QFrCoLA: a Quebec-French Corpus of Linguistic Acceptability Judgments

David Beauchemin and **Richard Khoury**
Group for Research in Artificial Intelligence of Laval University (GRAIL)
Université Laval, Québec, Canada
{david.beauchemin, richard.khoury}@ift.ulaval.ca

Abstract

Large language models (LLM) perform outstandingly in various downstream tasks. However, there is limited understanding regarding how these models internalize linguistic knowledge, so various linguistic benchmarks have recently been proposed to facilitate syntactic evaluation of language models (LM) across languages. This paper introduces QFrCoLA (Quebec-French Corpus of Linguistic Acceptability Judgments), a normative binary acceptability judgments dataset comprising 25,153 in-domain and 2,675 out-of-domain sentences. Our study leverages the QFrCoLA dataset and seven other linguistic binary acceptability judgments corpus to benchmark eight LM. The results demonstrate that, on average, fine-tuned Transformer-based LM are strong baselines for most languages and that zero-shot binary classification LLM perform worse than the naive baseline on the task. However, for the QFrCoLA benchmark, on average, a fine-tuned Transformer-based LM outperformed other methods tested. It also shows that pre-trained cross-lingual LLMs selected for our experimentation do not seem to have acquired linguistic judgment capabilities during their pre-training for Quebec French. Finally, our experiment results on QFrCoLA show that our dataset, built from examples that illustrate linguistic norms rather than speakers' feelings, is similar to linguistic acceptability judgment; it is a challenging dataset that can benchmark LM on their linguistic judgment capabilities.

1 Introduction

The introduction of large language models (LLM) (Touvron et al., 2023) and Transformer-based language model (LM) (Vaswani et al., 2017) has led to significant progress in natural language processing (NLP), substantially increasing the performance of most NLP tasks (Zhang et al., 2023). LLMs were initially introduced for English (Kenton and Toutanova, 2019; Brown et al., 2020; Touvron et al., 2023), but many other languages were later introduced, such as Russian (Kuratov and Arkhipov, 2019), French (Martin et al., 2020), and Norwegian (Kummervold et al., 2021). NLP research has approached the competencies evaluation of various natural language tasks of LM with various benchmark corpora such as the English benchmarks GLUE (Wang et al., 2018), SuperGLUE (Wang et al., 2019), and GLGE (Liu et al., 2021) to name a few. These corpora are collections of resources for training, evaluating, and analyzing LM (Gao et al., 2023; Chang et al., 2023). For example, GLUE aims to benchmark an NLP system's capabilities for natural language understanding (NLU) (Wang et al., 2018). At the same time, GLGE focuses on natural language generation (NLG) tasks such as document summarization (Liu et al., 2021).

Recently, much effort has been put into creating linguistic acceptability resources to assess and benchmark LM linguistic competency, where recent NLP research formulate linguistic competency as a binary classification task (Cherniavskii et al., 2022; Proskurina et al., 2023). That is the ability, from a native speaker's perspective, to distinguish the correct form and naturalness of an acceptable sentence from an unacceptable one (Chomsky, 2014). Recently, similar non-English resources have been proposed to answer this question in typologically diverse languages such as Japanese (Someya et al., 2023), Norwegian (Jentoft and Samuel, 2023), and Chinese (Hu et al., 2023). However, the ability of LMs to perform linguistic acceptability judgments in French remains understudied.

To this end, we introduce the **Q**uebec-**Fr**ench **C**orpus **o**f **L**inguistic **A**cceptability Judgments (QFrCoLA), a corpus consisting of 25,153 in-domain and 2,675 out-of-domain normative acceptability judgment sentences, making it the second largest linguistic acceptability resources available in the NLP literature. The main contributions of this work are therefore

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 119–130
November 4-9, 2025

1. The creation and release of QFrCoLA[1], a dataset of normative grammatical and ungrammatical sentences with binary labels;
2. A set of experiments to assess the performance of LM on QFrCoLA;
3. A cross-lingual benchmarking of LM on eight languages, including French, that opens up novel multi-language research perspectives.

It is outlined as follows: first, we study the available linguistic binary acceptability corpus and related binary classification LM research in Section 2. Then, we propose the QFrCoLA in Section 3, and in Section 4 and Section 5 we present a set of experiments aimed at testing the performance of LM binary classifiers on all the linguistic acceptability resource corpora. Finally, in Section 6, we conclude and discuss our future work.

2 Related Work

Linguistic acceptability judgment evaluates one capacity to distinguish the correct form and naturalness of an acceptable sentence from an unacceptable one. For instance, individuals can inherently distinguish between two sentences and identify the one that is more acceptable or natural-sounding. This assessment is the primary behavioural benchmark employed by generative linguists to investigate the underlying structure of human language (Chomsky, 2014). Through benchmarking linguistic acceptability judgments of LLM, one can assess their linguistic robustness.

2.1 Language Model Evaluation

Historically, evaluation of LMs has been conducted using metrics or benchmark corpora (Chang et al., 2023). The first approach relies either on task-agnostic metrics, such as perplexity (Jelinek et al., 1977) which measures the quality of the probability distribution of words in a given corpus by a model, or on task-specific metrics, like the BLEU score that evaluates a model's performance for machine translation (Papineni et al., 2002). The second approach relies on large corpora designed for NLU or NLG downstream tasks. For example, the GLUE benchmark (Wang et al., 2018) is used to assess a model's NLU performance on tasks such as semantic similarity, linguistic acceptability judgment and sentiment analysis. In contrast, GLGE (Liu et al., 2021) evaluates language generation tasks such as summarization and question answering.

[1] https://github.com/GRAAL-Research/QFrCoLA

Acceptable Sentence	Not Acceptable Sentence
The cats annoy Tim.	The cats annoys Tim.

Table 1: Example of a minimal pair (Warstadt and Bowman, 2019).

2.2 Language Model Linguistic Acceptability Judgments Evaluation

Recently, NLP researchers started using linguistic acceptability judgment tasks to assess the robustness of LMs against grammatical errors (Miaschi et al., 2023) and to probe their grammatical knowledge (Choshen et al., 2022; Mikhailov et al., 2022). Two approaches are used to perform this evaluation: minimal pairs and binary classification acceptability judgments (Chang et al., 2023).

In the first approach, a set of minimal pairs of grammatically acceptable and unacceptable sentences, such as the pair illustrated in Table 1, is presented to an LM. By observing which sentences the LM assigns a higher correctness probability to, one can assess which grammatical phenomena it is sensitive to (Warstadt and Bowman, 2019). Corpus such as BLiMP in English (Warstadt and Bowman, 2019) and CLiMP in Chinese (Xiang et al., 2021) have been proposed to enable the evaluation of LM on a wide range of linguistic phenomena.

