MELA: Multilingual Evaluation of Linguistic Acceptability

Ziyin Zhang[†*], **Yikang Liu**[‡*], **Weifang Huang**[‡], **Junyu Mao**[◇], **Rui Wang**[†], **Hai Hu**[‡]
[†] Dept. of Computer Science and Engineering, Shanghai Jiao Tong University
[‡] School of Foreign Languages, Shanghai Jiao Tong University
[◇] School of Arabic Studies, Beijing Foreign Studies University
{daenerystargaryen;yikangliu;huangweifang;wangrui12;hu.hai}@sjtu.edu.cn
maojunyu@bfsu.edu.cn

Abstract

In this work, we present the largest benchmark to date on linguistic acceptability: Multilingual Evaluation of Linguistic Acceptability—MELA, with 46K 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 conduct probing experiments with fine-tuned XLM-R to explore the process of syntax capability acquisition. Our results show that GPT-4o exhibits a strong multilingual ability, outperforming fine-tuned XLM-R, while open-source multilingual models lag behind by a noticeable gap. Cross-lingual transfer experiments show that transfer in acceptability judgment is non-trivial: 500 Icelandic fine-tuning examples lead to 23 MCC performance in a completely unrelated language—Chinese. Results of our probing experiments indicate that training on MELA improves the performance of XLM-R on syntax-related tasks.

https://github.com/sjtu-compling/MELA

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, the ability to tell well-formed sentences from ill-formed ones is one of the first that a good language model should have.

Many corpora and benchmarks have been built to evaluate language models' syntactic ability, 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., 2020b).

Recently, there has been growing interest in expanding 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 almost no multilingual benchmarks in this area that can be used to systematically test such abilities of multilingual models.

On the other hand, recently introduced 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., 2021a; Srivastava et al., 2022), math reasoning (Cobbe et al., 2021; Hendrycks et al., 2021b), and code generation (Chen et al., 2021; Austin et al., 2021; Zhang et al., 2023). Few studies, however, have investigated these models from a more linguistics-oriented aspect.

To address these gaps, we introduce MELA—Multilingual Evaluation of Linguistic Acceptability, the first large-scale multilingual acceptability benchmark with 46k examples covering 10 languages from a diverse set of language families. 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 frac-

[*] First two authors contributed equally to this work. Corresponding authors: Rui Wang and Hai Hu.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2658–2674
August 11-16, 2024 ©2024 Association for Computational Linguistics

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.

tion of Russian sentences from Mikhailov et al. (2022).

We propose three possible usage cases for MELA, and make preliminary explorations in this paper:

Benchmarking We benchmark various multilingual language models (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-4o (OpenAI, 2023).

Cross-lingual transfer We train XLM-R (Conneau et al., 2020) on different language combinations, finding non-trivial cross-lingual transfer performance even between unrelated language pairs, despite the vast difference in the basic syntax of the 10 languages in MELA.

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

In the rest of this paper, we first review relevant literature in §2, and then describe how MELA was constructed in §3. Next, we use MELA to benchmark several open-source and close-source LLMs in §4. We investigate cross-lingual transfer and multilingual fine-tuning in §5. Finally, we probe the XLM-Rs trained on MELA for their syntax-related capacity in §6.

2 Related Work

2.1 Linguistic Acceptability

As we mentioned in §1, large-scale linguistic acceptability datasets are currently available for four languages: CoLA for English (Warstadt et al., 2019), ItaCoLA for Italian (Trotta et al., 2021), RuCoLA for Russian (Mikhailov et al., 2022), CoLAC for Chinese (Hu et al., 2023), NoCoLA for Norwegian (Jentoft and Samuel, 2023), and JCoLA (Someya et al., 2024) for Japenese. Sentences from these datasets are taken from academic works by theoretical syntacticians and are therefore annotated by expert linguists.[1]

Another line of work in linguistic acceptability is based on minimal pairs, consisting of two near-identical sentences with minimal differences. Language models are expected to assign a higher probability to the acceptable sentence than the unacceptable one. The minimal pair paradigm is adopted to evaluate specific syntactic issues such as subject-verb agreement (Linzen et al., 2016; Marvin and Linzen, 2018; Varda and Marelli, 2023), reflexive anaphora (Futrell et al., 2019; Hu et al., 2020a), negative polarity licensing (Wilcox et al., 2019; Jumelet and Hupkes, 2018), long-distance dependency (Wilcox et al., 2018; Chowdhury and Zamparelli, 2018), and argument structure (Kann et al., 2019; Tjuatja et al., 2023). Following these works,

[1] CoLAC also comes with an additional set of crowd labels; Unacceptable sentences in NoCoLA are sourced from grammatical mistakes made by Norwegian learners.

comprehensive benchmarks of minimal pairs are constructed in English resources (Warstadt et al., 2020; Hu et al., 2020b), and then expanded into other languages (Xiang et al., 2021; Song et al., 2022; Someya and Oseki, 2023; Nielsen, 2023).

In this work, we follow CoLA in constructing our benchmark as an initial step towards multilingual evaluation in acceptability judgment, as our goal is to have a wide coverage of syntactic phenomena in the languages selected.

2.2 Multilingual Evaluation Benchmarks

XTREME (Hu et al., 2020c) 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 the popularization of multitask instruction finetuning 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). We refer to Qin et al. (2024) for a more comprehensive review of recent multilingual resources.

