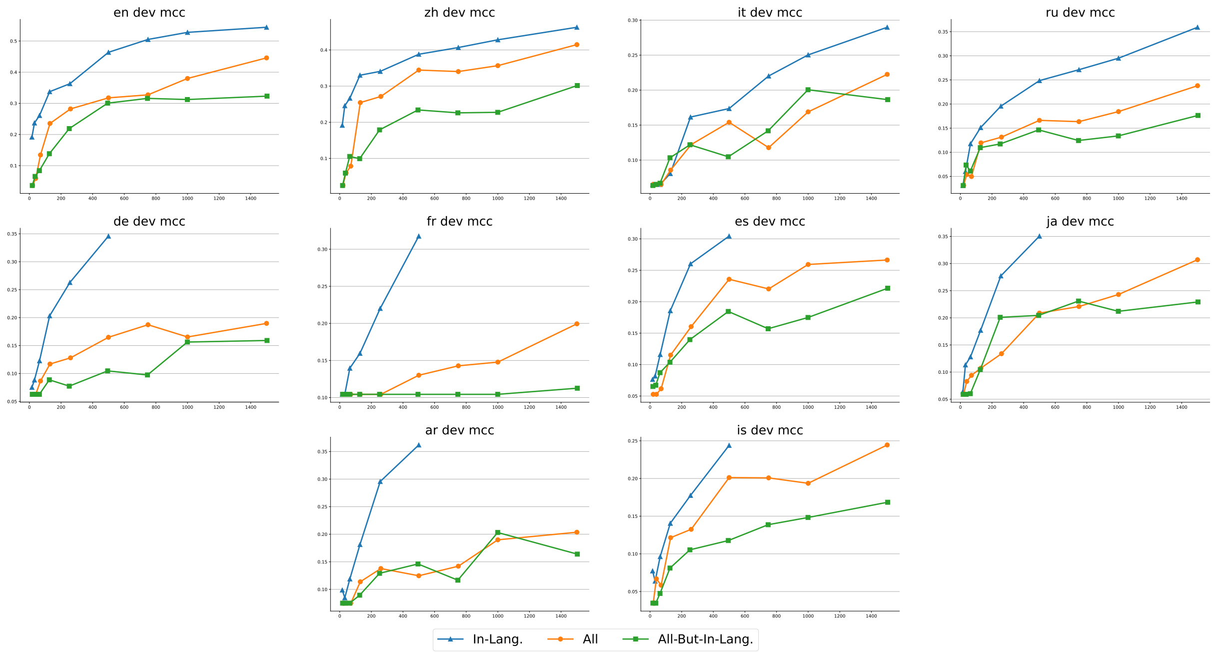

Figure 1: Performance of XLM-R when finetuned on different languages. The horizontal axis indicates the number of training samples per language. For example, for "all" curves, the point at 500 indicates the model is trained on 500 sentences, with 50 from each language. For "All-but-in-lang." curves, the point at 495 indicates the model is trained on 495 sentences, with 45 from each of the nine language except the one being evaluated on.

ples. When trained on sentences from other nine languages, however, the performance saturates at around 500-1000 training samples, consistent with previous findings about multi-task finetuning (Wang et al., 2022). When trained on all ten languages, the performance scales more steadily than all-but-in-language training, but still lags behind in-language training by a large margin, indicating the importance of in-language training data.

7 Edge Probing

In this section, we adopt *edge probing* (Tenney et al., 2019b,a) to explore whether training on linguistic acceptability tasks improves syntax-related capacity to the pre-trained XLM-R.

7.1 Preliminaries

Edge probing is designed to investigate how much encoders encode syntactic and semantic information, which is highly related to the acceptability judgment from a generative linguistic perspective.

To achieve this goal, edge probing focuses on structural labeling tasks in form of span labeling. Given one or two spans, the probing classifier is trained to predict the label with span representations encoded by pre-trained encoders (XLM-R in our case).

Task	base	en	it	ru	zh
`pos`	**92.87**	93.77	93.47	93.17	93.95
`dep`	**89.41**	90.34	90.13	89.92	89.86
`const`	78.54	79.10	**78.44**	79.26	78.96
`name`	93.49	94.23	**93.34**	94.53	94.08
`srl`	**77.93**	82.34	80.00	81.24	80.28
`coref`	**83.84**	85.55	84.12	83.98	84.53
avg	**86.01**	87.56	86.58	87.02	86.94

Table 5: F1 scores of Experiment 1 on six edge probing tasks. **Bold** denotes the lowest score in one task. We train probing classifiers using span representations from different XLM-R variants (base, en, it, ru, and zh).