In the second approach, a set of sentences that are either grammatical or ungrammatical, such as the two shown in Table 2, are provided to an LM which must perform a binary classification (Warstadt et al., 2019). Seven corpora have been proposed to assess LMs' capabilities to discriminate proper grammar from improper in their respective languages: CoLA for English (Warstadt et al., 2019), DaLAJ for Swedish (Volodina et al., 2021), 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 for Japanese (Someya et al., 2023). However, as of yet, no such corpus exists for French.

Typically, the datasets in the second approach comprise sentences collected from syntax textbooks and linguistics journals. These datasets propose "in-domain" train-dev-test splits to train and evaluate machine learning models. CoLA, RuCoLA, CoLAC, and JCoLA corpora also include an "out-of-domain" (OOD) split to assess whether a model suffers from overfitting. However, the definition of OOD varies depending on the corpus. CoLA includes sources of varying degrees

Label	Sentence
0 (Ungrammatical)	Edoardo returned to his last year city
1 (Grammatical)	This woman has impressed me

Table 2: Example sentences from the ItaCoLA dataset (Trotta et al., 2021).

of domain specificity and time period compared to those used for the primary dataset (Warstadt and Bowman, 2019). For RuCoLA, they are sentences generated by an automatic machine translation system and paraphrase generation models and annotated by a human annotator (Mikhailov et al., 2022). While JCoLA comprises sentences from the Journal of East Asian Linguistics, a source with typically more complex linguistic phenomena than the other reference use of the in-domain splits (Someya et al., 2023).

3 QFrCoLA: Quebec-French Corpus of Linguistic Acceptability Judgments

In this work, we introduce the **Q**uebec-**Fr**ench **C**orpus **o**f **L**inguistic **A**cceptability Judgments (QFrCoLA), which will be the first large-scale normative binary linguistic acceptability judgments dataset for the Quebec-French language and the second-largest corpus in any language.

3.1 Sources

QFrCoLA consists of French normative grammatical or ungrammatical sentences taken from two online French sources: the "*Banque de dépannage linguistique*" (BDL) and the *Académie française*. The first source is our "in-domain" Quebec-French sentences for the train-dev-test splits, while the second is our OOD hold-out split. Both sources are publicly available online, and we obtained authorization to publish them under a CC-BY-NC 4.0 license.

3.1.1 In-Domain Source

The BDL is an official online resource created by the "*Office québécois de la langue française*" (OQLF), a provincial government public organization in Canada[2], making it a reliable normative French resource. It is a normative grammatical

[2]The Quebec government created the OQLF to "protect" the French Quebec culture (Molinari, 2008; Bobowska-Nastarzewska, 2009), therefore it can be considered as a "political initiative". (Dahlet, 2010). Consequently, its BDL initiative can be perceived as a biased French grammatical resource. However, the accepted grammar of the BDL is similar to other French native communities such as Belgium and Switzerland (Saint, 2013).

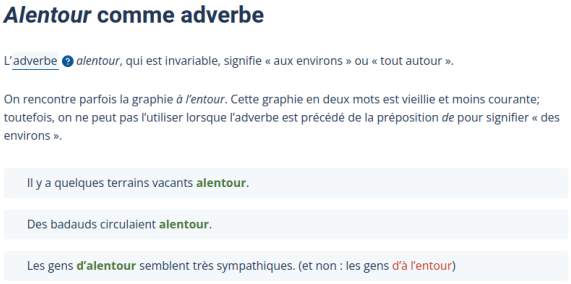

Figure 1: Snipped of the BDL article for the French adverb "*alentour*". The text is in French.

on dit	on ne dit pas
Je pense qu'on a fait un bon match	*Je pense on a fait un bon match*
Je trouve que c'est dur quand même	*Je trouve c'est dur quand même*
Tu crois que le professeur viendra ?	*Tu crois le professeur viendra ?*

Figure 2: Snipped of an *Académie française* article for the "*Omission de la conjonction « que »* (Omission of the conjunction "that")". The text is in French.

resource of 2,667 articles divided into eleven categories, such as "*orthographe*" (spelling), and "*syntaxe*" (syntax). These articles explain various normative linguistic phenomena that the OQLF considers correct or incorrect. It uses examples written by French linguists to illustrate both cases based on linguistic phenomenal observation. For example, the "*adverbes*" (adverbs) category includes an article about the linguistic phenomenon of proper and improper use of the adverb "*alentour*)" (surrounding). Figure 1 displays examples of well-written sentences using the adverb (in green) and an example of an erroneous usage (in red).

3.1.2 Out-Of-Domain Source

Our second source is the *Académie française*, a France-based organization acting as a "society of scholars" in science and literature (Académie française, 2024b). It publishes monthly in their online *La langue française: Dire, Ne pas dire* journal that presents 1,013 articles on normative grammar with examples of proper and improper use of French. These examples are sorted into three categories: "*néologismes and anglicismes*" (neologisms and anglicisms), "*emplois fautifs*" (wrongful employment), and "*extensions de sens abusives*" (abusive extensions of meaning). Figure 2 displays examples of proper (left) and wrongful (right) employments of the conjunction "*que*" (that).

Like CoLA, RuCoLA, CoLAC, and JCoLA, our corpus includes an OOD split using a similar approach as JCoLA and CoLA. Namely, we use a substantially different source to build it. Indeed,

French in Quebec differs from France (Fagyal et al., 2006). For example, the feminization of titles differs between the two; the feminization of *auteur* (author) in Quebec is accepted as *autrice* or *auteure* (OQLF, 2024), while in France it is only accepted as *auteure* (Académie française, 2024a). However, both countries have similar linguistic phenomena, such as syntax and plurals (Dankova, 2017).

3.2 Data Collection

3.2.1 In-Domain

We examined all 2,667 articles and manually extracted 25,153 normative linguistic acceptability judgment sentences. Each sentence was labelled `0` (ungrammatical) or `1` (grammatical) following the BDL green/red colour scheme as illustrated in Figure 1. Furthermore, since the BDL uses a fine-grained category structure to sort various linguistic phenomena, we collected these categories and associated them to labels according to the French linguistic literature (Fagyal et al., 2006; Chesley, 2010; Boivin and Pinsonneault, 2020; Feldhausen and Buchczyk, 2021), and labelled each extracted sentence accordingly. Our linguistic phenomena labels are listed below, and Table 3 presents QFrCoLA statistics for each one, along with an example. Our categories are unevenly distributed, with nearly 43% being in the morphology category. Moreover, the percentage of acceptability labels is also unevenly distributed, ranging from 58.26% to 77.56%. It is due to the nature of our dataset, where the BDL, in many cases, presents proper normative use of French rather than improper use. It is shown for the "anglicism" where nearly every sentence presents a proper and improper case.