3 MELA: Multilingual Evaluation of Linguistic Acceptability

MELA consists of more than 46 thousand acceptability samples across 10 languages from a diversity of language families and groups. 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.[2] 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, 2011) (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* (Haider, 2010), *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 labeled 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 is composed of only a phrase, while keeping the judgment.

For Japanese, we remove examples from 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 handbooks are

[2] NoCoLA and JCoLA are not included for they are concurrent with this work.

ISO code	English en	Chinese zh	Italian it	Russian ru	German de	French fr	Spanish es	Japanese ja	Arabic ar	Icelandic is
Train$_{v1.0}$	8551	6072	7801	7869	500	500	500	500	500	500
Dev$_{v1.0}$	527	492	946	1405	272	466	295	580	258	899
Test$_{v1.0}$	516	931	975	2227	273	467	293	581	259	899
Train$_{v1.1}$	8551	6072	7801	7869	-	-	-	-	-	-
Dev$_{v1.1}$	527	492	946	1405	100	100	100	100	100	100
Test$_{v1.1}$	516	931	975	2227	945	1333	988	1561	917	2198
acceptablelen (char)	33.1	10.7	30.0	47.9	39.9	22.9	26.2	14.7	18.1	26.4
len (byte)	34.1	34.3	31.3	95.7	41.5	24.6	28.4	47.0	38.3	31.1
len (token)	10.5	9.5	9.7	15.3	11.4	8.1	8.7	10.9	8.2	9.4

Table 2: Statistics of MELA: train/dev/test splits (in number of sentences), acceptable rate, and average sentence length by characters, bytes, and tokens (using the tokenizer of XLM-R (Conneau et al., 2020)). Subscripts denote the version of data splits: $v1.0$ is used for XLM-R fine-tuning and $v1.1$ is used for LLM zero/few-shot experiments.

overviews of the 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.

3.2 Resulting Corpus and Data Split

The resulting corpus contains more than 46k 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) 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 decide to adopt two splits for two purposes: fine-tuning smaller models such as XLM-R ($v1.0$) and benchmarking LLMs ($v1.1$).[3] For $v1.0$, with the purpose for fine-tuning, we randomly sample 500 sentences from each of these languages to construct a training set, and divide the remaining sentences roughly equally between validation and test sets. For $v1.1$, we reserve 100 samples from each language as the validation set, and keep all the rest of the examples in the test set, thus producing a larger test set which we believe will make the evaluation more stable. See Table 2 for details of the two splits.

3.3 Evaluation Metric

Following previous works in linguistic acceptability, we evaluate the performance on MELA by Matthews Correlation Coefficient (MCC, Matthews, 1975), which is a measure of similarity between binary distributions taking values from -1 to 1 and always yielding 0 for any two uncorrelated distributions, regardless of class imbalance.

3.4 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. First, the task of linguistic acceptability is highly language-dependent, and syntactic structures acceptable in one language may not be acceptable in another, and thus there is no easy way of translating existing corpora into other languages while keeping the target syntactic phenomena. Second, as Clark et al. (2020) and Hu and Kübler (2021) argue, translation introduces artifacts into multilingual benchmarks and often results in translationese.

4 Evaluating LLMs with MELA

In this section, we report the performance of fine-tuned XLM-R and several LLMs, open-source or close-source, on MELA.

4.1 Experimental Settings

To establish a supervised baseline, we use XLM-RoBERTa (Conneau et al., 2020), which is a multilingual version of RoBERTa (Liu et al., 2019) pre-trained on 2.5TB CommonCrawl corpus covering one hundred languages. XLM-R is fine-tuned on the combined training sets of all languages in

[3] Performance of LLMs on two splits of these data are similar (see Table 8).

