Cross-Lingual Text Classification with Large Language Models

Bin Han
bh193@uw.edu
University of Washington
Seattle, Washington, USA

Sean T. Yang
sean.yang01@yahooinc.com
Yahoo Research
Sunnyvale, California, USA

Christopher LuVogt
luvogt@yahoo.com
Yahoo Research
Sunnyvale, California, USA

Abstract

Cross-lingual text classification involves using a model trained on data in one language to classify text in another, which is crucial in global web applications where labeled data is scarce in certain languages. Although multilingual language models have been leveraged for such tasks, few studies investigate the ability of large language models in cross-lingual classification. In this paper, we evaluate the performance of four large language models and one smaller encoder-only model on cross-lingual text classification tasks using two benchmark datasets. We assess three task settings—direct-test without fine-tuning, direct-test with fine-tuning, and translate-test with fine-tuning. Our findings demonstrate that fine-tuning consistently improves performance, and the translate-test method slightly outperforms others. Furthermore, while large language models experience a slight performance drop on non-English languages, they outperform the baseline encoder-only model in many cases. This study provides insights into the effectiveness of large language models for cross-lingual classification in practical applications.

CCS Concepts

• **Information systems** → **Web applications**; • **Computing methodologies** → **Supervised learning by classification**; **Natural language processing**; *Machine translation.*

Keywords

Large Language Model (LLM); Multilingual Classification; Web Application; Machine Translation; Transferability

ACM Reference Format:
Bin Han, Sean T. Yang, and Christopher LuVogt. 2025. Cross-Lingual Text Classification with Large Language Models. In *Companion of the 16th ACM/SPEC International Conference on Performance Engineering (WWW Companion '25), April 28-May 2, 2025, Sydney, NSW, Australia.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3701716.3715567

1 Introduction

Cross-lingual text classification is a task where a model trained on data in one language is used to classify text in another language. This is essential when labeled data is available only in one language, but the task involves multiple languages, which is common in global web applications. It also plays a key role in developing language-agnostic systems that can operate across various languages with minimal additional training. While the intuitive solution to extend tasks across different languages is to obtain training data in the target languages, this approach is not scalable, especially when dealing with numerous languages.

Previous studies have used multilingual language models (LM) [4, 9] for cross-lingual NLP tasks. Multilingual LMs are pre-trained with monolingual corpora in multiple languages. For downstream applications, they are usually fine-tuned on task data in a high-resource language (e.g., English) and then directly used for inference in the target language, thanks to the aligned representations learned at pre-training time. Other popular approaches leverage machine translation models to assist the task. For example, the translate-train approach [5] involves translating training data from a high-resource language into the target language, which is then used to train the model. In contrast, the translate-test approach [2] translates test data from the target language into a high-resource language, enabling the use of models pre-trained on the high-resource language.

In recent era, Large Language Models (LLMs), such as Generative PreTrained Transformers (GPT) [1], and Meta's LLaMA [13], have shown great capabilities in various NLP tasks [12, 15]. However, as far as we acknowledge, there is no existing study that evaluates their performances in cross-lingual classification tasks. Given the large-scale pre-training and outstanding performances in other NLP tasks [2, 16], we are interested in evaluating LLMs' capabilities in the cross-lingual classification tasks. In this study, we evaluate four LLMs and one smaller Language Model (SLM) on two recent cross-lingual classification datasets with real-world utility. Two task settings, classification or generative, are evaluated for LLMs. We test three methods, "direct-test, no finetune", "direct-test, with finetune", and "translate-test, with finetune".

In this paper, we aim to investigate the following questions: 1) How do LLMs perform in cross-lingual classification compare to a smaller baseline encoder-only language model? 2) What cross-lingual technique generates best results with LLMs? 3) What patterns can be observed that are applicable to practical applications in cross-lingual classification? Overall, our contributions are:

- We benchmark performances of fours LLMs and one SLM on cross-lingual text classification task, providing insights on their capability in working with different languages.
- We conduct comprehensive evaluations on two benchmark datasets, comparing three methods and two task settings. Our findings show that fine-tuning consistently improves performance on both datasets, while the translate-test approach with LLMs delivers slightly better performance compared to other methods.
- We investigate the generalizability of LLMs across three languages, comparing them to an encoder-only SLM.