For instance, dependency labeling is a typical probing task, but it should be discriminated from dependency parsing. In dependency parsing, the parser should find out: a) the head and dependent, and b) the dependency relation between them. On the contrast, in dependency labeling the head and dependent are given so that the task is only about predicting the label between two words. Other tasks follow the same labeling scheme.

7.2 Experiment Settings

We hypothesize that training on MELA can improve the syntax-related capacity of XLM-

Probing task		Part-of-speech tagging					Depedency labeling				
↓eval / train→		en	it	ru	zh	**avg**	en	it	ru	zh	**avg**
en	base	92.87	75.77	65.63	43.33	69.40	89.41	74.99	60.67	40.05	66.28
	XLM-R$_{en}$	93.77	81.43	68.22	44.66	**72.02**	90.34	77.40	61.84	45.44	**68.76**
it	base	83.26	94.61	66.90	38.73	70.88	78.17	91.50	60.65	32.35	65.67
	XLM-R$_{it}$	85.6	95.71	63.73	39.70	**71.19**	83.56	92.46	62.85	37.31	**69.05**
ru	base	82.97	79.90	95.53	53.18	77.90	77.72	78.86	90.90	42.77	72.56
	XLM-R$_{ru}$	85.42	81.01	95.43	54.06	**78.98**	80.65	81.27	92.04	46.10	**75.02**
zh	base	61.19	58.57	64.43	93.88	69.52	50.16	43.42	43.12	86.06	55.69
	XLM-R$_{zh}$	64.55	55.60	63.98	94.35	**69.62**	55.42	44.52	44.16	87.73	**57.96**

Table 6: F1 scores of Experiment 2 on part-of-speech tagging and depedency labeling in a cross-lingual setting. **Bold** denotes a better performance in average between XLM-R$_{base}$ and XLM-R$_{lang}$. We conduct pair comparison between XLM-R$_{base}$ and XLM-R$_{lang}$ trained on MELA of one language to investigate whether linguistic accepability helps the cross-lingual transfer in above two probing tasks.

RoBERTa. To verify this intuition, we design following experiments.

Probing Tasks We choose six edge probing tasks: 1) part-of-speech tagging, 2) dependency labeling, 3) constituency labeling, 4) named entity labeling, 5) semantic role labeling, and 6) co-reference.

We incorporate multilingual data into probing. POS tagging and dependency labeling are from Universal Dependencies V2.13 (De Marneffe et al., 2021), including four MELA-high-resource language (i.e., English, Italian, Russian, and Chinese). The other four monolingual English tasks are sampled from OntoNotes 5.0 (Weischedel et al., 2013).

Experiment 1 We train probing classifiers using span representations from XLM-Rs on English probing tasks. We set the pre-trained XLM-R$_{base}$ as control group, and the other four MELA-finetuned XLM-R$_{lang}$ (trained respectively on four MELA-high-resource languages) as test group.

Experiment 2 For two tasks (*pos* and *dep*) with mulitlingual data available, we experiment on zero-shot cross-lingual transfer. We train probing classifiers on representations from XLM-R$_{base}$ in each of four high-resource languages, and run zero-shot evaluation on a target language (*lang*). We repeat the procedure on XLM-R$_{lang}$ (see more details in Appendix E.)

7.3 Results

Training on the linguistic acceptability judgment task indeed improves the performance of XLM-R on syntax-related probing tasks, which supports our hypothesis driven by linguistic intuition.

In Experiment 1, we train probing classifiers using representations from different XLM-R variants. The average performance of XLM-R$_{base}$ is the lowest across the six edge probing tasks (see in Table 5). In Experiment 2, we compare performances of cross-lingual transfer between XLM-R$_{base}$ and XLM-R$_{lang}$ (see in Table 6). The results indicate that training on MELA of one language helps zero-shot transfer to that language in part-of-speech tagging and dependency labeling.

These results match our linguistic intuition. In generative linguistics, the scheme of analyzing the grammaticality and acceptability relies on sub tasks, including categorization of words, combination of lexical items into constituency, and assignment of semantic role to arguments. Therefore, we assume that there might be a similar pattern with human regarding syntax acquisition.

