Towards Multilingual LLM Evaluation for Baltic and Nordic languages: A study on Lithuanian History

Yevhen Kostiuk[1,2]
[1]ARG-tech, University of Dundee
[2]OpenBabylon
ykostiuk001@dundee.ac.uk

Oxana Vitman
University of Bremen

Łukasz Gagała
Georg-August Universität Göttingen

Artur Kiulian
OpenBabylon

Abstract

In this work, we evaluated Lithuanian and general history knowledge of multilingual Large Language Models (LLMs) on a multiple-choice question-answering task. The models were tested on a dataset of Lithuanian national and general history questions translated into Baltic, Nordic, and other languages (English, Ukrainian, Arabic) to assess the knowledge sharing from culturally and historically connected groups. We evaluated GPT-4o, LLaMa3.1 8b and 70b, QWEN2.5 7b and 72b, Mistral Nemo 12b, LLaMa3 8b, Mistral 7b, LLaMa3.2 3b, and Nordic fine-tuned models (GPT-SW3 and LLaMa3 8b).

Our results show that GPT-4o consistently outperformed all other models across language groups, with slightly better results for Baltic and Nordic languages. Larger open-source models like QWEN2.5 72b and LLaMa3.1 70b performed well but showed weaker alignment with Baltic languages. Smaller models (Mistral Nemo 12b, LLaMa3.2 3b, QWEN 7B, LLaMa3.1 8B, and LLaMa3 8b) demonstrated gaps with Lithuanian national history related questions (LT-related) alignment with Baltic languages while performing better on Nordic and other languages. The Nordic fine-tuned models did not surpass multilingual models, indicating that shared cultural or historical context alone does not guarantee better performance.

1 Introduction

Large Language Models provide a functional framework for tackling various natural language processing (NLP) tasks, such as question-answering (Izacard et al., 2022; Dong et al., 2024), machine translation (Zhu et al., 2023; Kocmi et al., 2024) and so on. However, LLMs have shown less reliable results for low-resource languages (Ranjan et al., 2024; Sakib and Das, 2024) due to the smaller fraction of available data in comparison to English and a few other widely spoken languages.

Benchmarking multilingual LLMs across languages is essential for evaluating their capabilities. However, the availability of high-quality, culturally aligned datasets remains a challenge. This need for culturally aligned high-quality datasets becomes even more critical when evaluating historical knowledge, where ensuring linguistic and cultural fairness adds a layer of complexity.

Verifying comparability of the results on historical knowledge QA datasets requires that a single set of historical events is queried in all languages. The choice of that set is likely to be biased to events more represented in widely spoken languages. Conversely, events that are more region- or cultural-specific are less likely to occur in the benchmarks. Finding and addressing these gaps is important to improve the fairness of LLMs and highlight historical and cultural biases.

In this work, we focus on evaluating multilingual LLMs on Lithuanian and general history. Our goal is to determine how LLMs perform on Lithuanian history exam questions when prompted in different languages and explore the alignment between languages and historical awareness, particularly within the Nordic and Baltic language groups.

Our contribution is the following:

- We automatically translated publicly available Lithuanian history exam question-answering dataset into Nordic (Danish, Finnish, and Swedish), Baltic (Estonian and Latvian), and other (Arabic, Ukrainian, and English) languages and partially manually

Proceedings of the 1st NB-REAL Workshop 2025 - Nordic-Baltic Responsible Evaluation and Alignment of Language models, pages 1–11
March 2, 2025 ©2025 Association for Computational Linguistics

evaluated it.

- We tested GPT-4o (OpenAI et al., 2024), LLaMa3.2 3b, LLaMa3 8b, LLaMa3.1 8b and 70b (Dubey et al., 2024), Mistral Nemo 12b (Jiang et al., 2023), QWEN2.5 7b and 72b (Team, 2024; Yang et al., 2024), and GPT-SW3 and Nordic-trained LLaMa3 8b (Ekgren et al., 2023) models and compared their achieved accuracy scores per language and its average per language group.