model	size	examples	en	zh	it	ru	de	fr	es	ja	ar	is	avg
Supervised													
XLM-R	550M	-	60.64	54.94	53.53	49.37	26.72	19.04	34.08	29.32	14.12	35.41	37.72
Open-sourced													
BLOOMZ^0	7.1B	-	4.49	14.13	4.83	4.77	1.63	7.08	10.12	3.27	8.12	0.00	5.85
BLOOMZ^2	7.1B	in-lang.	-1.11	7.65	5.67	5.38	3.90	5.19	6.76	3.83	6.22	-0.35	4.31
BLOOMZ^2	7.1B	en	-1.11	7.90	4.22	4.74	0.96	4.72	8.07	2.45	4.65	-0.09	3.65
mT0^0	13B	-	-5.42	9.68	12.68	5.57	11.33	8.24	2.88	13.20	6.77	1.22	6.62
mT0^2	13B	in-lang.	3.54	8.76	10.93	9.04	5.35	6.66	4.72	12.41	8.95	6.61	7.70
mT0^2	13B	en	3.54	8.06	10.22	10.68	7.17	7.40	6.09	10.46	2.91	4.86	7.14
mTk^0	13B	-	5.25	-1.49	-3.62	5.90	1.25	3.81	5.82	1.04	1.85	2.59	2.24
mTk^2	13B	in-lang.	22.74	8.47	10.24	16.66	11.96	9.28	13.34	12.00	4.87	10.93	12.05
mTk^2	13B	en	22.74	8.36	8.98	15.69	14.54	12.30	9.28	10.92	6.52	5.99	11.53
Baichuan2-Base^0	13B	-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Baichuan2-Base^2	13B	in-lang.	46.11	47.36	24.01	28.84	13.40	17.41	21.95	20.68	13.90	1.81	23.55
Baichuan2-Base^2	13B	en	46.11	35.16	17.84	25.88	6.42	15.95	16.57	13.48	11.41	-3.80	18.50
Baichuan2-Chat^0	13B	-	37.15	33.56	10.08	7.93	-6.49	8.41	18.32	11.15	0.00	2.01	12.21
Baichuan2-Chat^2	13B	in-lang.	41.12	29.25	18.10	19.46	6.46	18.57	20.81	14.18	13.97	-1.51	18.04
Baichuan2-Chat^2	13B	en	41.12	27.02	12.22	14.11	2.80	9.49	14.62	11.40	7.00	-5.03	13.47
Close-sourced													
GPT-3.5^0	-	-	64.60	17.01	14.33	18.05	23.01	31.66	24.35	16.61	9.57	4.69	22.39
GPT-3.5^2	-	in-lang.	64.11	25.32	38.66	21.59	21.62	29.52	44.20	21.48	6.19	9.70	28.24
GPT-3.5^2	-	en	64.11	30.25	25.27	24.91	24.54	29.88	37.75	21.70	6.43	0.56	26.54
GPT-4o^0	-	-	69.05	**62.38**	53.01	55.24	36.61	37.01	58.13	50.32	29.86	40.63	49.22
GPT-4o^2	-	in-lang.	**72.14**	59.01	**54.86**	**59.17**	**39.66**	37.19	**61.36**	**52.03**	**32.38**	**43.64**	**51.14**
GPT-4o^2	-	en	**72.14**	54.77	52.51	52.96	39.20	**39.50**	52.61	47.98	30.08	31.61	47.34

Table 3: Performance of large language models on MELA, in comparison with XLM-R finetuned on MELA training set (all 10 languages). Superscripts denote the number of in-context examples. Note that XLM-R is fine-tuned and evaluated on $v1.0$ while LLMs are evaluated on $v1.1$. However, the performance of LLMs on the two versions is consistent (see Table 8). Thus we report results from different data splits in the same table. We also evaluate mTk with its origin CoLA prompt in its training set (see Table 7).

MELA $v1.0$ with the hyper-parameters described in Appendix A.2.

For open-source LLMs, 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) along with its base model. BLOOMZ is both pre-trained and fine-tuned on 46 languages, which only covers 5 languages in MELA: English, Chinese, French, Spanish, and Arabic[4]. The pre-training corpus of mT5 includes all 10 languages in MELA, but mT0 is fine-tuned on the same instruction dataset as BLOOMZ. mTk's fine-tuning data, on the other hand, covers nine languages in MELA (except for Icelandic) and includes the English CoLA dataset. For Baichuan2, the exact language distribution of pre-training and fine-tuning data is not disclosed. For close-source models, we consider GPT-3.5 and GPT-4o (OpenAI, 2023).

There are several decisions to make when evaluating the above LLMs on MELA: prompt selection and the number of examples in the few-shot scenario. After some pilot experiments, which we describe in Appendix A, we opt to use a binary-choice method with the best performing prompt on the development set, and report the results on the test set in both zero-shot and two-shot scenarios.

4.2 Main results

Results of fine-tuned XLM-R and LLMs evaluated on MELA are given in Table 3. We make the following observations.

Observation 1: GPT-4o exhibits a strong multilingual ability for acceptability judgement. It achieves the best performance on each individual language in MELA, exceeding supervised fine-tuned XLM-R. Its performance is 11 points

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

↓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.

higher than finetuned XLM-R even in a 0-shot setting. There is a bigger gap between GPT-4o and finetuned XLM-R on low-resource languages than high-resource ones, likely due to the small amount of training data (500 examples) for XLM-R. Compared to GPT-4o, GPT-3.5 seems to be more English-centric, with drastic performance drop in non-English languages.

Observation 2: LLMs benefit more from in-language examples in two-shot setting. Our results suggest that prompting with two English examples (most of the time) leads to a lower performance than prompting with in-language examples. On Icelandic, for example, the MCC of the 2-shot setting with English in-context examples is even lower than the 0-shot performance for Baichuan2, GPT-3.5, and GPT-4o.

Observation 3: Baichuan2-Base requires in-context examples As shown in Table 3, under zero-shot setting, Baichuan2-Base shows random performance[5], while its Chat model exhibits non-trivial performance, even for languages such as Spanish. The Base model benefits more from in-context learning examples though, surpassing the Chat model in two-shot settings.

5 Cross-lingual Transfer and Multilingual Fine-tuning

In this section, we investigate cross-lingual transfer and multilingual fine-tuning of linguistic acceptability with XLM-R. All training and evaluation are done on MELA $v1.0$ as it requires a training set.