8 Conclusion

In this work we present MELA, the first multilingual acceptability judgement benchmark covering a diverse set of languages, all annotated by expert linguists. By benchmarking multilingual LLMs on MELA and finetuning XLM-R in different cross-lingual settings, we find that GPT-4 performs on par with supervised XLM-R, and in-language data is crucial, both for few-shot evaluation and supervised finetuning. We probe MELA-finetuned XLM-R for syntax capacity, finding that training on MELA improves the performance on syntax-related probing tasks, which indicates that language models acquire syntax knowledge during training on linguistic acceptability judgements.

Limitations

Due to the large amount of human labor involved in transcribing and examining the sentences in MELA, the dataset only covers ten languages, of which six are low-resource, with only a small number of training samples. In the future, we intend to expand the dataset by additionally collecting data in other languages, especially non-Latin and non-Indo-European languages, which are currently underrepresented in MELA.

Also, in this work we focused on introducing the MELA dataset and showcasing some of its usages, such as serving as a benchmark for evaluating LLMs and providing a data resource for cross-lingual researches in computational linguistics. We did not propose any new theory, method, or model to improve the understanding of linguistic acceptability in humans or language models. We leave the exploration of other use cases of MELA to future works.

Ethics Statement

Sentences in our dataset MELA, including those in English, Italian, Russian, and Chinese consolidated from previous works, are sourced from renounced linguistics publications such as syntax textbooks and journal articles. Therefore, we believe they do not raise any ethical issues such as leak of personal identifiable information.

The sentences in MELA, both acceptable and unacceptable, are only intended for researches concerning the acquisition and evaluation of linguistic capabilities (of either humans or language models), and should not be interpreted otherwise. We release MELA under Apache 2.0 license, and note that for the four existing acceptability datasets, RuCoLA is available under Apache 2.0 license, while the authors of CoLA, ItaCoLA, and CoLAC did not provide any license information along with their released datasets. For the individual sentences in MELA, the copyright (where applicable) remains with the original authors or publishers.

References

Joseph E. Aoun, Elabbas Benmamoun, and Lina Choueiri. 2009. *The Syntax of Arabic*. Cambridge Syntax Guides. Cambridge University Press.

John Frederick Bailyn. 2011a. *The Syntax of Russian*. Cambridge Syntax Guides. Cambridge University Press.

John Frederick Bailyn. 2011b. *The Syntax of Russian*. Cambridge Syntax Guides. Cambridge University Press.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*, 50 edition. The MIT Press.

Jonathan H. Clark, Jennimaria Palomaki, Vitaly Nikolaev, Eunsol Choi, Dan Garrette, Michael Collins, and Tom Kwiatkowski. 2020. Tydi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Trans. Assoc. Comput. Linguistics*, 8:454–470.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168.

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, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2475–2485. Association for Computational Linguistics.

Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational linguistics*, 47(2):255–308.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22:107:1–107:48.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Trans. Assoc. Comput. Linguistics*, 10:522–538.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Hai Hu, Ziyin Zhang, Weifang Huang, Jackie Yan-Ki Lai, Aini Li, Yina Ma, Jiahui Huang, Peng Zhang, and Rui Wang. 2023. Revisiting acceptability judgements. *CoRR*, abs/2305.14091.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020a. 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.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020b. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *CoRR*, abs/2003.11080.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, Online. Association for Computational Linguistics.

Patrick S. H. Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7315–7330. Association for Computational Linguistics.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Vladislav Mikhailov, Tatiana Shamardina, Max Ryabinin, Alena Pestova, Ivan Smurov, and Ekaterina Artemova. 2022. Rucola: Russian corpus of linguistic acceptability. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 5207–5227. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M. Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 15991–16111. 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.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

John Ross. 1967. *Constraints on variables in syntax*. Ph.D. thesis, MIT.

Paul Rowlett. 2007. *The Syntax of French*. Cambridge Syntax Guides. Cambridge University Press.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. MLSUM: the multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 8051–8067. Association for Computational Linguistics.

Masayoshi Shibatani, Shigeru Miyagawa, and Hisashi Noda, editors. 2017. *Handbook of Japanese Syntax*. De Gruyter Mouton, Berlin, Boston.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew K. Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakas, and et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *CoRR*, abs/2206.04615.