Our findings revealed that GPT-4o consistently outperformed other models across all evaluated languages and language groups on a dataset of LT-related and general history questions. Larger open-source models, such as LLaMa3.1 70b and QWEN2.5 72b, also demonstrated strong and consistent performance in all languages. In contrast, smaller models like Mistral Nemo 12b, LLaMa3 8b, LLaMa3.2 3b, and LLaMa3.1 8b showed notable gaps in their historical knowledge from a Lithuanian perspective, particularly with Baltic languages, despite Lithuanian being part of this group. The best performance was observed in the Nordic language group, suggesting that cultural or historical alignment alone does not ensure higher accuracy. Interestingly, the Nordic pretrained models failed to surpass the multilingual model.

The code and data are available in our GitHub repository[1].

2 Related Work

Pre-trained LLMs have exhibited a remarkable ability to encode and retrieve factual and common knowledge across different languages (Wang et al., 2023; Zhao et al., 2024). However, there is a notable variation in model performance across languages, with a strong shift toward high-resource languages (Qi et al., 2023), particularly languages with Latin scripts (Ifergan et al., 2024).

The datasets used for benchmarking multilingual LLMs are created using either one of the two approaches: human annotation (Kocmi et al., 2023; Goyal et al., 2022) or translating existing annotated datasets using LLMs (Lai et al., 2023).

Although datasets created by human annotators provide accurate translations and task-specific precision, they require considerable investment of both time and finances (Yang et al., 2019).

On the other hand, with an advancement of LLMs, the translation performance of automatic tools has been significantly boosted lately. For example, ChatGPT demonstrates fewer errors with the launch of the GPT-4 engine, even for distant languages (Jiao et al., 2023). The quality control research on the DeepL translation tool found that DeepL[2] performed well in terms of translation accuracy, fluency, and naturalness, reaching an overall semantic similarity score 94.13 (Linlin, 2024).

This improvement elevated the creation of benchmark datasets on various tasks. DeepL was used for creating the X-FACT multilingual factual knowledge dataset translated in 25 languages (Gupta and Srikumar, 2021). In the research (Thellmann et al., 2024), five well-known datasets of various tasks were translated by DeepL into 21 European languages. LLMs with different numbers of parameters were evaluated on newly introduced datasets. The authors observed that models generally achieve higher performance on Romance and Germanic languages compared to Slavic languages.

ChatGPT was utilized to translate the 158K English instructions into 26 languages, including 7 low-resource languages (Lai et al., 2023). The data was used to instruction-tune LLM for multiple languages using reinforcement learning from human feedback. The resulting framework, Okapi, was also evaluated on datasets translated by ChatGPT from English into 26 selected languages.

3 Methodology

In this paper, we investigate performance consistency of LLMs within Nordic and Baltic language groups on the Lithuanian history exams questions. We hypothesized that the LLMs perform better in this domain, when presented with questions in languages from Nordic and Baltic groups than from other due to the cultural, linguistic and historical similarities.

The methodology consists of two steps: *data preparation* and *models' benchmarking*.

Data Preparation. To test the hypothesis, we chose EXAMS (Hardalov et al., 2020) dataset. Specifically, we used samples that correspond to Lithuanian history. Each sample contains a ques-

[1] https://github.com/OpenBabylon/NoDaLiDa2025-LT-History-Eval

[2] https://www.deepl.com/

Question

Kuria kalba parašytas Trečiasis Lietuvos Statutas?

EN Translation:
In which language was the Third Statute of Lithuania written?

Choices	*EN Translation*
A) Lietuvių.	*A) Lithuanian*
B) Lenkų.	*B) Polish*
C) Lotynų.	*C) Latin*
D) Rusėnų.	*D) Ruthenian*

Correct Answer: D

Figure 1: Example of the dataset sample in Lithuanian.

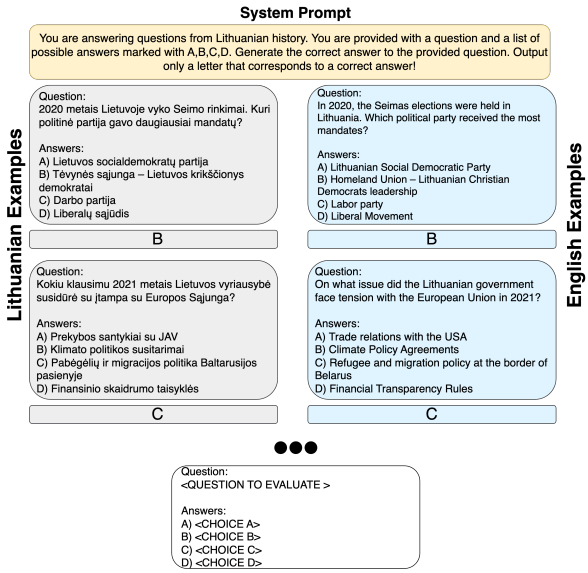

Figure 2: Example of the chat prompts that the model was presented to evaluate the dataset in Lithuanian and English languages.$< \ldots >$ is the actual question that the model is evaluated on.