WWW Companion '25, April 28-May 2, 2025, Sydney, NSW, Australia.

ACM ISBN 979-8-4007-1331-6/25/04
https://doi.org/10.1145/3701716.3715567

2 Experiments

In this section, we articulate the experimental settings of our study, including the definition of the cross-lingual classification problem in §2.1, the descriptions of the two cross-lingual text classification datasets in §2.2, the language models that we evaluate in §2.3, and three methods for the cross-lingual text classification task in §2.4. In §2.5, we describe the details of fine-tuning and evaluation metrics.

2.1 Problem Definition

We first define the cross-lingual classification problem. Denote:

- X^s: text input in the source language
- X^t: text input in the target language
- Y^s: labels corresponding to the input text X^s
- Y^t: labels corresponding to the input text X^t

The problem is to learn a classification function f_θ with data in source language s. The function can then be used on data in the target language t to predict labels:

$$f(X^s, Y^s, X^t; \theta) \rightarrow Y^t \tag{1}$$

2.2 Datasets

For our analysis, we focus on evaluating the candidate models using two recent datasets that offer practical, real-world utility:

- **MultiEURLEX[3]**: a multilingual legal, multi-label document classification dataset. It comprises 65,000 EU laws in 23 official EU languages, with 21 labels representing legal concepts. For each language, the original dataset comprises 55,000 samples for training, and 5,000 for validation and test split respectively. This analysis focuses on scenarios with limited resources, using only 25% of the original dataset splits—11,000 for training, 1,000 for validation, and 1,000 for testing. Each sample across different languages shares approximately the same semantic meaning, making MultiEURLEX a suitable dataset for evaluating multilingual classification problems.
- **AmazonReview[8]**: a multilingual, single-label Amazon review classification dataset. The corpus contains reviews in six languages, with five labels representing five ratings. Each sample in the dataset contains the review text and the star rating. For each language, we use 25,000 training, 1,250 validation and 5,000 test samples to reduce the computational burden. Unlike MultiEURLEX, this dataset does not share semantic meaning across different languages.

We evaluate three languages in our study — English (en), which serves as the baseline to assess and compare performance; French (fr) and Spanish (es), representing foreign languages. All three languages are generally recognized in the literature as high-resource languages. The motivation is to examine cross-lingual classification tasks in contexts that align with real-world web applications, which predominantly focus on high-resource languages. The statistics of both datasets are presented in Table 1. MultiEURLEX has more labels and longer average texts compared to the AmazonReview.

Dataset	# of Labels	Avg. # of Tokens		
		en	fr	es
MultiEURLEX	21	1,184	1,269	1,297
AmazonReview	5	38	31	31

Table 1: Statistics of MultiEURLEX and AmazonReview.

2.3 Language Models

We evaluate the following four LLMs and one SLM:

- **Mistral-7b-Instruct-v0.3 (mistral-7b)**[1]
- **Meta-Llama-3-8B-Instruct (llama3-8b)**[2]
- **Mixtral-8x7B-Instruct-v0.1 (mixtral-8x7b)**[3]
- **Meta-Llama-3-70B-Instruct (llama3-70b)**[4]
- **XLM-RoBERTa-L (XLM)**[5]: a multilingual version of RoBERTa, pre-trained on 2.5TB of filtered CommonCrawl data containing 100 languages. It is evaluated in the state-of-the-art work [2].

We chose the above open-sourced LLMs for three key reasons: 1) Their leading performances in NLP tasks from benchmark leader boards; 2) They are open-sourced and free to use, allowing more thorough and comprehensive evaluation without incurring expensive cost; and 3) Fine-tuning state-of-the-art models like GPT-4 requires sending data through the OpenAI API, which is not suitable for handling private data. All models are evaluated under two different task settings — generative or classification:

- **Generative Setting (g)**: the LLMs share the causal language modeling (next-word prediction) objective introduced by Radford et al. [11]. We provide instructions of how to accomplish the classification tasks to the model in natural language, and receive a natural language response from the model [10]. The generative setting only applies to the decoder-only models.
- **Classification Setting(c)**: the classification setting takes the text embedding from the models, and sends it through a classification head to predict the labels. It is applicable to all aforementioend models. In our experiments, we selectively evaluate XLM, llama3-8b, and llama3-70b under the classification setting.