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

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

Höskuldur Thráinsson. 2007. *The Syntax of Icelandic*. Cambridge Syntax Guides. Cambridge University Press.

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, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 2929–2940. Association for Computational Linguistics.

Andrea Gregor de Varda and Marco Marelli. 2023. Data-driven Cross-lingual Syntax: An Agreement Study with Massively Multilingual Models. *Computational Linguistics*, 49(2):261–299.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022. Super-naturalinstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 5085–5109. 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. *Trans. Assoc. Comput. Linguistics*, 8:377–392.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Trans. Assoc. Comput. Linguistics*, 7:625–641.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Ralph Weischedel et al. 2013. Ontonotes release 5.0. Web Download. LDC2013T19.

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, EACL 2021, Online, April 19 - 23, 2021*, pages 2784–2790. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 483–498. Association for Computational Linguistics.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Haizhou Zhao, Hang Xu, Haoze Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, Juntao Dai, Kun Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang Wang, Mickel Liu, MingAn Lin, Nuolan Nie, Peidong Guo, Ruiyang Sun, Tao Zhang, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xiangrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yanjun Shen, Yiding Wang, Yiyu Li, Youxin Jiang, Yuchen Gao, Yupeng Zhang, Zenan Zhou, and Zhiying Wu. 2023. Baichuan 2: Open large-scale language models. *CoRR*, abs/2309.10305.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3685–3690. Association for Computational Linguistics.

Karen Zagona. 2001. *The Syntax of Spanish*. Cambridge Syntax Guides. Cambridge University Press.

Ziyin Zhang, Chaoyu Chen, Bingchang Liu, Cong Liao, Zi Gong, Hang Yu, Jianguo Li, and Rui Wang. 2023. Unifying the perspectives of nlp and software engineering: A survey on language models for code. *CoRR*, abs/2311.07989.

A Prompts

A.1 mTk

The prompt for evaluating mTk is presented in Figure 2, which reuses the prompt for Supernatural Instruction task 616[5]. In Table 7 we present the results on validation sets when prompting with examples from different languages' training sets. Each figure is the median of three sets of randomly selected prompts (always one acceptable and one unacceptable).

A.2 Other Open-Source models

Since linguistic acceptability is not included in the finetuning mixture of mT0, BLOOMZ, and Baichuan2, we experiment with several prompts imitating the style of both mTk's prompt for CoLA and the prompts given by Muennighoff et al. (2023), and the results are presented in Table 13. For the main experiments we use the best prompt selected on this subset of training set, i.e. the seventh prompt in Table 13. For few-shot evaluation, we experiment with both in-language examples and English-only examples, and report the median of three sets of prompts. For the instruction itself we always use English.

A.3 OpenAI Models

For OpenAI models, we use the 0613 version of GPT-3.5-turbo and GPT-4. Due to the limited budget, we choose to use the fifth prompt in Table 13 without further ablations, which is found to perform reasonably well in preliminary experiments. For few-shot evaluation we also experiment with in-language examples and English-only examples. We experimented with different sets of in-context examples with GPT-3.5, and found it to have limited impact: the average MCC for 2-shot evaluation of GPT-3.5 with in-language examples reported in Table 3 is 35.80, and two other runs with different examples yield 37.20 and 36.56, respectively.

B Training Details

For experiments concerning XLM-R in §5, §4 and §6, we finetune with learning rate 7.5e-6, weight decay 0.075 and batch size 32. To minimize confounding variables and accentuate the interaction across languages in terms of linguistic acceptability performance, we train the model for 15k steps for all experiments in §4 and §5 with 750 steps of linear warmup and cosine learning rate decay over 0.4 cycles, and take the best checkpoint based on validation results. For experiments on downsampled data in §6 and Appendix D.1 the model is trained for 5K steps instead, and validation is performed every 250 steps. The training is conducted on a single RTX 3090 with 24GB RAM.

We note that these hyperparameters are chosen based on previous works on similar tasks (Liu et al., 2019; Hu et al., 2023) and our preliminary experiments. The sheer amount of experiments covered in our work makes it impossible to finetune hyperparameters on each combination of training data, and we thus decide to keep them fixed across all experiments for a fair comparison across languages, which may be suboptimal for certain cases. Hu et al. (2023), for example, report 56.45 MCC for XLM-R on CoLAC development set, while our result is 52.71 with the same training data.