tion, four different answer choices marked with the labels A,B,C and D with an indication of the correct one (see Figure 1). Questions and choices are in Lithuanian. We manually removed the questions that require an image to answer it, obtaining 550 samples.

The dataset was machine translated into Nordic (Danish, Finnish, and Swedish), Baltic (Estonian and Latvian), and outside of Nordic-Baltic, multilingual language group: Ukrainian, English, and Arabic. In more details, the dataset was translated from Lithuanian to English, and then the English translations were translated in other languages. We used GPT-4o (OpenAI et al., 2024) and DeepL as translation algorithms, as they are proven to have a good machine translation performance from- and to-English rather than between underrepresented languages (Wang, 2024; Hendy et al., 2023).

After that, we separated dataset into 2 parts: Lithuanian national history related questions (LT-related) and general history questions. We assigned a question to the LT-related group if it specifically mentions Lithuania, mentions Lithuanian historic figure or a question about the country that Lithuania was a part of or occupied by (e.g. Polish-Lithuanian Commonwealth, USSR after 1940 etc.). Other questions were assigned to a general history questions group.

To ensure quality, a subset of the dataset was evaluated manually by a group of native speakers. Annotators were presented with 100 English and translated language pairs (50 from LT-related and 50 from General history dataset) with 20 samples being the same for all the annotators to measure the annotators' agreement. For more details, see Appendix A.

Models' benchmarking. We experimented on the following models: LLaMa3 8b, LLaMa3.1 70b and 8b (Dubey et al., 2024), LLaMa3.2 3b (Dubey et al., 2024), Mistral Nemo 12b (Jiang et al., 2023), GPT-4o (OpenAI et al., 2024), QWEN2.5 7b and 72b (Team, 2024; Yang et al., 2024), and families of instruct pre-trained models developed by AI Sweden[3]: GPT-SW3 (Ekgren et al., 2023) (126m, 356m, 1.3b, 6.7b) pre-trained for and LLaMa3 8b fine-tuned for Swedish, Norwegian and Danish. GPT-SW3 models were pre-trained for Swedish, Norwegian, Danish, Icelandic, English, and programming code and LLaMa3 8b (we will refer to this model as NRD LLaMa3 from now on to avoid confusion) was fine-tuned for Swedish, Norwegian and Danish.

In our experiments, we used multilingual instruct LLMs. During generation, all the parameters were set to defaults, except for random seed, which we set to 2 with Ollama [4] framework for open source models. For Nordic models, we

[3] https://huggingface.co/AI-Sweden-Models

[4] https://github.com/ollama/ollama

used implementation from the *transformers* (Wolf et al., 2020). Specifically, we used GPT-SW3 and Swedish LLaMa3 8b[5]. The models were shown the same set of questions, translated to the corresponding language.

Another limitation of the approach is that we used GPT-4o for both translation and evaluation. To ensure that there is no data leakage, the model was shown only one sentence from the dataset at a time during translation and evaluation (along with manually crafted few-shot examples).

For each language in the translated dataset, the model was evaluated on the multiple choices question-answering task. The model was presented with a system message in English explaining the task, four question-answering examples in the corresponding language, and finally, a question with answer choices that the model has to answer. The examples were presented in the same format as the final question and consisted of question, four answer choices marked with A, B, C, D, and the correct letter for an answer as an expected output. Examples were taken from the modern (later than 2020) history of Lithuania, and do not intersect with questions in the dataset. Everything, except the system prompt, was presented to the model in the evaluated language (see Figure 2).