2.4 Methods

We evaluate the following three methods that are commonly used in the cross-lingual classification literature:

- **Direct-Test, No Finetune (DT, w/o F.T.)** [14]: also known as Zero-Shot Transfer. We use the pre-trained LLMs as they are, by providing them with task instructions in English, and data in target languages. The models then output responses that are post-processed to generate the predicted labels. We only give the models instruction and the data to be classified, without providing any additional examples. This method can only be applied with LLMs under the generative setting.
- **Direct-Test, With Finetune (DT, w F.T.)**: We fine-tune both LLMs and XLM on labeled English data to help the models better learn the specific classification tasks. After fine-tuning, we directly input data in the target languages into the models. This fine-tuning process is performed in both task settings.
- **Translate-Test, With Finetune (TT, w F.T.) [2]**: During the inference time, we first translate data from target language to English using Google Translate [17]. The translated data then goes through a fine-tuned model to predict the labels, under either generative or classification setting. In our experiment, we use Translate-Test method on the fine-tuned models based.

[1] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3
[2] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
[3] https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1
[4] https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct
[5] https://huggingface.co/FacebookAI/xlm-roberta-large

Methods	Setting	Models	MultiEURLEX			AmazonReview		
			F1			Accuracy		
			en	fr	es	en	fr	es
Direct-Test. No Finetune	Generative	mistral-7b(g)	0.267	0.281	0.295	0.560	0.514	0.496
		llama3-8b(g)	0.345	0.354	0.336	0.530	0.494	0.491
		mixtral-8x7b(g)	0.258	0.285	0.303	0.531	0.486	0.480
		llama3-70b(g)	0.490	0.485	0.485	0.629	0.568	0.571
Direct-Test, With Finetune	Generative	mistral-7b(g)	0.678	0.665	0.667	**0.686**	0.596	0.596
		llama3-8b(g)	0.714	0.699	0.696	0.673	0.582	0.584
		mixtral-8x7b(g)	0.658	0.620	0.627	0.680	**0.610**	**0.611**
		llama3-70b(g)	0.733	0.726	0.716	**0.686**	**0.610**	0.608
	Classification	llama3-8b(c)	0.807	0.784	0.766	0.659	0.550	0.561
		llama3-70b(c)	0.805	0.787	0.753	0.659	0.478	0.509
		XLM(c)	**0.809**	0.796	0.798	0.644	0.585	0.584
Translate-Test, With Finetune	Generative	mistral-7b(g)	-	0.677	0.676	-	0.598	0.602
		llama3-8b(g)	-	0.699	0.693	-	0.591	0.601
		mixtral-8x7b(g)	-	0.657	0.665	-	0.601	0.608
		llama3-70b(g)	-	0.729	0.729	-	0.596	0.607
	Classification	llama3-8b(c)	-	0.801	0.793	-	0.576	0.578
		llama3-70b(c)	-	0.793	0.796	-	0.542	0.554
		XLM(c)	-	**0.806**	**0.807**	-	0.579	0.579

Table 2: Classification results on two datasets. F1 and accuracy are reported on MultiEURLEX and AmazonReview respectively. Bold numbers represent the best results for each language. Underlined numbers represent the second best.

2.5 Fine-tuning & Evaluation

To fine-tune LLMs on both datasets, we use qLoRA [6, 7] to tune all layers, with LoRA matrix rank, scaling factor, and dropout set to 16, 32, and 0.05 respectively. We used gradient checkpointing during fine-tuning to save memory. Batch size, warmup ratio. learning rate are set to 1, 0.03 and 2e-4. For all tasks we set the batch size to 1 due to memory limitations. We fine-tune all LLMs for just one epoch given two reasons: 1) the models converged already with one epoch of fine-tuning on the validation set. 2) preliminary experiments on llama3-8b have shown that fine-tuning for five epochs provides similar results as one epoch of fine-tuning. For XLM, we fine-tune the full model for five epochs, with batch size and learning rate set to 8 and 2e-5. For MultiEURLEX dataset, the max sequence length is set to 4096, while it is 512 for AmazonReview. We evaluate F1 values for MultiEURLEX (multi-label classification task) and accuracy for AmazonReview (single-label classification task).