5.1 Experimental Settings

Cross-lingual Transfer To observe the transfer of acceptability judgements across languages, we fine-tune XLM-R on one language, and evaluate on all 10 development sets. We report the median MCC of seven runs for all results to mitigate inter-run variance.[6]

Multilingual Fine-tuning We downsample sentences in each language to the same number, and fine-tune XLM-R in three settings: (1) in-language, where the fine-tuning and evaluation languages are the same; (2) all-language, where the model is fine-tuned on a mixture of data containing an equal number of sentences from ten languages; and (3) all-but-in-language, where the model is fine-tuned on a mixture of data containing an equal number of sentences from nine languages, except the one being evaluated on.

[5] Baichuan2-Base always chooses "B. Unacceptable" as the answer, under all prompts we tested.

[6] Training details can be found in Appendix A.2.

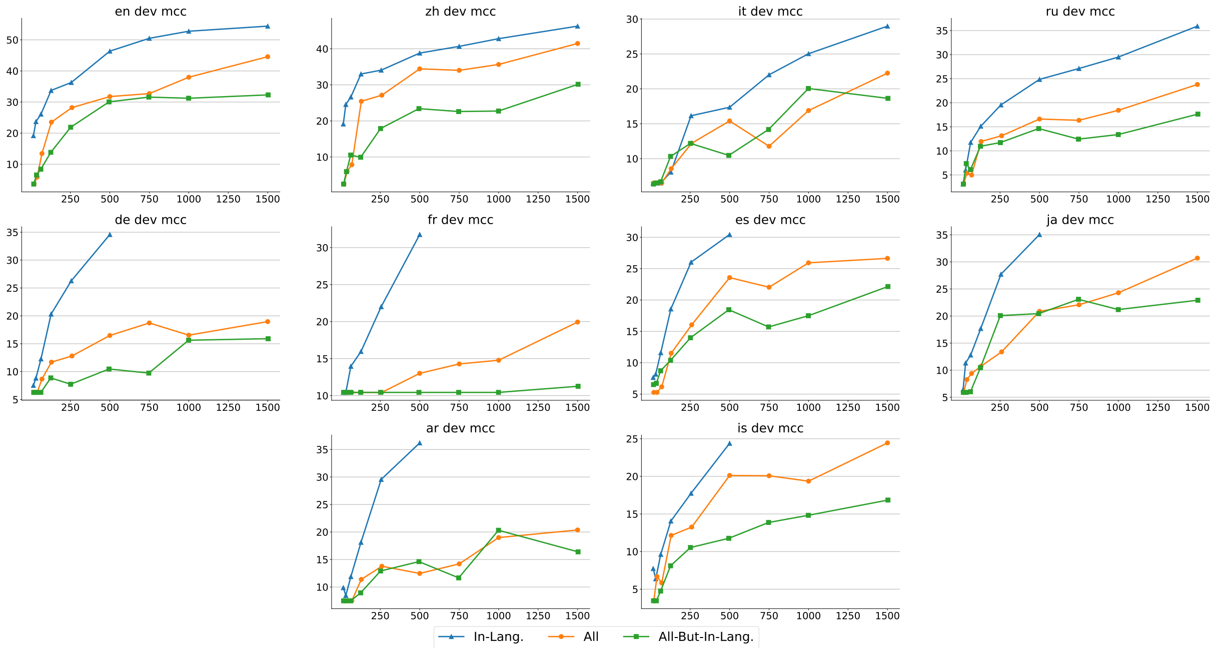

Figure 1: Performance of XLM-R when fine-tuned on different languages. The horizontal axis indicates the number of training samples. For example, for “all” curves, the point at 500 indicates the model is trained on 500 sentences, with 50 from each language. For “All-but-in-lang.” curves, the point at 495 indicates the model is trained on 495 sentences, with 55 from each of the nine languages except the one being evaluated on.

5.2 Results

Results for cross-lingual transfer and multilingual finetuning of XLM-R are reported in Table 4 and Figure 1. We make the following observations.

Observation 1: Cross-lingual transfer is non-trivial. In Table 4, we see that all numbers are (much) greater than 0, suggesting that transferring from language A to language B is possible, even for acceptability judgment tasks. For instance, fine-tuning XLM-R on 500 Icelandic examples results in 23.16 MCC for a completely unrelated language, Mandarin Chinese. A similar conclusion can be drawn from Figure 1, where the green line, which has no in-language training data, demonstrates an increasing trend for all languages, except for French, which plateaus at around 10 MCC.

Observation 2: Size of training set matters, but not always. The overall performance when high-resource languages are used as training data (>6k training examples), 29.69 MCC, is higher than when low-resource ones are used (=500 examples), 18.18 MCC, as shown in the last block of Table 4. However, it must be pointed out that sometimes a (14 times) larger training set does not lead to better performance. For instance, in the second column of Table 4, when evaluated on Chinese, 500 examples of German or Spanish achieve roughly 37 MCC, which is on par with having more than 7,000 training examples for Italian and Russian, with 34.86 and 39.77 MCC respectively. Similarly, when evaluated on Icelandic, 500 German examples again demonstrate a performance on par with many thousands of Chinese, Italian or Russian examples (second to last column of Table 4). Thus the transferring performance between two languages seems to be a result of both the language pair in question as well as the size of the training set.

Observation 3: Among low-resource languages, Arabic training data has the lowest average performance (Table 4). This is likely due to the fact that Arabic is from a different language family from all other nine languages. From the last column of Table 4, we observe that German and Spanish training data have the best performance, likely because MELA has three Germanic languages and three Romance languages, which may make cross-lingual transfer easier among these cognate languages.