The results were parsed in the following way. If the model generated more than one letter, the generated text was separated into words. From these words, only capital letters were kept that corresponds to possible choice letters A,B,C, or D. If only one letter was present, it was considered as a final output. Otherwise, we assume that the model failed to produce a reasonable output and record it as if the model's answer was incorrect. As a result, we measured accuracy score for each model and each language.

During the evaluation, the translation quality can influence the final results. Since the original dataset was in Lithuanian, we would expect the models perform better on Lithuanian, as it did not go through translation steps. It is can be viewed as a possible advantage for Lithuanian over other languages, particularly for LT-related history questions.

[5] https://huggingface.co/collections/AI-Sweden-Models/

4 Results and Discussion

For each LLM, we grouped its results per language group into Nordic (Danish, Finnish, and Swedish), Baltic (Lithuanian, Latvian, Estonian), and multilingual (Ukrainian, English, and Arabic). The accuracy scores per language and averages scores per language group are presented in Tables 1 and 2 and on Figures 3, 4, and 5.

Our results demonstrated that all the models except for GPT-4o obtained better scores for general history questions rather than for LT-related ones. This observation is expected due to a biased training datasets for such models towards English-centric data.

The largest evaluated model, GPT-4o, performed consistently better than other models for LT-related and general history questions in all language groups. The model achieved a maximum average score of 0.88 for LT-related history questions for the Baltic group (BLT) and performed similarly for Nordic languages (NRD) with a score of 0.87, though it showed slightly weaker performance in the multilingual group (MLT), scoring 0.84. These results suggest better knowledge representation for Nordic and Baltic language groups in LT-related history exams. Among the individual languages, English and Lithuanian were the best-performing languages for both LT-related and general history questions.

The 70b group of models (QWEN2.5 72b and LLaMa3.1 70b) demonstrated second best performance across all the types of questions. QWEN2.5 showed lower accuracy for Baltic languages on average, obtaining similar scores for MLT and NRD groups. Also, in both types of questions, QWEN2.5 showed similar trends of receiving lower scores in Estonian and Latvian, but higher scores for Nordic languages. Additionally, its performance was better in general questions across all languages, but when it comes to the alignment with LT-related, model was able to output better results for English, Swedish, and Danish rather than for Lithuanian or Baltic languages. In contrast, LLaMa3.1 70b did not performed at par with diff language groups. The results are similar for all languages in all questions with Arabic being the weakest and Lithuanian with English slightly stronger than others.

In case of Mistral Nemo 12b, the model scored the smallest scores comparing to other, even smaller (7-8b, 3b) models. It showed similar re-

	NRD			BLT			MLT		
	LT	*G*	*LT+G*	*LT*	*G*	*LT+G*	*LT*	*G*	*LT+G*
GPT-4o	0.87	0.89	0.88	0.88	0.89	0.89	0.84	0.88	0.86
QWEN2.5 72b	0.74	0.87	0.81	0.71	0.83	0.77	0.76	0.87	0.82
LLaMa3.1 70b	0.72	0.82	0.77	0.72	0.81	0.76	0.72	0.81	0.77
M Nemo 12b	0.36	0.49	0.43	0.36	0.42	0.39	0.41	0.55	0.48
LLaMa3.1 8b	0.47	0.62	0.54	0.44	0.57	0.50	0.50	0.66	0.58
LLaMa3 8b	0.45	0.48	0.46	0.39	0.40	0.40	0.48	0.53	0.50
QWEN2.5 7b	0.49	0.62	0.56	0.46	0.48	0.47	0.58	0.73	0.65
LLaMa3.2 3b	0.40	0.50	0.45	0.34	0.33	0.34	0.42	0.47	0.45

Table 1: Average accuracy results per language group and model. **NRD** stands for Nordic, **BLT** stands for Baltic, and **MLT** stands for multilingual language groups. *LT-R*, *G*, and *LT+G* stand for Lithuania-related history questions, general history questions and merged history questions respectively. **M Nemo 12b** refers to Mistral Nemo 12b model.