3.2.2 Out-Of-Domain

OOD sentences were manually extracted from the journal's 1,013 articles. We extracted 2,675 sentences from those articles and only binary labelled them following the table scheme (left/right) as illustrated in Figure 2. We discuss the dataset statistics in the following section.

3.3 Comparison With Other Similar Corpora

This section compares our corpus with all related ones. Table 4 present in-domain number of sentences, percentage of acceptable sentences and vocabulary size for the train, dev and test sets[3] and for the entire corpus. The total vocabulary sizes were computed using language-specific SpaCy tokenizers (Honnibal et al., 2020) that split each sentence into individual words or punctuation. We can see that QFrCoLA is the second largest corpus in terms of the number of sentences it contains, behind only NoCoLA, and is approximately twice the size of all the other corpora. Moreover, it has a similar frequency of acceptable sentences to the CoLA, CoLAC, and RuCoLA datasets, and like the other corpora, all splits have a similar frequency of acceptable sentences. Finally, we can see that QFrCoLA has the third-largest vocabulary size compared to the other datasets.

Table 5 present, for the OOD split, the number of sentences, vocabulary size and percentage of acceptable sentences of all linguistic corpora with an available OOD split. However, since other corpora do not distribute their hold-out labels, we could not compute the percentage of acceptable sentences. We also note that for JCoLA, the OOD hold-out split was unavailable in their official dataset GitHub repository. Once again, we can see that QFrCoLA is the second largest corpus in terms of number of sentences and vocabulary size, with nearly as many sentences as RuCoLA. Compared to the main QFrCoLA corpus in Table 4, we can see that the OOD split comprises a much less diverse vocabulary, making it well distinct from the other splits. Finally, the OOD hold-out split has a percentage of acceptable sentences nearly 15% lower than the overall corpus, making it more robust to highlight overfitting cases in machine learning models.

4 Experiments

We train and evaluate three fine-tuned approaches and evaluate eight LLMs in a zero-shot binary classification setup. We then benchmark these models against a baseline.

4.1 Evaluation Metrics

Following Warstadt et al. (2019), performance is measured using the accuracy score and Matthews correlation coefficient (MCC) (Matthews, 1975). Accuracy on the dev set is used as the target metric for hyperparameter tuning and early stopping. We

[3]It is worth mentioning that for CoLA, RuCoLA and JCoLA, their in-domain test set labels are not available to reduce the risk of overfitting. Thus, like other related work (Cherniavskii et al., 2022), we use their out-of-domain dev sets as the test sets. Also, CoLAC does not provide an OOD set nor label for their test set. Thus, per the authors' recommendation, the in-domain train and dev set was resampled using a 60-10-30% split with seed 42 to create new splits.

Category	BDL Fine-Grained Categories	# Sen % Acp	Example
Syntax	Agreement violations, corruption of word order, misconstruction of syntactic clauses and phrases, incorrect use of appositions, violations of verb transitivity or argument structure, ellipsis, missing grammatical constituencies or words	5,152 77.24	*Dès son arrivée, on s'empressa de lui poser des questions à propos de son voyage.* (translated) As soon as he arrived, people were quick to ask him questions about his trip. ***Dès en arrivant, on s'empressa de lui poser des questions à propos de son voyage.*** (translated) **As soon as he arrived, they were quick to ask him questions about his trip.**
Morphology	Incorrect derivation or word building, non-existent words	10,642 68.26	*Sa maison est neuve.* (translated) His house is new. ***Sa maison est neuf.*** (translated) **His house is new.**
Semantic	Incorrect use of negation or violates the verb's semantic argument structure	5,442 72.97	*Quand la parade est passée, le vieil homme s'est levé pour aller voir à la fenêtre.* (translated) When the parade was over, the old man got up to look out the window. ***Quand la parade est passée, le vieil homme s'est levé debout pour aller voir à la fenêtre.*** (translated) **When the parade passed, the old man stood up to look out the window.**
Anglicism	Word and syntactical structure borrowed from English grammar	3,917 57.18	*Sauront-Ils répondre aux les besoins de l'enfant?* (translated) Will they be able to meet the child's needs? ***Sauront-Ils rencontrer les besoins de l'enfant?*** (translated) **Will they be able to meet the child's needs?**

Table 3: Number of sentences (# Sen) and the percentage of acceptable sentences (% Acp) per category in QFrCoLA (all three splits), and example of a positive and a negative (**bolded** with error underlined) along with their translation in each category.

	Language	Train			Dev			OOD/Test			Total		
		# Sen	% Acp	Vocab	# Sen	% Acp	Vocab	# Sen	% Acp	Vocab	# Sen	% Acp	Vocab
CoLA (Warstadt et al., 2019)	English	8,551	70.44	5,778	527	69.26	1,375	516	68.60	988	9,594	70.27	6,097
DaLAJ (Volodina et al., 2021)	Swedish	7,682	50.00	6,841	890	50.00	1,799	888	50.00	1,661	9,460	50.00	7,884
ITACoLA (Trotta et al., 2021)	Italian	7,801	84.39	5,825	946	85.41	1,844	1,888	84.21	1,888	9,722	84.47	6,402
RuCoLA (Mikhailov et al., 2022)	Russian	7,869	74.52	19,057	983	74.57	4,140	1,804	63.69	9,353	10,656	72.69	26,382
CoLAC (Hu et al., 2023)	Chinese	4,134	66.09	3,835	460	66.96	1,024	1,970	67.82	2,636	6,564	66.67	4,759
NoCoLA (Jentoft and Samuel, 2023)	Norwegian	116,195	31.46	32,561	14,289	32.59	8,865	14,383	31.58	8,600	144,867	31.58	37,319
JCoLA (Someya et al., 2023)	Japanese	6,919	83.38	3,730	865	83.93	1,483	684	73.28	896	8,469	82.62	4,146
QFrCoLA	French	15,846	69.49	18,350	1,761	69.51	5,369	7,546	69.49	12,690	25,153	69.49	22,131

Table 4: Comparison of QFrCoLA and related corpora for the number of sentences (# Sen), percentage of acceptable sentences (% Acp), and vocabulary size (Vocab). "OOD" stands for "out-of-domain".