We also note that finetuning language models on linguistic acceptability data leads to large performance variations, regardless of the specific lan-

[5] https://github.com/allenai/natural-instructions/blob/master/tasks/task616_cola_classification.json

Definition: You’re given a sentence and your task is to classify whether the sentence is acceptable or not. Any sentence which is grammatically correct, has a naturalistic text, is written by a native speaker and which minimizes superfluous content is acceptable, otherwise unacceptable. If the sentence is acceptable then write "acceptable", otherwise "unacceptable".
Positive Example 1–
    input: {example1}
    output: acceptable
Positive Example 2–
    input: {example2}
    output: unacceptable
Now complete the following example–
    input: {sent}
    output:

Figure 2: Prompt for evaluating mTk.

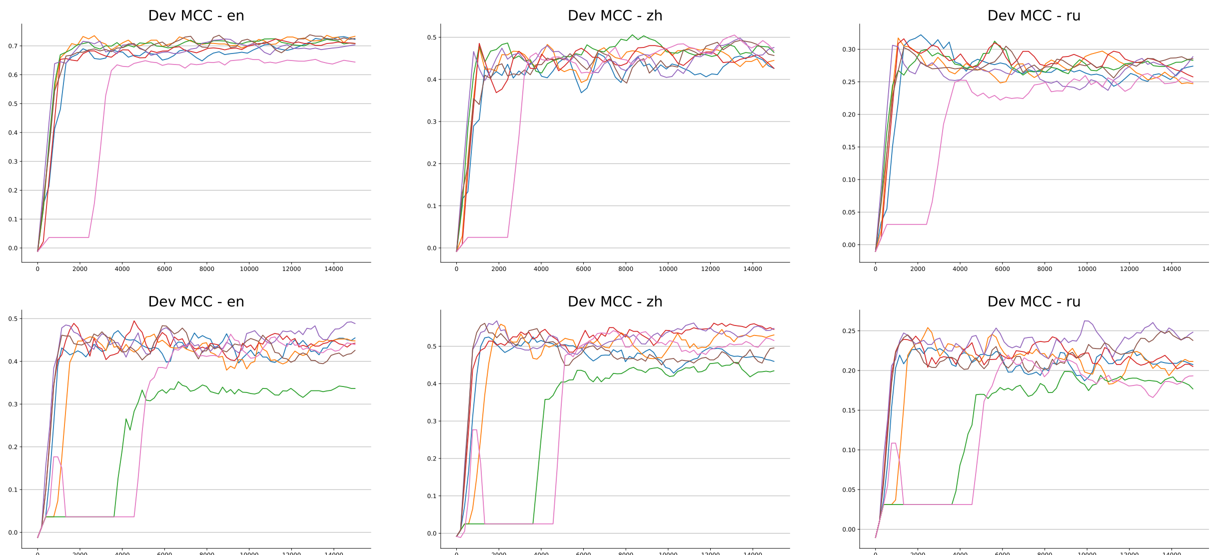

Figure 3: Interrun variance when finetuning XLM-R on English (first row) and Chinese (second row) training data. Each subfigure plots the validation MCC of seven runs with different random seeds on one language. After taking the median of these seven runs, this variance is mitigated to a large extent.

prompt	en	zh	it	ru	de	fr	es	ja	ar	is	avg
en	**39.13**	31.48	12.12	**14.92**	**16.46**	12.81	15.77	15.17	6.34	11.21	17.54
zh	18.82	**32.18**	11.42	8.94	2.96	3.50	17.13	16.92	8.27	13.81	13.39
it	11.67	25.64	**18.26**	11.54	4.75	8.45	24.30	20.01	9.43	12.62	14.67
ru	14.32	26.47	11.37	11.83	4.39	10.34	15.56	20.60	10.01	**16.09**	14.10
de	15.39	24.40	12.09	9.05	9.91	6.80	19.59	15.63	7.89	13.40	13.42
fr	13.29	25.41	13.15	12.02	9.29	**13.09**	17.52	13.42	8.86	13.04	13.91
es	15.09	26.78	14.15	13.83	6.78	6.99	**24.42**	18.14	14.76	11.64	15.26
ja	13.52	26.74	12.15	2.99	0.59	2.52	11.68	**22.45**	4.00	12.26	10.89
ar	22.14	25.56	16.81	13.32	9.05	12.61	17.82	16.30	**12.72**	7.37	15.37
is	9.89	23.25	9.16	5.54	4.79	5.60	13.14	15.53	6.84	15.54	10.93
in-lang.	**39.13**	**32.18**	**18.26**	11.83	9.91	**13.09**	**24.42**	**22.45**	**12.72**	15.54	**19.95**

Table 7: Validation performance of mTk, with in-context examples from different languages.