	SW			DN			EN		
	LTR	*G*	*LTR+G*	*LTR*	*G*	*LTR+G*	*LTR*	*G*	*LTR+G*
NRD LLaMa3 8b	0.43	0.50	0.46	0.42	0.51	0.46	–	–	–
GPT-SW3 126m	0.27	0.22	0.24	0.27	0.21	0.24	0.27	0.20	0.24
GPT-SW3 356m	0.27	0.21	0.24	0.27	0.21	0.24	0.23	0.20	0.21
GPT-SW3 1.3b	0.23	0.24	0.23	0.23	0.23	0.23	0.23	0.24	0.24
GPT-SW3 6.7b	0.32	0.27	0.29	0.33	0.29	0.29	0.24	0.28	0.26

Table 2: Accuracy results for Nordic fine-tuned models. **NRD LLaMa3 8b** refers to pre-trained LLaMa3 8b by AI Sweden. *LTR*, *G*, and *LTR+G* stand for Lithuania-related history questions, general history questions and merged history questions respectively. **SW** (Swedish), **DN** (Danish), **EN** (English) indicate a language that was used for evaluating the model.

sults across all language groups, obtaining the same average accuracy scores (36%) on Baltic and Nordic group on LT-related questions and a better performance for Nordic group on general questions than for Baltic. The average of MLT group was better, even though neither score was higher than 64%.

LLaMa3 8b, QWEN2.5 7b, and LLaMa3.1 8b demonstrated a weaker performance when tested on BLT group across all questions. Using Lithuanian showed a better results. Similarly, Swedish and Danish helped QWEN2.5 7b obtain a better score. This results indicate that these models are better aligned with Lithuanian national history when asked in a language from Nordic group or in Lithuanian. LLaMa3.2 3b showed similar performance on NRD group to Mistral Nemo, but in MLT and BLT settings it received the lowest scores.

The Nordic-specific models performed similarly on all their supported languages. From the considered models, NRD LLaMa3 is a clear winner. It demonstrated a similar performance across its supported languages and is very close to LLaMa3.2 performance on Swedish and Danish, but still underperformed LLaMa3.1 8b and QWEN2.5 7b on the corresponding languages. When it comes to a family of GPT-SW3, the greater the amount of parameters - the better performance. GPT-SW3 6.7b outperformed other versions of the model across Swedish and Danish. However, on English, GPT-SW3 with 126m performed better on LT-related questions.

While our findings suggest that shared cultural or historical context does not guarantee better model performance, the other factors could potentially play a role. The evaluated multilingual models were trained on disproportionately larger datasets for Nordic languages due to its better availability (e.g. Wikipedia articles for Swedish and Danish etc.). This disproportion can explain the performance gaps, even for general

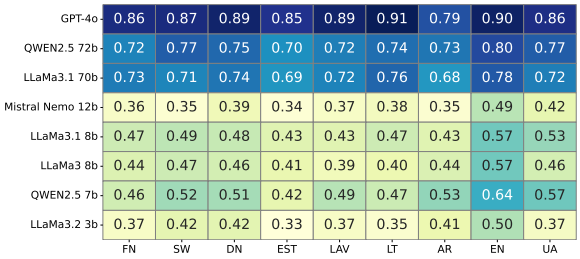

Figure 3: Accuracy results per language for LT-related history questions.

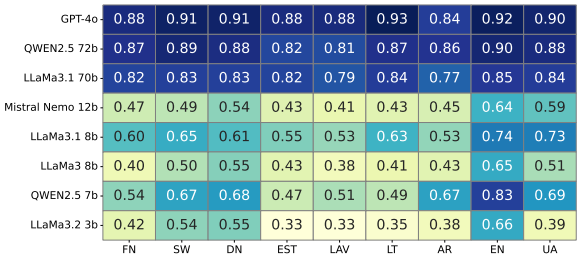

Figure 4: Accuracy results per language for general history questions.

knowledge questions. For instance, in our results, smaller models consistently achieved higher accuracy on Swedish and Danish compared to Lithuanian across both general and LT-related questions. These differences highlight the importance of training data availability and linguistic representation, in addition to cultural and historical alignment, in shaping LLM performance. Future work should further investigate the interaction between these factors to better address the challenges of underrepresented languages.