	OOD Hold-Out		
	# Sen	Vocab	% Acp
CoLA	533	1035	N/A
RuCoLA	2,789	12,211	N/A
CoLAC	931	1,168	N/A
JCoLA	N/A	N/A	N/A
QFrCoLA	2,675	1,651	53.91

Table 5: Comparison of QFrCoLA with all related corpus with an out-of-domain (OOD) hold-out set for the number of sentences (# Sen), the vocabulary size (Vocab) and the % of acceptable sentences (% Acp).

Language	Model Name
En	bert-base-cased (Kenton and Toutanova, 2019)
SV	bert-base-swedish-cased (Malmsten et al., 2020)
IT	bert-base-Instructtalian-cased (Schweter, 2020)
RU	ruBert-base (Zmitrovich et al., 2023)
ZH	bert-base-chinese (Cui et al., 2021)
NO	nb-bert-base (Kummervold et al., 2021)
JA	bert-base-japanese (Suzuki and Takahashi, 2019)
FR	camembert-base (Martin et al., 2020)

Table 6: Selected pre-trained transformer models per language using ISO-2 letter format.

report the results averaged over ten restarts from different random seeds (i.e. $[42, 43, \cdots, 50, 51]$).

4.2 Models

As our baseline, we selected the trivial approach to always select class `1` (`Baseline`). Namely, this model accuracy equals the percentage of acceptable sentences (% Acp) illustrated in Table 3.

4.2.1 Monolingual Language Model

Monolingual We selected a state-of-the-art (SOTA) pre-trained monolingual LM for each language based on their benchmark performance on various tasks (Chang et al., 2023) as our monolingual baseline (`BERT`). We detail the selected language-specific model name in Table 6.

State-Of-The-Art The SOTA approach to binary linguistic acceptability judgments is the topological data analysis (TDA) proposed by Cherniavskii et al. (2022) (`LA-TDA`). This approach extracts the attention maps of a fine-tuned Transformers-based LM to use as linguistic features to train a binary logistic regression. The authors report that this approach significantly outperformed previous approaches, in-

creasing the MCC score on linguistic acceptability for English, Italian, and Swedish by up to 0.24. In our case, we use the attention maps from the monolingual fine-tuned models. We selected this approach since it is the SOTA approach.

4.3 Cross-Lingual Language Model

To assess whether cross-lingual LM approaches can benefit from using linguistic phenomena from various languages, we compare a Transformer-based cross-lingual baseline against four cross-lingual LLMs. Our objective is to evaluate cross-lingual LM linguistic capabilities across various languages.

Fine-Tuned Transformer-Based Cross-Lingual Language Model For our cross-lingual baseline, we use XLM-RoBERTa-base (Conneau et al., 2020), a Transformer-based approach.

Zero-Shot Large Language Model Benchmarking all available LLM was outside the scope of this article due to a lack of resources to process the evaluation. LLM benchmark articles have reported using many SOTA GPU devices to do such evaluation (Kew et al., 2023), which we do not have at our disposal. We instead selected five LLMs that were 1) open-source, 2) around 7B parameters, and 3) have been shown to perform well on various benchmark (Kew et al., 2023; Xu et al., 2023; Malode, 2024), or optimized for generation of French text, namely BLOOM-7B (Le Scao et al., 2023), BLOOMZ-7B (Yong et al., 2022), Mistral-7B-v0.3 (Jiang et al., 2023), LLama-3.1-8B (Dubey et al., 2024), and Lucie (Gouvert et al., 2025) (optimized for French) along with their instruct variants (I), if available. We benchmarked all LLMs using HuggingFace's zero-shot-classification.

4.4 Training Settings

Each BERT LM is fine-tuned using the language-specific train and dev split, while RoBERTa LM uses all the languages train and dev splits. All models are evaluated using the test or, if available, OOD split following the standard procedure under the HuggingFace library (Wolf et al., 2020). Each model is fine-tuned for four epochs and uses the AdamW optimizer (Loshchilov and Hutter, 2018), with a learning rate of $3e{-}5$ and a weights decay of $1e{-}2$. Since the corpora are unbalanced, we use a weighted balanced loss based on the train split percentage of acceptable sentences. We use a batch size of 32 and the HuggingFace default train hyperparameters. For each LM, we use the default tokenizer with a maximum sequence length of 64 tokens without lowercasing during tokenization.

5 Results and Discussion

5.1 In-domain Results

Table 7 presents the accuracy and the MCC of all models for each benchmark dataset on the dev and test sets, with **bolded** value indicating the best score per benchmark. Except for the zero-shot evaluation setup, the table reports the average and one standard deviation over the ten restarts. We observe that, for most languages, on average LA-TDA outperforms other fine-tuned methods, but not on all metrics and with a smaller margin than reported by Cherniavskii et al. (2022). The two exceptions to this are CoLA and QFrCoLA. QFrCoLA performs slightly better using the fine-tuned BERT model. Considering that LA-TDA is computed asymptotically in quadratic time (Cherniavskii et al., 2022), the performance gains seem marginal compared to the added computational expense. These results show that fine-tuned Transformer-based LM are strong baselines for the binary linguistic acceptability classification tasks.

Moreover, LLM accuracy performances are either worse than the baselines or at par with it for all languages except Norwegian. In the case of Norwegian, performance is slightly better than the baseline. Llama achieves the worst performance across all languages; however, BLOOMZ and Mistral perform best for most languages. We also observed that, for all LLMs, the instruct (I) version of the LLM performs better than the non-instruct one by, for most of them, a large margin (i.e. double or less the performance). Furthermore, all LLM achieve poor MCC on all splits, with scores close to 0, meaning a negligible correlation between the prediction and the labels. Our experimentation results show that pre-trained cross-lingual LLMs selected for our experimentation do not seem to have acquired linguistic judgment capabilities during their pre-training, nor French optimized LLM (Lucie). Indeed, we can see that even Lucie performed poorly on the task, with an accuracy below the naive approach. Moreover, even our fine-tuned approach (RoBERTa) does not seem to acquire cross-lingual linguistic capabilities from potentially similar linguistic phenomena amongst languages. It shows that leveraging multilingual linguistic corpus to train a multilingual acceptability judgment LM is complex, and more work needs

to be done to achieve better performance than the monolingual approach. Most tested languages do not share a common grammatical language or alphabet (e.g., Japanese and Italian). Thus, it highlights that training LMs on a multilingual dataset without proper grammar assessment could lead to LMs not fully comprehending language linguistics.

Finally, our experiment results on QFrCoLA show that our dataset, which is built from examples that illustrate linguistic norms rather than speakers' feelings, is similar to linguistic acceptability judgment; namely, it is a challenging dataset that can be used to benchmark LM on their linguistic judgment capabilities.