3 Results

The quantitative results of both datasets are reported in Table 2. It is important to note that absolute performance numbers are not the focus of this analysis. The objective is to evaluate performance differences between the target languages and the source language (English), rather than the overall performance in isolation. We also note that performance in target languages is not expected to match that in English. Instead, our goal is to provide a thorough analysis of models and methods to enable informed decision-making.

Models. We observe varying performance across models. While XLM slightly outperforms LLMs on MultiEURLEX, LLMs deliver superior results in the generative setting for the AmazonReview dataset. Among LLMs, Llama models generally outperform Mistral models. Additionally, larger models (e.g., Llama3-70B, Mistral-8x7B) tend to excel in generative settings, whereas smaller models (e.g., Llama3-8B) often perform better in classification settings. We conjure that compared to larger models with vast number of parameters, the parameter count of the classification layer is much closer to that of smaller models. Therefore, the classification layer in smaller models may adapt more easily to the task and have much more impact on the classification results. Additionally, computational cost is a crucial factor when deploying these models in real-world applications. While larger models generally achieve marginally better performance, they also come with significantly higher computational demands in terms of both time and cost. For example, fine-tuning llama3-70b under the generative setting takes over 20 hours, whereas its 8b counterpart requires only a few hours.

	DT, w/o F.T.		DT, w. F.T.		TT, w F.T.	
Dataset	**GEN**	**CLS**	**GEN**	**CLS**	**GEN**	**CLS**
MultiEURLEX	0.349	-	0.677	0.772	0.690	0.796
AmazonReview	0.529	-	0.599	0.525	0.601	0.562

Table 3: Aggregated results across three different cross-lingual techniques. GEN indicates generation setting while CLS means classification setting.

Methods. We compare the results of different cross-lingual methods by averaging performance of all LLMs in Table 3. Translate-Test emerges as the superior technique across both datasets, showing a 2-point improvement over direct-test, with finetune approach. The trade-off between performance gains and the increased latency from translating data into English for Translate-Test should be evaluated on a case-by-case basis.

Cross-lingual Transferability. With the AmazonReview dataset, we observe that all models and methods struggle to generalize to target languages when fine-tuned on English data. Performance on English data significantly surpasses that on French and Spanish by

an average of 7pts, a trend not seen in MultiEURLEX. Even using the translate-test method, results remain inferior to the original English performance. A possible explanation is that MultiEURLEX data is constructed by translating each legal document into multiple languages, ensuring the test sets for English, French, and Spanish share identical content. In contrast, the test sets for product reviews in AmazonReview differ across languages, which introduces performance variances.

To validate this hypothesis, we conduct target fine-tuning by fine-tuning language models using training data in the target language. This approach helps establish a "ceiling" for performance on the test set in the target language. We evaluate the target fine-tuned models on the corresponding test splits using llama3-70(g), llama3-70(c), and XLM. The results are presented in Table 4. The best performances in French and Spanish are achieved by llama3-70b(g), with accuracies of 0.634 and 0.629, reducing the cross-lingual transfer performance gap to under 3 points. This aligns with our MultiEURLEX observations. While XLM demonstrates minimal drops, with decreases of 0.5 and 0.6 points for French and Spanish, respectively, it still underperforms compared to llama3-70b.

Target Language	Fine-Tune Language	llama3-70b(g)	llama3-70b(c)	XLM
fr	**en**	0.610	0.478	0.585
	fr	0.634	0.629	0.590
es	**en**	0.608	0.509	0.584
	es	0.629	0.600	0.600

Table 4: Classification results of target fine-tuning.