Observation 4: For Italian, Japanese and Arabic, during multi-task training, adding in-language data does not affect performance much. From Figure 1, we see that the green and orange lines cross for these three languages, suggesting that *when training with mixed-language* *data*, in-language examples may not be very critical for these languages.

Task	base	en	it	ru	zh
pos	92.87	+0.90	+0.60	+0.30	+1.08
dep	89.41	+0.93	+0.72	+0.51	+0.45
const	78.54	+0.56	−0.10	+0.72	+0.42
name	93.49	+0.74	−0.15	+1.04	+0.59
srl	77.93	+4.41	+2.07	+3.31	+2.35
coref	83.84	+1.71	+0.28	+0.14	+0.69
avg	86.01	+1.54	+0.57	+1.00	+0.93

Table 5: We report the F1 score of XLM-R$_{base}$ (base) on each probing task and the differences between the probing results between MELA-fine-tuned XLM-Rs (en, it, ru and zh) and the base model.

6 Edge Probing

In this section, we adopt edge probing (Tenney et al., 2019a,b) to explore whether fine-tuning on acceptability judgment tasks injects syntax-related information into the pre-trained XLM-R.

6.1 Experimental settings

Edge probing focuses on structural labeling tasks in the form of span labeling. We choose following tasks: 1) part-of-speech tagging, 2) dependency labeling, 3) constituency labeling, 4) named entity labeling, 5) semantic role labeling, and 6) coreference.[7] Take dependency labeling as an example, representations of a dependent and its head, encoded by an XLM-RoBERTa, are used to train a probe classifier to predict the dependency relation between the two words.

We hypothesize that training on MELA can improve the performance of XLM-R on the syntax-related probing tasks above, and design the following experiments.

Experiment 1 We train probing classifiers using span representations from XLM-Rs on English probing tasks. We set the pre-trained M$_{base}$ (pre-trained XLM-R) as the control group, and the other four MELA-fine-tuned M_{mela}^{lang} as the test group, where $lang$ specifies the training data of which language from MELA were used for fine-tuning.

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

6.2 Results

In Experiment 1 we train probing classifiers using representations from different XLM-R variants, some of which have been fine-tuned on MELA while the base model has not. Results in Table 5 show that the average performance of XLM-R$_{base}$ on the six probing tasks is the lowest across the six edge probing tasks (see the last row). We further observe from Table 5 that the semantic role labeling task benefits most from MELA-fine-tuning.

In Experiment 2, the performances of probing tasks are evaluated in the cross-lingual transfer setting. We compare the the pre-trained XLM-R model (M$_{base}$) and XLM-Rs fine-tuned on the linguistic acceptability judgement task of a specific language (M_{mela}^{lang}) (see Table 6). The results indicate that the probing classifiers trained on span representations from fine-tuned XLM-R models achieve better performance than the base model. Fine-tuning on one language of MELA helps the model transfer to that language, and more often than not other languages in part-of-speech tagging and dependency labeling.

MELA-fine-tuned XLM-Rs perform better on syntax-related probing tasks in mono-lingual and cross-lingual settings, supporting our hypothesis.

7 Conclusion

In this work we present MELA, the first multilingual acceptability judgement benchmark covering a diverse set of 10 languages, all annotated by expert linguists. By benchmarking multilingual LLMs on MELA and fine-tuning XLM-R in different cross-lingual settings, we find that (1) GPT-4o ourperforms supervised XLM-R, especially on low-resource languages, that (2) in-language data is crucial for few-shot evaluation and that (3) cross-lingual transfer is non-trivial for all language pairs in supervised fine-tuning. We probe MELA-fine-tuned XLM-R for the syntax information encoded, finding that training on MELA improves the performance on syntax-related probing tasks, which indicates that language models acquire syntactic knowledge during training on linguistic acceptability judgements.

[7] Tasks 1-2 are from UD (De Marneffe et al., 2021); Tasks 3-6 are from OntoNotes (Weischedel et al., 2013).

Probing task		Part-of-speech tagging					Depedency labeling				
↓eval / train→		en	it	ru	zh	**avg**	en	it	ru	zh	**avg**
en	M_{base}	92.87	75.77	65.63	43.33	69.40	89.41	74.99	60.67	40.05	66.28
	M_{mela}^{en}	93.77	81.43	68.22	44.66	**72.02**	90.34	77.40	61.84	45.44	**68.76**
it	M_{base}	83.26	94.61	66.90	38.73	70.88	78.17	91.50	60.65	32.35	65.67
	M_{mela}^{it}	85.60	95.71	63.73	39.70	**71.19**	83.56	92.46	62.85	37.31	**69.05**
ru	M_{base}	82.97	79.90	95.53	53.18	77.90	77.72	78.86	90.90	42.77	72.56
	M_{mela}^{ru}	85.42	81.01	95.43	54.06	**78.98**	80.65	81.27	92.04	46.10	**75.02**
zh	M_{base}	61.19	58.57	64.43	93.88	69.52	50.16	43.42	43.12	86.06	55.69
	M_{mela}^{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. M_{base} refers to the pre-trained XLM-R model; M_{mela}^{en} refers the XLM-R fine-tuned on the English MELA. **Bold** denotes a better performance in average between M_{base} and M_{mela}^{lang}. We conduct a pair comparison between M_{base} and M_{mela}^{lang} trained on MELA of one language to investigate whether linguistic acceptability helps the cross-lingual transfer in the above two probing tasks. For each cell, probing classifiers are trained on span representations in the language denoted in the second row, which are encoded by the model denoted in the second column, and evaluated on the probing tasks in the same language on which XLM-R is fine-tuned (the first column).