5.2 Out-Of-Domain Results

We present in Table 8 the accuracy and the MCC of our three models trained using QFrCoLA over the dataset's four categories along with the six LLM evaluated in a zero-shot binary classification setup. Except for the LLM, the table reports the average and one standard deviation over the ten restarts. We can see that the category "anglicism" has the lowest performance for the Transformer-based LM. For the two approaches using monolingual LLM (i.e. `BERT` and `LA-TDA`), we hypothesize that this situation is due to occurrences of anglicism in the LLM training dataset. Indeed, using word and syntactical structure borrowed from English grammar is more common over web-based (Laviosa, 2010; Planchon and Stockemer, 2019; Solano, 2021; Šukalić et al., 2022) and even official educational text (Șimon et al., 2021). Thus, fine-tuning the pre-trained LLM model can be more challenging, considering that the "anglicism" category contains the least examples. For the cross-lingual approach, since the LLM has learned word representation over English during training, we hypothesize that sentences using English words or syntax are considered more probable for the model; thus, it is more challenging for the classifier to classify these examples correctly. For the LLM, the "anglicism" performances are worse than the other category and the baseline.

Our experimentation results show that pre-trained cross-lingual or French optimized LMs selected for our experimentation do not seem to have acquired linguistic judgment capabilities during their pre-training, even on the more dominant France-French. Indeed, France has more publicly available datasets online to train LM on, such as OSCAR (Abadji et al., 2022). It shows that these

Model	Dev		Test/OOD	
	Acc (%) (↑)	MCC (↑)	Acc (%) (↑)	MCC (↑)
CoLA				
`Baseline`	69.26	0.000	68.60	0.000
`BERT`	83.61 ± 2.56	**0.639 ± 0.030**	**80.89 ± 1.15**	**0.544 ± 0.025**
`LA-TDA`	**84.91 ± 1.24**	0.633 ± 0.031	80.70 ± 1.38	0.532 ± 0.034
`RoBERTa`	82.24 ± 1.35	0.575 ± 0.033	77.25 ± 2.42	0.452 ± 0.041
`BLOOM`	31.88	0.019	32.56	0.051
`BLOOMZ`	64.14	0.151	60.47	0.044
`Mistral`	30.93	-0.039	33.72	0.073
`Mistral-I`	63.57	0.005	62.02	-0.043
`Llama`	55.03	-0.003	58.53	0.021
`Llama-I`	56.93	-0.003	52.71	-0.039
DaLAJ				
`Baseline`	50.00	0.000	50.00	0.000
`BERT`	69.12 ± 1.53	**0.411 ± 0.029**	72.33 ± 1.40	0.467 ± 0.025
`LA-TDA`	**70.08 ± 1.24**	**0.411 ± 0.024**	**73.54 ± 1.05**	**0.475 ± 0.020**
`RoBERTa`	55.18 ± 5.90	0.131 ± 0.144	55.21 ± 5.89	0.124 ± 0.137
`BLOOM`	50.45	0.010	49.21	-0.020
`BLOOMZ`	50.90	0.047	49.77	-0.011
`Mistral`	65.52	-0.016	66.63	-0.014
`Mistral-I`	52.17	-0.072	51.05	-0.093
`Llama`	38.46	-0.068	37.22	-0.075
`Llama-I`	61.89	0.009	62.57	0.009
ITACoLA				
`Baseline`	85.41	0.000	84.21	0.000
`BERT`	83.29 ± 3.71	0.420 ± 0.051	83.45 ± 3.34	**0.446 ± 0.050**
`LA-TDA`	**87.51 ± 0.88**	**0.423 ± 0.050**	**86.59 ± 0.93**	0.422 ± 0.054
`RoBERTa`	79.97 ± 6.22	0.105 ± 0.121	79.12 ± 5.99	0.117 ± 0.124
`BLOOM`	73.15	0.006	69.00	-0.095
`BLOOMZ`	54.97	-0.058	55.28	-0.052
`Mistral`	15.33	0.036	16.72	-0.014
`Mistral-I`	63.53	-0.036	58.87	-0.032
`Llama`	37.32	0.010	34.26	-0.044
`Llama-I`	32.77	-0.012	30.46	-0.071
RuCoLA				
`Baseline`	74.57	0.000	63.69	0.000
`BERT`	74.49 ± 2.56	**0.352 ± 0.027**	66.81 ± 3.56	0.379 ± 0.030
`LA-TDA`	**77.56 ± 0.61**	0.337 ± 0.022	**71.09 ± 0.92**	**0.382 ± 0.018**
`RoBERTa`	71.84 ± 3.00	0.276 ± 0.038	56.81 ± 3.18	0.189 ± 0.026
`BLOOM`	37.44	-0.084	47.56	-0.012
`BLOOMZ`	59.91	0.014	51.05	-0.040
`Mistral`	26.25	0.036	36.97	0.014
`Mistral-I`	61.65	-0.052	58.76	-0.055
`Llama`	61.95	0.028	53.10	0.049
`Llama-I`	34.99	0.008	44.57	-0.037
CoLAC				
`Baseline`	66.96	0.000	67.82	0.000
`BERT`	75.93 ± 1.35	0.444 ± 0.027	77.78 ± 1.43	0.482 ± 0.023
`LA-TDA`	**77.33 ± 1.79**	**0.469 ± 0.044**	**79.01 ± 0.86**	**0.502 ± 0.023**
`RoBERTa`	73.37 ± 2.72	0.337 ± 0.022	71.09 ± 0.92	0.382 ± 0.018
`BLOOM`	66.96	0.000	67.71	0.001
`BLOOMZ`	63.91	-0.029	65.03	-0.015
`Mistral`	32.83	-0.064	33.15	0.005
`Mistral-I`	38.91	-0.003	37.41	-0.016
`Llama`	62.61	-0.040	64.67	-0.007
`Llama-I`	63.48	0.026	63.76	0.005
NoCoLA				
`Baseline`	32.59	0.000	31.58	0.000
`BERT`	77.90 ± 0.96	0.560 ± 0.009	77.90 ± 0.98	0.560 ± 0.009
`LA-TDA`	**81.58 ± 0.29**	**0.582 ± 0.007**	**82.01 ± 0.31**	**0.589 ± 0.009**
`RoBERTa`	73.92 ± 1.40	0.504 ± 0.017	73.79 ± 1.37	0.505 ± 0.015
`BLOOM`	61.10	0.013	61.31	0.003
`BLOOMZ`	35.92	-0.047	36.92	-0.033
`Mistral`	65.52	-0.016	66.63	-0.014
`Mistral-I`	52.17	-0.072	51.05	-0.093
`Llama`	38.46	-0.068	37.22	-0.075
`Llama-I`	61.89	0.009	62.57	0.009
JCoLA				
`Baseline`	83.93	0.000	73.28	0.000
`BERT`	81.34 ± 4.48	0.039 ± 0.062	73.17 ± 0.61	0.067 ± 0.111
`LA-TDA`	**83.49 ± 0.68**	0.252 ± 0.051	**75.30 ± 1.25**	0.230 ± 0.070
`RoBERTa`	72.64 ± 8.11	**0.262 ± 0.058**	72.86 ± 4.61	**0.328 ± 0.059**
`BLOOM`	24.51	0.036	31.82	0.000
`BLOOMZ`	81.39	-0.002	70.22	-0.007
`Mistral`	18.84	0.031	29.05	0.054
`Mistral-I`	25.09	-0.016	33.43	0.035
`Llama`	31.33	0.006	36.64	0.000
`Llama-I`	62.54	0.001	56.20	-0.126
QFrCoLA				
`Baseline`	69.51	0.000	69.49	0.000
`BERT`	**84.51 ± 0.78**	**0.619 ± 0.02**	**82.92 ± 0.61**	**0.578 ± 0.015**
`LA-TDA`	84.00 ± 0.48	0.606 ± 0.013	82.79 ± 0.45	0.574 ± 0.012
`RoBERTa`	70.67 ± 15.13	0.243 ± 0.263	69.91 ± 14.61	0.222 ± 0.240
`BLOOM`	32.71	0.007	32.94	0.020
`BLOOMZ`	64.00	0.043	61.75	-0.011
`Mistral`	33.50	-0.005	33.45	-0.002
`Mistral-I`	63.03	-0.020	63.61	-0.007
`Llama`	45.43	-0.019	45.44	0.000
`Llama-I`	46.45	-0.026	48.25	-0.001
`Lucie`	60.14	0.041	58.18	-0.008
`Lucie-I`	36.40	-0.034	38.87	0.011