In conclusion, our experiments show that GPT-4o performs consistently better across all tested languages and language groups on LT-related and general history questions. Larger open source models, LLaMa3.1 70b and QWEN2.5 72b also performed consistently well in all languages. Mistral Nemo 12b, LLaMa3 8b, LLaMa3.2 3b, and

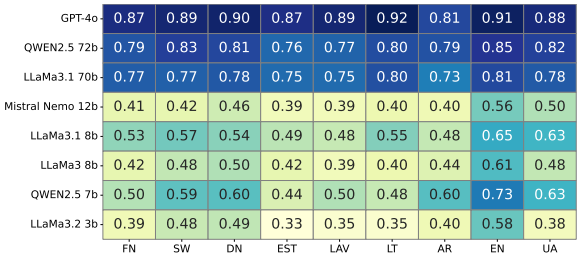

Figure 5: Accuracy results per language for merged LT-related and general history questions.

LLaMa3.1 8b demonstrated significant gaps in their historical knowledge for LT-related history questions within Baltic language group, even when Lithuanian is part of this group. The better performance was obtained in Nordic language group, indicating that cultural or historical alignment alone does not guarantee higher accuracy for these models. The Nordic pre-trained models were not able to outperform the multilingual model, rejecting our initial hypothesis.

5 Conclusion

This study evaluated the performance of Large Language Models (LLMs) on Lithuanian historical multiple-choice question-answering tasks, focusing on Baltic, Nordic, and other language groups. The models were evaluated on the Lithuanian national history related (LT-related) questions and a general history questions.

Our findings showed that GPT-4o consistently outperformed all other tested models across languages, achieving the highest scores for LT-related and general history questions, with slightly better results for Baltic and Nordic languages. Among open-source models, larger models QWEN2.5 72b and LLaMa3.1 70b performed well but did not match GPT-4o, especially in Baltic languages. Smaller models, including Mistral Nemo 12b, LLaMa3.2 3b, QWEN 7B, LLaMa3.1 8B, and LLaMa3 8b demonstrated weaker results with Baltic languages, including Lithuanian, while performing better in Nordic and multilingual groups.

Nordic fine-tuned models performed consistently across their supported languages but failed to surpass general multilingual models, even within their specialized domain. These findings highlight that shared cultural or historical context alone does not guarantee better model performance. To bridge these gaps, further efforts are needed to develop targeted datasets and fine-tuning strategies to improve LLM alignment with less-resourced languages like those in the Baltic language group.

Acknowledgments

This chapter is a product of the research conducted in the Collaborative Research Center 1342 “Global Dynamics of Social Policy”. The center is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—project number 374666841—SFB 1342. We would like

to express our gratitude to the following organizations for their generous support, which made this work possible: GCP (Google Cloud) for providing credits used for model training and inference, and Tensorwave for providing AMD MI300X instance for inference and evaluations.

References

Mohamad Adam Bujang and Nurakmal Baharum. 2017. Guidelines of the minimum sample size requirements for cohen's kappa. *Epidemiology Biostatistics and Public Health*, 14:e12267–1.

Guanting Dong, Yutao Zhu, Chenghao Zhang, Zechen Wang, Zhicheng Dou, and Ji-Rong Wen. 2024. Understand what LLM needs: Dual preference alignment for retrieval-augmented generation. *CoRR*, abs/2406.18676.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanaz-

eri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models.

Ariel Ekgren, Amaru Cuba Gyllensten, Felix Stollenwerk, Joey Öhman, Tim Isbister, Evangelia Gogoulou, Fredrik Carlsson, Alice Heiman, Judit Casademont, and Magnus Sahlgren. 2023. Gpt-sw3: An autoregressive language model for the nordic languages. *arXiv preprint arXiv:2305.12987*.

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. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Ashim Gupta and Vivek Srikumar. 2021. X-fact: A new benchmark dataset for multilingual fact checking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 675–682, Online. Association for Computational Linguistics.

Momchil Hardalov, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and Preslav Nakov. 2020. EXAMS: A multi-subject high school examinations dataset for cross-lingual and multilingual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, EMNLP '20, pages 5427–5444, Online. Association for Computational Linguistics.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Maxim Ifergan, Leshem Choshen, Roee Aharoni, Idan Szpektor, and Omri Abend. 2024. Beneath the surface of consistency: Exploring cross-lingual knowledge representation sharing in llms. *arXiv preprint arXiv:2408.10646*.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Few-shot Learning with Retrieval Augmented Language Models.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *arXiv preprint arXiv:2301.08745*.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327, Singapore. Association for Computational Linguistics.