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 benchmarking LLMs and providing a data resource for cross-lingual research in computational linguistics. We leave the exploration of other use cases of MELA to future work.

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 research concerning the acquisition and evaluation of linguistic capabilities (of either humans or language models), and should not be interpreted otherwise. For individual sentences in MELA, the copyright (where applicable) remains with the original authors or publishers. We ask researchers who use MELA to also cite the original source, i.e., CoLA (Warstadt et al., 2019), ItaCoLA (Trotta et al., 2021), RuCoLA (Mikhailov et al., 2022) and CoLAC (Hu et al., 2023).

Acknowledgments

We thank Bingjie Mao, Jinchi Jiang, Wen Shi, Jialin Guo, Chenhui Liu and Licen Liu for their help in data collection. We also appreciate the suggestions and comments from the anonymous reviewers. This study is supported by Shanghai Pujiang Program awarded to Hai Hu (22PJC063). Ziyin Zhang and Rui Wang are partially supported by the National Natural Science Foundation of China (62176153) and the Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102, as the MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University).

References

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

Jacob Austin, Augustus Odena, Maxwell I. Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie J. Cai, Michael Terry, Quoc V. Le, and Charles Sutton. 2021. Program synthesis with large language models. *CoRR*, abs/2108.07732.

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

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harrison Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code. *CoRR*, abs/2107.03374.

Shammur Absar Chowdhury and Roberto Zamparelli. 2018. Rnn simulations of grammaticality judgments on long-distance dependencies. In *Proceedings of the 27th international conference on computational linguistics*, pages 133–144.

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.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. 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 32–42.

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.

Hubert Haider. 2010. *The Syntax of German*. Cambridge Syntax Guides. Cambridge University Press.

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

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Hai Hu and Sandra Kübler. 2021. Investigating translated Chinese and its variants using machine learning. *Natural Language Engineering*, 27(3):339–372.

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, Sherry Yong Chen, and Roger Levy. 2020a. A closer look at the performance of neural language models on reflexive anaphor licensing. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 323–333.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020b. 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. 2020c. XTREME: A massively multilingual multitask benchmark for evaluating cross-lingual generalization. *CoRR*, abs/2003.11080.

Matias Jentoft and David Samuel. 2023. Nocola: The norwegian corpus of linguistic acceptability. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 610–617.

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.

Katharina Kann, Alex Warstadt, Adina Williams, and Samuel Bowman. 2019. Verb argument structure alternations in word and sentence embeddings. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, pages 287–297.

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.

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.

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.

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.

B.W. Matthews. 1975. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442–451.

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.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. 2024. Multilingual large language model: A survey of resources, taxonomy and frontiers. *CoRR*, abs/2404.04925.

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.

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.

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

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.

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.

Lindia Tjuatja, Emmy Liu, Lori Levin, and Graham Neubig. 2023. Syntax and semantics meet in the "middle": Probing the syntax-semantics interface of lms through agentivity. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (* SEM 2023)*, pages 149–164.

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.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do rnn language models learn about filler–gap dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 211–221.

Ethan Wilcox, Peng Qian, Richard Futrell, Miguel Ballesteros, and Roger Levy. 2019. Structural supervision improves learning of non-local grammatical dependencies. *arXiv preprint arXiv:1903.00943*.

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 Benchmark Details

In this section, we provide details about how we make decisions to benchmark LLMs, which includes prompt selection, and the number of examples in the few-shot scenario, along with details of fine-tuning XLM-R.

A.1 Large Language Models

Following MMLU (Hendrycks et al., 2021a), we evaluate MELA in the multiple-choice format. We use prompts that end with “Answer:”. Models are required to produce probabilities for index tokens “A” and “B”. The index token with a higher probability is regarded as the decision of models. We tune prompts with open-sourced LLMs on the validation set in our pilot experiment.

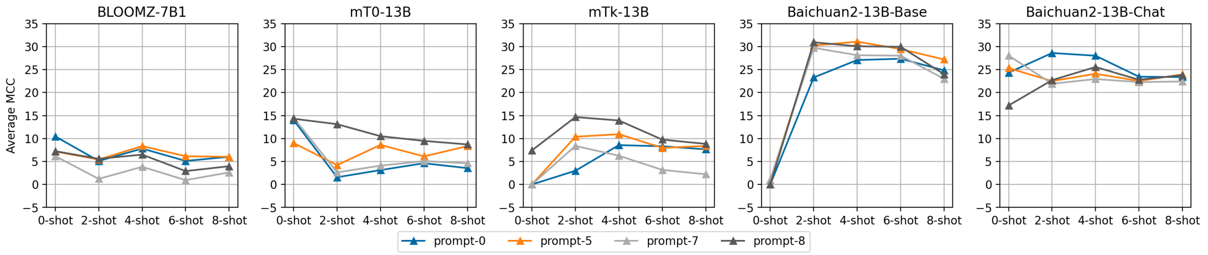

Figure 2: Prompt selection results. We experiment with 4 prompts adapted from previous CoLA-prompts from `promptsource` and `lm-evaluation-harness`.