Table 7: Acceptability binary classification results and MCC by language. The best score per benchmark is **bolded**. "OOD" stands for "out-of-domain". ↑ means higher is better

Model	Category			
	Syntax	Morphology	Semantic	Anglicism
	Test Accuracy (%) (↑)			
Baseline	77.24	68.26	72.97	57.18
BERT	**88.59 ± 0.60**	**81.76 ± 0.74**	**85.82 ± 0.40**	**74.36 ± 1.40**
LA-TDA	88.40 ± 0.23	81.49 ± 0.51	85.39 ± 0.53	74.18 ± 1.44
RoBERTa	83.31 ± 4.31	74.93 ± 4.70	79.84 ± 4.88	63.79 ± 4.66
BLOOM	57.67	56.33	55.03	57.36
BLOOMZ	65.66	61.02	64.36	54.61
Mistral	26.53	34.08	30.97	44.86
Mistral-I	67.52	63.76	64.97	55.76
Llama	42.14	46.00	45.70	48.05
Llama-I	46.74	48.81	47.88	49.29
Lucie	59.78	59.27	58.06	53.01
Lucie-I	33.38	40.92	35.15	40.92
	Test MCC (↑)			
Baseline	0.000	0.000	0.000	0.000
BERT	**0.654 ± 0.018**	**0.563 ± 0.017**	**0.620 ± 0.011**	**0.506 ± 0.028**
LA-TDA	0.649 ± 0.009	0.555 ± 0.013	0.609 ± 0.014	0.405 ± 0.026
RoBERTa	0.403 ± 0.279	0.327 ± 0.226	0.378 ± 0.261	0.223 ± 0.156
BLOOM	-0.017	0.024	0.002	0.140
BLOOMZ	-0.044	-0.003	-0.024	0.008
Mistral	0.002	-0.016	-0.002	0.034
Mistral-I	-0.084	0.006	-0.011	0.029
Llama	-0.062	0.016	0.017	-0.004
Llama-I	-0.032	0.017	-0.007	-0.005
Lucie	-0.014	0.005	-0.019	-0.001
Lucie-I	0.009	0.010	0.027	0.015

Table 8: Acceptability binary classification results and MCC for QFrCoLA per category. The best score is **bolded**. ↑ means higher is better.

tested LMs do not seem to have acquired linguistic capabilities from their monolingual training nor from other languages.

Moreover, LLM accuracy performance is always worse for all categories than the baseline, and predictions correlate weakly with labels. It shows again that the benchmarked LLMs do not seem to have a linguistic understanding of Quebec French.

Finally, we present in Table 9 the accuracy and the MCC of our three models trained using QFrCoLA but evaluated using our OOD hold-out set. The table reports the average and one standard deviation over the ten restarts. We can see that, once again, the BERT model outperforms the LA-TDA model. However, all three models show significant performance drops, of nearly 22% in accuracy and nearly 50% for the MCC. It shows that the fine-tuned models have overfitted over the train and dev dataset. As stated before, it is also worth noting that the French in Quebec differ from the French in France. These differences could explain the lower performance observed on the OOD split.

6 Conclusion and Future Works

This article introduced QFrCoLA, the Quebec-French Corpus of Linguistic Acceptability Judgments, a dataset comprising 25,153 in-domain and 2,675 OOD sentences annotated with binary acceptability manually extracted from two official online linguistic normative resources. It is the first such

	OOD Hold-Out	
	Acc (%) (↑)	MCC (↑)
Baseline	53.91	0.000
BERT	**62.69 ± 1.13**	**0.286 ± 0.020**
LA-TDA	61.36 ± 0.90	0.090 ± 0.019
RoBERTa	55.99 ± 4.36	0.107 ± 0.088
BLOOM	45.42	-0.048
BLOOMZ	53.73	0.028
Mistral	46.34	-0.003
Mistral-I	53.06	0.002
Llama	49.30	0.017
Llama-I	49.30	-0.019
Lucie	51.61	-0.018
Lucie-I	47.55	-0.022

Table 9: Acceptability binary classification result on the QFrCoLA out-of-domain (OOD) hold-out set. The best score per benchmark is **bolded**. ↑ means higher is better.

corpus in French and the second-biggest one in any language. We have evaluated the linguistic performances of two monolingual and one cross-lingual fine-tuned Transformer-based LM approaches and four cross-lingual LLM on eight binary acceptability judgement datasets.