Model	Examples	en	zh	it	ru	de	fr	es	ja	ar	is	avg
	None	7.32	20.13	10.83	1.95	10.28	0.39	9.32	13.71	0.04	4.95	7.89
mT0	In-language	-3.64	-2.51	-6.52	-3.12	-6.27	-10.44	-5.27	-5.90	-7.50	-3.46	-5.46
	English	-3.64	-2.51	-6.52	-3.12	-6.27	-10.44	-5.27	-5.90	-7.50	-3.46	-5.46

Table 8: 0-shot (rows marked as 'None') and 2-shot validation performance of mT0 13B. The model performs exactly the same when prompted with English and in-language examples.

guages (see Figure 3), which corresponds with previous findings in the literature (Raffel et al., 2020). We thus train with seven different random seeds for every experiment in this work to reduce this variance, and the reported scores are computed by first taking the median of these seven runs at each checkpointing step, and then maxing over all the aggregated checkpoints. For experiments on downsampled data in §6, each run also select a different subset of training data.

C Alternative Labels for CoLAC

Hu et al. (2023) propose two sets of labels for the Chinese acceptability corpus CoLAC, and in MELA we adopt the crowd label following their suggestions. In Table 9 we present additional experimental results of finetuning and evaluating XLM-R on the linguist label of CoLAC, with other languages' validation samples kept the same as Table 4. The results suggest that from the perspective of cross-lingual transfer, label0 (crowd label) of CoLAC has higher quality then label1 (linguist label).

D Additional Results

D.1 Bilingual training

Apart from the multilingual training in §6, we also perform bilingual training with MELA, where the fine-tuning data come from two languages, each with 250 randomly examples.

Results are shown in Table 10. We find from the last column that English is most helpful when transferring to other languages, which is not surprising since it makes up of the largest portion in the model's pretraining data. From the last row, on the other hand, we find that French benefits the least from other languages, which is consistent with the results in Figure 1.

D.2 ScaLA

As we noted in §2, Nielsen (2023) recently introduce ScaLA, an automatically constructed linguistic acceptability dataset covering six Scandinavian languages. Here we extend our experiments by evaluating the transfer between MELA and ScaLA with XLM-R. We also evaluate BLOOMZ and the two instruction finetuned mT5 on ScaLA for reference. For finetuning XLM-R, we use the full training set of ScaLA, but discard sentences with more than 100 tokens to avoid running out of mem-

CoLAC label	en	zh	it	ru	de	fr	es	ja	ar	is	avg
label1	45.72	52.71	23.18	22.80	21.31	17.61	29.01	31.48	22.16	20.57	28.65
label0	39.56	36.59	19.47	22.33	16.29	17.84	20.99	28.91	10.19	14.89	22.71

Table 9: Performance of XLM-R when trained and evaluated on label1 (first row, same as Table 4) and label0 (second row) of CoLAC. The validation sets of other languages are kept fixed. We note that the columns "zh" and "avg" are not directly comparable between the two rows, since the Chinese validation set are evaluated with different labels. However, of the other nine languages eight have higher performance on the first row than the second row. French is the only exception.