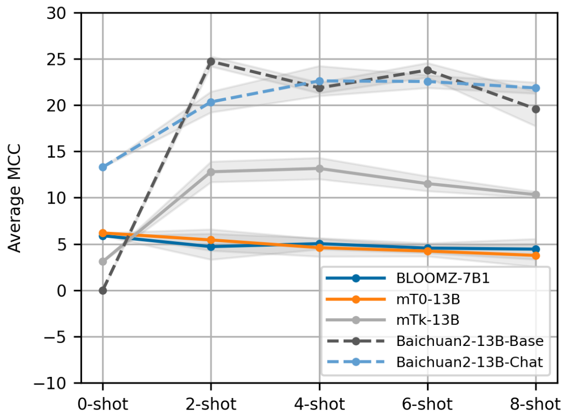

Figure 3: Average performance across languages with different numbers of in-context examples. We average the MCC and report standard deviations over 5 seeds. Gray bands denote standard deviations.

We first experiment on 1,000 samples of MELA (50 samples per label per language), using 4 different prompts in 0/2/4/6/8-shot (equal number of positive and negative examples) scenarios with in-language examples provided. The prompt with the highest average MCC is selected to evaluate LLMs on the whole MELA test set (see Figure 2 and 3). The results indicate that 1) `prompt-8` is better than others and 2) models no longer improve with more than 2 in-context examples.

Therefore, we carry out formal experiments with `prompt-8` (see Figure 4) for both open-sourced LLMs and GPT models in zero and two-shot scenarios (see §4).

Note that mTk includes the 2-shot CoLA task in its training data. We also reuse the prompt for Supernatural Instruction task 616[8]. We compare the results of mTk's origin CoLA prompt (see Figure 4) and our multiple-choice prompt (see Table 7).

A.2 XLM-R Fine-tuning Details

For experiments concerning XLM-R in §4 and §5, 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 5k 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.

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 languages (see Figure 5), 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 §5, each run also selects a different subset of training data.

A.3 Comparison Between Two Splits

In §3, we have to split the MELA datasets into train, development, and test sets to fine-tune XLM-

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

```
# 0−shot multiple−choice prompt
Determine whether the following sentence(s) violate certain linguistic constraints. If yes, then it is "
unacceptable"; otherwise, "acceptable".

Sentence: {target sentence}.
Determine whether this sentence is acceptable or unacceptable?
A. Acceptable
B. Unacceptable
Answer:
```

```
# 2−shot multiple−choice prompt
Determine whether the following sentence(s) violate certain linguistic constraints. If yes, then it is "
unacceptable"; otherwise, "acceptable".

Sentence: {positive example1}.
Determine whether this sentence is acceptable or unacceptable?
A. Acceptable
B. Unacceptable
Answer: A

Sentence: {negative example2}.
Determine whether this sentence is acceptable or unacceptable?
A. Acceptable
B. Unacceptable
Answer: B

Sentence: {target sentence}.
Determine whether this sentence is acceptable or unacceptable?
A. Acceptable
B. Unacceptable
Answer:
```

```
# 2−shot mTk origin prompt
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: {positive example1}
        output: acceptable
Positive Example 2−
        input: {negative example2}
        output: unacceptable
Now complete the following example−
        input: {target sentence}
        output:
```

Figure 4: Prompt used for evaluating LLMs.

prompt	model	size	examples	en	zh	it	ru	de	fr	es	ja	ar	is	avg
ours	mTk^2	13B	in-lang.	22.74	8.47	10.24	16.66	11.96	9.28	13.34	12.00	4.87	10.93	12.05
	mTk^2	13B	en	22.74	8.36	8.98	15.69	14.54	12.30	9.28	10.92	6.52	5.99	11.53
origin	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

Table 7: We compare the results of mTk on our prompt and the origin prompt in its training data.