Our results demonstrated that Transformer-based LM achieves high results on the binary classification task and are strong baselines. When fined-tuned on QFrCoLA, a Transformer-based LM even outperforms the SOTA LA-TDA method proposed by Cherniavskii et al. (2022). It also shows that pre-trained cross-lingual LLMs selected for our experimentation do not seem to have acquired linguistic judgment capabilities during their pre-training for Quebec French. Finally, our experiment results on QFrCoLA show that our dataset, which is built from examples that illustrate linguistic norms rather than speakers' feelings, is similar to linguistic acceptability judgment; namely, it is a challenging dataset that can be used to benchmark LM on their linguistic judgment capabilities.

In our future works, we plan to extend the granularity of our dataset linguistic phenomena and generate the complementary grammatical or ungrammatical sentence of each sentence in the dataset to create the first French minimal pair benchmark dataset. Moreover, we would also like to explore the linguistic phenomena errors generated by the LLM qualitatively.

Limitations

All the sentences in QFrCoLA have been extracted from official linguistic sources on theoretical syn-

tax and normative grammar. Therefore, those sentences are guaranteed to be theoretically meaningful, making QFrCoLA a challenging dataset. However, the categories extracted automatically from the official source are skewed. Indeed, as shown in Table 3, nearly 42% of the dataset comprises morphological linguistic phenomena. This imbalance means overrepresenting morphology examples, which could provide an incomplete evaluation of a LM's ability to perform the task. Moreover, as discussed, the dataset is based on the OQLF, a Quebec-French government organization, and the *Académie française*; thus, the dataset represents normative grammar. Furthermore, Quebec and France share a common grammar base but differ in some points, such as feminization (e.g. *auteure*/*autrice*). Thus, as discussed, the out-of-domain hold-out is a challenging split since it might represent accepted grammar use in Quebec rather than in France.

Ethical Considerations

QFrCoLA may serve as training data for binary linguistic acceptability judgment classifiers (Batra et al., 2021), which may benefit the quality of generated texts. We acknowledge that such text generation progress could lead to misusing LLMs for malicious purposes, such as disinformation or harmful text generation and online harassment (Weidinger et al., 2021; Bender et al., 2021). Nevertheless, our corpus can be used to train adversarial defence against such misuse and to train artificial text detection models (Lewis and White, 2023; Kumar et al., 2023).

Acknowledgements

This research was made possible thanks to the support of a Canadian insurance company, NSERC research grant RDCPJ 537198-18 and FRQNT doctoral research grant. We thank the reviewers for their comments regarding our work. We also thank the *Office québécois de la langue française* for their help regarding the curation of the corpus.

References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a Cleaner Document-Oriented Multilingual Crawled Corpus. *arXiv:2201.06642*, page arXiv:2201.06642.

Académie française. 2024a. La bataille idéologique. Accessed: 2024-06-15.

Académie française. 2024b. L'institution et l'organisation *(The Institution and the Organization)*. Accessed: 2024-02-10.

Soumya Batra, Shashank Jain, Peyman Heidari, Ankit Arun, Catharine Youngs, Xintong Li, Pinar Donmez, Shawn Mei, Shiunzu Kuo, Vikas Bhardwaj, Anuj Kumar, and Michael White. 2021. Building adaptive acceptability classifiers for neural NLG. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 682–697, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the ACM conference on fairness, accountability, and transparency*, pages 610–623.

Patrycja Bobowska-Nastarzewska. 2009. Quebec French–the Struggle for National Identity. *SORUS SC Wydawnictwo i Drukarnia Cyfrowa*.

Marie-Claude Boivin and Reine Pinsonneault. 2020. La catégorisation des erreurs linguistiques: une grille de codage fondée sur la grammaire moderne. *Le français aujourd'hui*, 2(209):89–116.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models Are Few-Shot Learners. *Advances in neural information processing systems*, 33:1877–1901.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A Survey on Evaluation of Large Language Models. *ACM Transactions on Intelligent Systems and Technology*.

Daniil Cherniavskii, Eduard Tulchinskii, Vladislav Mikhailov, Irina Proskurina, Laida Kushnareva, Ekaterina Artemova, Serguei Barannikov, Irina Piontkovskaya, Dmitri Piontkovski, and Evgeny Burnaev. 2022. Acceptability judgements via examining the topology of attention maps. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 88–107, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Paula Chesley. 2010. Lexical Borrowings in French: Anglicisms as a Separate Phenomenon. *Journal of French Language Studies*, 20(3):231–251.

Noam Chomsky. 2014. *Aspects of the Theory of Syntax*. 11. MIT press.

Leshem Choshen, Guy Hacohen, Daphna Weinshall, and Omri Abend. 2022. The grammar-learning trajectories of neural language models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 8281–8297, Dublin, Ireland. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training With Whole Word Masking for Chinese Bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Véronique Braun Dahlet. 2010. L'orthographe française: entre langue et politique. *Synergies Brésil*, pages 159–166.

Natalia Dankova. 2017. Storytelling in French from France and French from Quebec. *Corela*, 15(2).

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv:2407.21783*.

Zsuzsanna Fagyal, Douglas Kibbee, and Frederic Jenkins. 2006. *French: A Linguistic Introduction*. Cambridge University Press.

Ingo Feldhausen and Sebastian Buchczyk. 2021. Revisiting Subjunctive Obviation in French: A Formal Acceptability Judgment Study. *Glossa: a journal of general linguistics. 6 (1): 59.*

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A Framework for Few-Shot Language Model Evaluation.

Olivier Gouvert, Julie Hunter, Jérôme Louradour, Evan Dufraisse, Yaya Sy, Pierre-Carl Langlais, Anastasia Stasenko, Laura Rivière, Christophe Cerisara, and Jean-Pierre Lorré. 2025. The lucie-7b llm and the lucie training dataset: open resources for multilingual language generation.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. SpaCy: Industrial-strength Natural Language Processing in Python.

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. *arXiv:2305.14091*.

Fred Jelinek, Robert L Mercer, Lalit R Bahl, and James K Baker. 1977. Perplexity—a Measure of the Difficulty of Speech Recognition Tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63.

Matias Jentoft and David Samuel. 2023. NoCoLA: The Norwegian Corpus of Linguistic Acceptability. In *Proceedings of the Nordic Conference on Computational Linguistics*, pages 610–617.

AQ Jiang, A Sablayrolles, A Mensch, C Bamford, DS Chaplot, D de las Casas, F Bressand, G Lengyel, G Lample, L Saulnier, et al. 2023. Mistral 7b (2023). *arXiv:2310.06825*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Tannon Kew, Alison Chi, Laura Vásquez-Rodríguez, Sweta Agrawal, Dennis Aumiller, Fernando Alva-Manchego, and Matthew Shardlow. 2023. BLESS: Benchmarking Large Language Models on Sentence Simplification. *arXiv:2310.15773*.