Li Linlin. 2024. Artificial intelligence translator deepl translation quality control. *Procedia Computer Science*, 247:710–717.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wain-

wright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. Cross-lingual consistency of factual knowledge in multilingual language models. *arXiv preprint arXiv:2310.10378*.

Rajesh Ranjan, Shailja Gupta, and Surya Narayan Singh. 2024. A comprehensive survey of bias in llms: Current landscape and future directions.

Shahnewaz Karim Sakib and Anindya Bijoy Das. 2024. Unveiling and mitigating bias in large language model recommendations: A path to fairness.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Klaudia Thellmann, Bernhard Stadler, Michael Fromm, Jasper Schulze Buschhoff, Alex Jude, Fabio Barth, Johannes Leveling, Nicolas Flores-Herr, Joachim Köhler, René Jäkel, et al. 2024. Towards multilingual llm evaluation for european languages. *arXiv preprint arXiv:2410.08928*.

Jiaan Wang, Yunlong Liang, Zengkui Sun, Yuxuan Cao, Jiarong Xu, and Fandong Meng. 2023. Cross-lingual knowledge editing in large language models. *arXiv preprint arXiv:2309.08952*.

Jingjing Wang. 2024. Exploring the potential of chatgpt-4o in translation quality assessment. *Journal of Theory and Practice in Humanities and Social Sciences*, 1(3):18–30.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

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)*, pages 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Xin Zhao, Naoki Yoshinaga, and Daisuke Oba. 2024. Tracing the roots of facts in multilingual language models: Independent, shared, and transferred knowledge. *arXiv preprint arXiv:2403.05189*.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Jiajun Chen, Lei Li, and Shujian Huang. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *ArXiv*, abs/2304.04675.

A Manual Translation Quality Evaluation

The annotation guidelines and examples can be found in our GitHub repository. For each translated language, we utilized the same following strategy. We recruited native speaker annotators, who are also are proficient in English. They were presented with 80 random samples from the dataset distinct for each annotator and 20 samples that are the same for each annotator. From those 80 samples, 40 were selected from a pool of Lithuanian history questions, and other 40 from the general history question. The same approach was applied for the remaining 20 samples: 10 were selected from a pool of Lithuanian history questions, and other 10 from the general history question.

The annotators were presented with the translated English question, its answer choices and the corresponding translation for the question and choices. in the case of Lithuanian to English translation, the pairs of Lithuanian and English were presented. They were instructed to determine if the translation is correct from the following standpoints. The translation accurately conveys the meaning of the English or Lithuanian text. The

Lang Pair	# Reject (A)	# Reject (B)	# Accept (A)	# Accept (B)	Intersect, %	Cohen Kappa
LT-EN	11	18	89	82	75	0.286
EN-UA	10	30	90	70	75	0.286
EN-AR	28	21	72	79	65	0.239
LT-EST*	13	54	87	46	0.55	0.0
EN-SW	1	6	99	94	0.9	-0.053
EN-DN	9	41	91	59	0.6	-0.013
EN-FN	15	33	78	67	0.73	0.189
LT-LAV*	27	43	73	57	0.7	0.381

Table 3: Annotation results. * indicates translation with DeepL from Lithuanian to the target language. # Reject and Accept refer to a number of rejected and accepted samples by the annotator (marked with letters A and B). Intersect indicates a percentage of samples that annotators assigned the same label.

order of answers (with respect to the letters) is the same in both languages. The names of historical figures, locations, dates, or events are correctly translated and align with conventions. Text semantics are clear and do not change the intent or emphasis of the question or answers. If the translation contains grammar or phrasing issues, or minor typos, they do not lead to confusion or ambiguity and do not change the semantics.

If the translation does not fit the requirements above, the translation is rejected. The annotation results and agreements (in a form of number of intersections and Cohen Kappa scores) are presented in the Table 3. During our experiments, chatGPT showed poor results when translating to Latvian and Estonian. Therefore, we used DeepL to translated Lithuanian to Latvian and Estonian. The annotation in the Table 3 corresponds to DeepL translation.

The obtained Cohen Kappa scores were not high, especially for Swedish and Danish. As we only had 20 samples for comparison (Bujang and Baharum, 2017), the Cohen Kappa score is not reliable in this case, we additionally calculated the number of intersections.