train 1 / eval $\rightarrow$ train 2 $\downarrow$	en	de	is	it	fr	es	ru	zh	ja	ar	avg
in-lang.	46.37	34.59	24.39	17.37	31.77	30.44	24.86	38.81	35.06	36.19	31.99
en	0.00	**-8.30**	-3.67	**4.36**	-9.38	-3.09	-1.50	-2.08	**-3.06**	-9.38	**-3.61**
de	-5.44	0.00	**-0.31**	0.10	-13.44	-8.04	-4.72	**-0.49**	-7.42	**-7.94**	-4.77
is	-9.70	-13.82	0.00	-1.18	-13.74	-9.30	-5.10	-2.16	-7.53	-10.11	-7.26
it	-10.40	-10.30	-2.94	0.00	-11.92	-7.52	-4.70	-0.57	-6.04	-9.79	-6.42
fr	-7.00	-9.94	-6.08	-2.63	0.00	-7.48	-5.15	-4.29	-4.34	-11.09	-5.80
es	**-0.49**	-10.61	-4.15	1.21	-11.33	0.00	-4.47	4.86	-3.73	-8.61	-3.73
ru	-4.81	-10.14	-3.61	-1.40	**-7.81**	-9.92	0.00	-0.63	-5.80	-9.31	-5.34
zh	-3.32	-10.30	-2.16	-0.68	-9.31	**-2.20**	**-0.02**	0.00	-3.81	-8.65	-4.05
ja	-8.83	-11.62	-5.57	1.11	-13.45	-8.19	-2.61	-1.01	0.00	-9.42	-5.96
ar	-9.12	-10.15	-4.34	-3.87	-13.38	-9.72	-6.08	-4.32	-4.71	0.00	-6.57
avg	-5.91	-9.52	-3.28	**-0.30**	-10.38	-6.55	-3.44	-1.07	-4.64	-8.43	

Table 10: Bilingual fine-tuning results. The first row (in-lang.) reports absolute MCC, while the rest report relative MCC w.r.t. the first row. Each cell indicates the result of fine-tuning on 2×250 examples in two languages (train1 and train2) and evaluating on train1. Diagonal cells show results when fine-tuning on 500 samples from a single language, and evaluating on this language.

ory[6].

The results are presented in Table 11. We make several observations:

- Transferring both from MELA to ScaLA and from ScaLA to MELA leads to notable performance drop, even for Icelandic, which is both in MELA and ScaLA. We attribute this to the fact that MELA and ScaLA are constructed differently: the negative sentences in MELA are written by expert linguists, while the negative samples in ScaLA are generated automatically.

- The relative performance of the three instruction finetuned models are consistent between MELA and ScaLA. mTk performance better than mT0 and BLOOMZ, and prompting with in-language examples leads to higher performance than prompting with English examples. BLOOMZ performs only at chance level on ScaLA, since these languages are not covered in its pretraining data.

- Finetuning experiments suggest that the automatically constructed negative samples in ScaLA are easier to distinguish than hand-written negative samples in MELA, as indicated by the much higher MCC score on ScaLA. Few-shot evaluation of instruction finetuned models, however, obtains lower MCC scores on ScaLA than MELA, suggesting that patterns easily captured by finetuning is not necessarily easy to perceive by LLMs in the few-shot setting.

[6] Due to the difference in data source, the average sentence length of ScaLA is significantly longer then MELA.

E Edge Probing Details

Probing Classifier We follow the same architecture of probing classifier as (Tenney et al., 2019b). We extract contextual representations from each layer of XLM-R (including the embedding layer),

			MELA											ScaLA						
model	train data	examples	en	zh	it	ru	de	fr	es	ja	ar	is	avg	da	fo	is	nb	nn	sv	avg
XLM-R	MELA	-	71.6	56.9	57.3	48.9	35.8	24.2	43.6	43.6	31.8	34.7	44.8	47.2	15.6	33.0	57.5	35.9	56.5	41.0
XLM-R	ScaLA	-	51.02	44.43	20.99	23.96	24.92	9.81	33.23	24.64	8.58	25.95	26.75	86.8	51.0	86.1	88.3	78.6	84.4	79.2
XLM-R	MELA + ScaLA	-	70.63	57.84	53.78	49.80	31.91	20.89	44.87	41.24	24.61	33.20	42.88	84.5	46.6	85.4	87.5	77.5	85.2	77.8
BLOOMZ 7B^0	-	-	-2.3	10.0	-1.3	-1.6	-0.9	-3.2	-0.9	-1.9	7.5	3.5	0.9	0.0	0.0	0.0	0.0	0.0	3.6	0.6
BLOOMZ 7B^2	-	in-lang.	7.7	17.6	4.9	-0.2	-0.1	-0.3	7.0	3.8	-1.9	-1.9	3.7	-0.8	2.3	-1.7	2.4	-0.8	0.0	0.3
BLOOMZ 7B^2	-	en	7.7	12.5	2.7	-0.5	3.9	-1.5	6.3	2.3	-2.4	-3.9	2.7	5.6	1.6	-2.3	-1.7	0.0	1.6	0.8
mT0^0	-	-	7.3	20.1	10.8	1.9	10.3	0.4	9.3	13.7	0.0	5.0	7.9	8.6	3.3	7.0	4.7	7.8	5.8	6.2
mTk^2	-	in-lang.	39.1	32.2	18.3	11.8	9.9	13.1	24.4	22.4	12.7	15.5	19.9	20.9	4.2	14.5	12.6	18.1	22.7	15.5
mTk^2	-	en	39.1	31.5	12.1	14.9	16.5	12.8	15.8	15.2	6.3	11.2	17.5	17.1	2.8	6.5	11.7	17.9	17.5	12.2