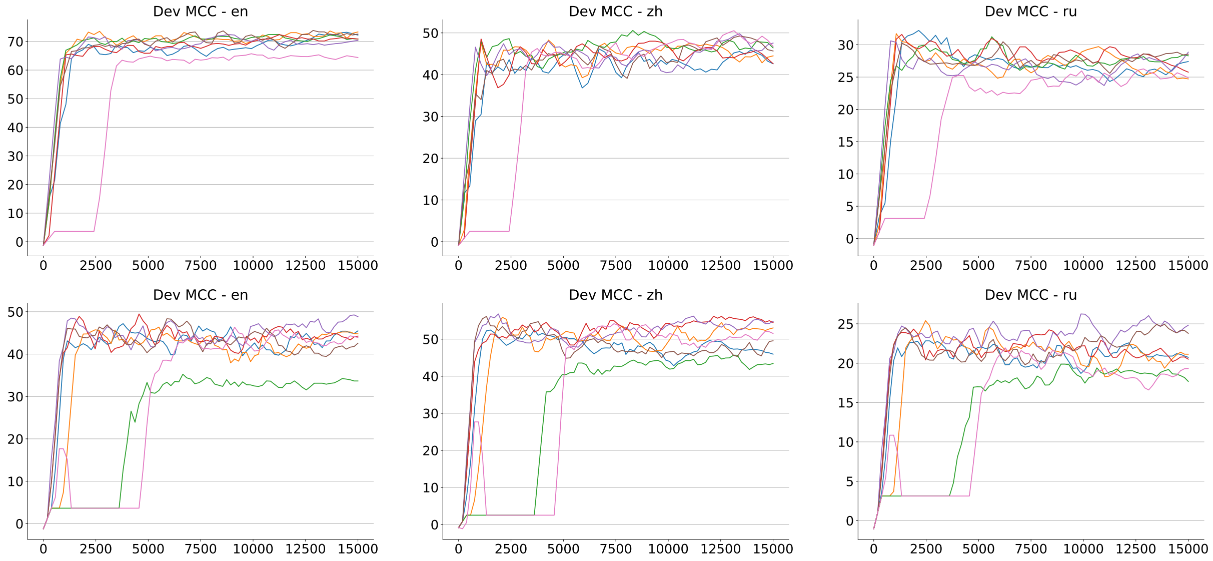

Figure 5: Interrun variance when finetuning XLM-R on English (first row) and Chinese (second row) training data. Each subfigure plots the validation MCC of seven runs with different random seeds on one language. After taking the median of these seven runs, this variance is mitigated to a large extent.

R. However, considering training data is no longer necessary for LLM evaluation. Therefore, we decide to make two different data splits. On the one hand, we want a fair comparison between supervised fine-tuning XLM-R and zero/few-shot LLMs. On the other hand, we want MELA to better fit the recent paradigm of LLM evaluation. In this case, we provide a comparison between the performance of LLMs on the two different versions of data (see Table 8). The results of the two versions are similar, by which we make it comparable between fine-tuned XLM-R on `v1.0`-test set and LLMs on `v1.1`-test set.

B 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), 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 downsample sentences to 2k. All datasets are split into train, development and test sets in a ratio 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 9.

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

model		BLOOMZ			mT0			mTk			Baichuan2-Base			Baichuan2-Chat		
n-shot		0-shot	2-shot	2-shot	0-shot	2-shot	2-shot	0-shot	2-shot	2-shot	0-shot	2-shot	2-shot	0-shot	2-shot	2-shot
ex. lang.		-	in-lang.	en	-	in-lang.	en	-	in-lang.	en	-	in-lang.	en	-	in-lang.	en
de$_{v1.0}$	273	-12.22	0.46	1.71	7.00	8.36	6.83	16.63	10.70	10.70	0.00	7.98	2.77	0.00	0.86	-3.05
de$_{v1.1}$	945	1.63	3.90	0.96	11.33	5.35	7.17	1.25	11.96	14.54	0.00	13.40	6.42	-6.49	6.46	2.80
fr$_{v1.0}$	467	11.32	2.83	4.84	5.83	4.91	8.55	2.14	8.95	7.71	0.00	17.82	13.87	9.58	16.99	10.03
fr$_{v1.1}$	1333	7.08	5.19	4.72	8.24	6.66	7.40	3.81	9.28	12.30	0.00	17.41	15.95	8.41	18.57	9.49
es$_{v1.0}$	293	10.83	9.20	8.38	7.18	11.10	10.44	8.65	21.07	17.67	0.00	24.74	21.19	20.85	24.35	14.29
es$_{v1.1}$	988	10.12	6.76	8.07	2.88	4.72	6.09	5.82	13.34	9.28	0.00	21.95	16.57	18.32	20.81	14.62
ja$_{v1.0}$	581	1.98	-0.31	2.19	12.32	12.58	9.51	-2.80	5.58	5.70	0.00	24.93	16.72	9.06	20.92	14.96
ja$_{v1.1}$	1561	3.27	3.83	2.45	13.20	12.41	10.46	1.04	12.00	10.92	0.00	20.68	13.48	11.15	14.18	11.40
ar$_{v1.0}$	259	5.53	6.83	7.43	2.95	5.66	-3.59	10.60	3.48	6.77	0.00	10.75	6.79	0.00	12.66	0.52
ar$_{v1.1}$	917	8.12	6.22	4.65	6.77	8.95	2.91	1.85	4.87	6.52	0.00	13.90	11.41	0.00	13.97	7.00
is$_{v1.0}$	899	0.00	0.91	0.95	3.85	4.08	4.18	7.74	14.04	6.81	0.00	3.99	0.75	2.68	3.14	-3.08
is$_{v1.1}$	2198	0.00	-0.35	-0.09	1.22	6.61	4.86	2.59	10.93	5.99	0.00	1.81	-3.80	2.01	-1.51	-5.03
avg$_{v1.0}$	-	2.91	3.32	4.25	6.52	7.78	5.99	7.16	10.64	9.23	0.00	15.03	10.35	7.03	13.15	5.61
avg$_{v1.1}$	-	5.04	4.26	3.46	7.27	7.45	6.48	2.73	10.40	9.93	0.00	14.86	10.01	5.57	12.08	6.71

Table 8: Comparison between the performance of open-sourced LLMs on two versions of data splits. We only report results on six low-resource languages since data for the four high-resource languages are the same between the two splits.

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