Sachin Kumar, Vidhisha Balachandran, Lucille Njoo, Antonios Anastasopoulos, and Yulia Tsvetkov. 2023. Mitigating societal harms in large language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 26–33, Singapore. Association for Computational Linguistics.

Per E Kummervold, Javier De la Rosa, Freddy Wetjen, and Svein Arne Brygfjeld. 2021. Operationalizing a national digital library: The case for a Norwegian transformer model. In *Proceedings of the Nordic Conference on Computational Linguistics*, pages 20–29, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. *arXiv:1905.07213*.

Sara Laviosa. 2010. Corpus-Based Translation Studies 15 Years On: Theory, Findings, Applications. *SYNAPS*, 24:3–12.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *arXiv:2211.05100*.

Ashley Lewis and Michael White. 2023. Mitigating harms of LLMs via knowledge distillation for a virtual museum tour guide. In *Proceedings of the Workshop on Taming Large Language Models: Controllability in the era of Interactive Assistants!*, pages 31–45, Prague, Czech Republic. Association for Computational Linguistics.

Dayiheng Liu, Yu Yan, Yeyun Gong, Weizhen Qi, Hang Zhang, Jian Jiao, Weizhu Chen, Jie Fu, Linjun Shou,

Ming Gong, et al. 2021. GLGE: A New General Language Generation Evaluation Benchmark. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP*, pages 408–420.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with Words at the National Library of Sweden – Making a Swedish BERT.

Vishal Manjunatha Malode. 2024. *Benchmarking Public Large Language Model*. Ph.D. thesis, Technische Hochschule Ingolstadt.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a Tasty French Language Model. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Brian 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.

Alessio Miaschi, Dominique Brunato, Felice Dell'Orletta, and Giulia Venturi. 2023. On Robustness and Sensitivity of a Neural Language Model: A Case Study on Italian L1 Learner Errors. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:426–438.

Vladislav Mikhailov, Tatiana Shamardina, Max Ryabinin, Alena Pestova, Ivan Smurov, and Ekaterina Artemova. 2022. RuCoLA: Russian Corpus of Linguistic Acceptability. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 5207–5227.

Chiara Molinari. 2008. Anglais et français au Québec: d'une relation conflictuelle à une interaction pacifique? *Etudes de linguistique appliquée*, 1(149):93–106.

Office québécois de la langue française OQLF. 2024. Féminin de auteur : *autrice* ou *auteure*. Accessed: 2024-06-15.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the annual meeting of the Association for Computational Linguistics*, pages 311–318.

Cecile Planchon and Daniel Stockemer. 2019. Anglicisms, French Equivalents, and Language Attitudes Among Quebec Undergraduates. *British Journal of Canadian Studies*, 32(12):93–118.

Irina Proskurina, Ekaterina Artemova, and Irina Piontkovskaya. 2023. Can BERT eat RuCoLA? Topological Data Analysis to Explain. In *Proceedings of the Workshop on Slavic Natural Language Processing*, pages 123–137.

Elizabeth C Saint. 2013. Les attitudes à l'égard de l'emprunt à l'anglais au Québec et en France: Le cas du domaine informatique. *Communication, lettres et sciences du langage*, 7(1):87–101.

Stefan Schweter. 2020. Italian BERT and ELECTRA Models.

Ramon Marti Solano. 2021. Anglicisms and Corpus Linguistics: Corpus-Aided Research into the Influence of English on European Languages. Introduction.

Taiga Someya, Yushi Sugimoto, and Yohei Oseki. 2023. JCoLA: Japanese Corpus of Linguistic Acceptability. *arXiv:2309.12676*.

Đelaludina Šukalić, Edina Rizvić-Eminović, and Adnan Bujak. 2022. A Corpus-Based Study of Anglicisms Across Different Text Types of Online News. *Journal of French Language Studies*, 20(3):231–251.

Masatoshi Suzuki and Ryo Takahashi. 2019. BERT base Japanese (IPA dictionary). `https://huggingface.co/tohoku-nlp/bert-base-japanese`. Accessed: 2024-02-10.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and Efficient Foundation Language Models. *arXiv:2302.13971*.

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*, pages 2929–2940.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *Advances in neural information processing systems*, 30.

Elena Volodina, Yousuf Ali Mohammed, and Julia Klezl. 2021. Dalaj–a dataset for linguistic acceptability judgments for swedish. In *Proceedings of the Workshop on NLP for Computer Assisted Language Learning*, pages 28–37.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A Stickier Benchmark for General-Purpose Language Understanding Systems. *Advances in neural information processing systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the*

EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Alex Warstadt and Samuel R Bowman. 2019. Linguistic Analysis of Pretrained Sentence Encoders With Acceptability Judgments. *arXiv:1901.03438*.

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

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and Social Risks of Harm From Language Models. *arXiv:2112.04359*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Beilei Xiang, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann. 2021. CLiMP: A Benchmark for Chinese Language Model Evaluation. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2784–2790.

Liang Xu, Anqi Li, Lei Zhu, Hang Xue, Changtai Zhu, Kangkang Zhao, Haonan He, Xuanwei Zhang, Qiyue Kang, and Zhenzhong Lan. 2023. Superclue: A Comprehensive Chinese Large Language Model Benchmark. *arXiv:2307.15020*.

Zheng-Xin Yong, Hailey Schoelkopf, Niklas Muennighoff, Alham Fikri Aji, David Ifeoluwa Adelani, Khalid Almubarak, M Saiful Bari, Lintang Sutawika, Jungo Kasai, Ahmed Baruwa, et al. 2022. BLOOM+ 1: Adding Language Support to Bloom for Zero-Shot Prompting. *arXiv:2212.09535*.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023. Instruction Tuning for Large Language Models: A Survey. *arXiv:2308.10792*.

Dmitry Zmitrovich, Alexander Abramov, Andrey Kalmykov, Maria Tikhonova, Ekaterina Taktasheva, Danil Astafurov, Mark Baushenko, Artem Snegirev, Tatiana Shavrina, Sergey Markov, Vladislav Mikhailov, and Alena Fenogenova. 2023. A Family of Pretrained Transformer Language Models for Russian.

Simona Șimon, Claudia E Stoian, Anca Dejica-Cartiș, and Andrea Kriston. 2021. The use of anglicisms in the field of education: A comparative analysis of Romanian, German, and French. *Sage Open*, 11(4):21582440211053241.