Table 11: Additional results of finetuned XLM-R and few-shot evaluation of BLOOMZ, mT0, mTk on ScaLA.

and get the scalar mixed representations (in 1,024-dim), see Equation (1) in (Tenney et al., 2019a). Then, the representations are projected in 512-dim with a CNN module. For two-span prediction, we concatenate representations of two spans into a 1,024-dim tensor. We pass the span representations to the probing classifier, which is a two-layer MLP (hidden state dimension is set to 512).

Probing Dataset For part-of-speech tagging and dependency labeling, we use PUD (parallel sentences in all four languages) in UD V2.13. For the other four tasks in OntoNotes 5.0, we down sample sentences to 2k. All dataset all split into train, development and test sets in a ration of 7:1.5:1.5. For each sentence, there might be multiple labels, so we present the numbers of sentences, words and labels in Table 12.

Training We train classifiers for all probing tasks with Adam optimizer at a starting learning rate of 5e-4 for 3,000 training steps with the batch size of 32, and evaluate on the development set every 50 traing steps, halving the learning rate if no improvement is seen in 5 evaluation during training.

Task	$\|L\|$	Sentences	Words	Total Labels
Part-of-speech	17	0.7k / 0.15k / 0.15k	14.7k / 3.2k / 3.3k	14.7k / 3.2k / 3.3k
Dependencies	36	0.7k / 0.15k / 0.15k	14.7k / 3.2k / 3.3k	14.7k / 3.2k / 3.3k
Constituencies	78	1.4k / 0.3k / 0.3k	27.0k / 5.9k / 5.7k	51.1k / 11.1k / 10.7k
Named Entities	18	1.4k / 0.3k / 0.3k	34.6k / 7.3k / 7.4k	3.7k / 0.8k / 0.7k
Semantic Roles	2	1.4k / 0.3k / 0.3k	29.9k / 6.4k / 6.6k	7.3k / 1.5k / 1.6k
Co-reference	66	1.4k / 0.3k / 0.3k	35.4k / 8.1k / 7.5k	3.6k / 0.8k / 0.7k

Table 12: The summary statistics for each split and for each English probing task.

Prompt	MCC
Determine if the following sentence is acceptable or not. Answer acceptable or unacceptable. {sent}	3.92
Determine if the following sentence is acceptable or not. Answer ‘acceptable’ or ‘unacceptable’. {sent}	4.78
Determine if the following sentence is acceptable or not. A sentence which is grammatically correct, has a naturalistic text, is written by a native speaker and which minimizes superfluous content is acceptable, otherwise unacceptable. {sent}	2.37
Determine if the following sentence is acceptable or not. A sentence which is grammatically correct, has a naturalistic text, is written by a native speaker and which minimizes superfluous content is acceptable, otherwise unacceptable. Answer acceptable or unacceptable. {sent}	0.56
Determine if the following sentence is acceptable or not. A sentence which is grammatically correct, has a naturalistic text, is written by a native speaker and which minimizes superfluous content is acceptable, otherwise unacceptable. Answer ‘acceptable’ or ‘unacceptable’. {sent}	3.93
{sent} Question: Is this sentence linguistically acceptable? Answer acceptable or unacceptable.	8.89
{sent} Question: Is this sentence linguistically acceptable? Answer ‘acceptable’ or ‘unacceptable’.	9.23
{sent} Question: Is this sentence linguistically acceptable? A sentence is acceptable if it is grammatically correct and has a naturalistic text. Answer acceptable or unacceptable.	5.30
{sent} Question: Is this sentence linguistically acceptable? A sentence is acceptable if it is grammatically correct and has a naturalistic text. Answer ‘acceptable’ or ‘unacceptable’.	6.47

Table 13: The performance of mT0 on 2560 MELA training samples (256 samples per language) with various prompts.