4 Discussion and Conclusion

In this study, we benchmark LLMs' capabilities in cross-lingual classification tasks. We evaluate four LLMs and one encoder-only SLM on two cross-lingual classification datasets with tangible real-world applications. Two task settings, classification and generative, are evaluated for LLMs. We test three methods, including "direct-test, no finetune", "direct-test, with finetune", and "translate-test, with finetune". We have observed some consistencies and variances of the results between the two datasets and across the three methods. Overall, we have the following findings:

- **Model-wise**, performance varies across different models. XLM demonstrates the best performance on the MultiEURLEX dataset in the classification setting and is significantly faster to fine-tune, lowering computational costs. However, LLMs consistently deliver more robust and reliable results across various datasets, whether in generative or classification settings.
- **Method-wise**, unsurprisingly, fine-tuning significantly improves results on both datasets by enabling the models to learn the specific classification tasks. While the translate-test approach generally offers slightly better performance, it incurs additional latency due to the translation steps. The choice between methods depends on balancing performance gains against these costs.
- **Cross-lingual transferability**, we observe a roughly two-point drop in performance with LLMs and less than a point drop with XLM. Moreover, while XLM performs best on the MultiEURLEX, LLMs produce the best results on AmazonReview. In practice, we recommend using XLM as a baseline and experimenting with LLMs on the target datasets to identify the best model.

We acknowledge several limitations in our study: 1) We only evaluate open-source LLMs, excluding state-of-the-art models like GPT-4o due to computational costs and privacy concerns with private datasets. Future studies should include these models for a more comprehensive understanding of LLMs' performance in cross-lingual classification tasks. 2) Our evaluation is limited to two high-resource languages from the Indo-European family, without testing languages from other families, such as certain Asian languages, or low-resource languages. Including these languages would provide a clearer picture of the generalizability of the language models.

We provide a baseline for cross-lingual classification with LLMs, unlocking opportunities for expanding applications across different languages on the worldwide web. Future work includes extending the evaluation to low-resource languages and different language families, as well as exploring other natural language processing tasks, such as extraction, information retrieval, data augmentation, and question answering, where the generative capabilities of LLMs can be leveraged.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv:2303.08774* (2023).

[2] Mikel Artetxe, Vedanuj Goswami, Shruti Bhosale, Angela Fan, and Luke Zettlemoyer. 2023. Revisiting Machine Translation for Cross-lingual Classification. In *EMNLP*.

[3] Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. 2021. MultiEURLEX - A multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In *EMNLP*.

[4] Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *NeuRIPs* (2019).

[5] Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating Cross-lingual Sentence Representations. In *EMNLP*. https://aclanthology.org/D18-1269

[6] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *NeuRIPs* (2024).

[7] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. In *ICLR*.

[8] Phillip Keung, Yichao Lu, György Szarvas, and Noah A Smith. 2020. The Multilingual Amazon Reviews Corpus. In *EMNLP*.

[9] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. [n. d.]. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics* ([n. d.]). https://aclanthology.org/2020.tacl-1.47

[10] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *NeuRIPs* (2022).

[11] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. (2018).

[12] Giridhar Kaushik Ramachandran, Yujuan Fu, Bin Han, Kevin Lybarger, Nic Dobbins, Ozlem Uzuner, and Meliha Yetisgen-Yildiz. 2023. Prompt-based Extraction of Social Determinants of Health Using Few-shot Learning. In *ClinicalNLP*. 385–393.

[13] 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* (2023).

[14] Inigo Jauregi Unanue, Gholamreza Haffari, and Massimo Piccardi. 2023. T3l: Translate-and-test transfer learning for cross-lingual text classification. *Transactions of the Association for Computational Linguistics* (2023).

[15] Robert Wolfe, Isaac Slaughter, Bin Han, Bingbing Wen, Yiwei Yang, Lucas Rosenblatt, Bernease Herman, Eva Brown, Zening Qu, Nic Weber, et al. 2024. Laboratory-Scale AI: Open-Weight Models are Competitive with ChatGPT Even in Low-Resource Settings. In *FAccT*. 1199–1210.

[16] Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. In *ICML*. PMLR.

[17] Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675* (2023).