A Comparative Study of Translation Bias and Accuracy in Multilingual Large Language Models for Cross-Language Claim Verification

Aryan Singhal **Veronica Shao** **Gary Sun** **Ryan Ding** **Jonathan Lu** **Kevin Zhu**
Algoverse AI Research
jonathan@algoverse.us, kevin@algoverse.us

Abstract

The rise of digital misinformation has heightened interest in using multilingual Large Language Models (LLMs) for fact-checking. This study systematically evaluates translation bias and the effectiveness of LLMs for cross-lingual claim verification across fifteen languages from five language families: Romance, Slavic, Turkic, Indo-Aryan, and Kartvelian. Using the XFACT dataset to assess their impact on accuracy and bias, we investigate two distinct translation methods: pre-translation and self-translation. We use mBERT's performance on the English dataset as a baseline to compare language-specific accuracies. Our findings reveal that low-resource languages exhibit significantly lower accuracy in direct inference due to underrepresentation in the training data. Furthermore, larger models demonstrate superior performance in self-translation, improving translation accuracy and reducing bias. These results highlight the need for balanced multilingual training, especially in low-resource languages, to promote equitable access to reliable fact-checking tools and minimize the risk of spreading misinformation in different linguistic contexts.

1 Introduction

Multilingual Large Language Models (LLMs), such as GPT-4 and Llama 3.1, have shown remarkable capabilities in various languages and tasks [Ahuja et al., 2024]. Thus, there has been increasing interest in possible usages of LLMs for claim verification across languages [Panchendrarajan and Zubiaga, 2024].

However, recent studies have revealed significant disparities in their performance and bias in different languages [Xu et al., 2024, Huang et al., 2024]. This variability is especially concerning given the importance of claim verification in combating misinformation [Sundriyal et al., 2023]. The performance discrepancies observed in LLMs often favor resource-rich languages like English, French, and German over resource-poor languages such as Kannada and Occitan [Robinson et al., 2023, Bawden and Yvon, 2023, Quelle and Bovet, 2024]. These differences stem from variations in accuracy and translation quality between languages. Although LLMs demonstrate impressive average performance in a wide range of languages, Li et al. [2024] highlights persistent gaps between high-resource and low-resource languages, emphasizing the need for more balanced data collection and training approaches.

Addressing misinformation for claim verification tasks is critical, as ineffective claim verification can spread false information between languages and vulnerable populations [Thorne and Vlachos, 2018]. Although advances in LLMs, such as Meta's Llama 3.1 models [Dubey et al., 2024], have improved multilingual capabilities, reliance on external translation methods in some contexts—especially by users or systems that use third-party services such as Google Translate or that rely on the LLM

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

in use and its multilingual capabilities—can still introduce biases. These biases can undermine the improvements made by LLMs and contribute to the spread of misinformation, particularly in resource-poor languages. Ensuring fair and accurate fact-checking in multiple languages is essential for equitable access to reliable information worldwide [Zhang et al., 2024].

This study evaluates pre-translation and self-translation methods across 15 languages, grouped into five language families—Romance, Slavic, Turkic, Indo-Aryan, and Kartvelian—spanning both high- and low-resource languages. We use mBERT's performance on the English dataset as a baseline to measure language-specific accuracy and the effectiveness of translation. The translation techniques are further explained in Section 3.4 and are evaluated against the XFACT dataset by Gupta and Srikumar [2021]. Our analysis aims to inform the development of more balanced LLMs and guide future efforts in claim verification, helping to close the performance gap between high- and low-resource languages and creating more equitable language technologies.

2 Related Works

2.1 English and Multilingual Fact-Checking

The application of LLMs for fact-checking tasks has emerged as a promising area of research. Quelle and Bovet [2024] demonstrated that the GPT-3.5 and GPT-4 models can achieve high accuracy in English fact-checking tasks when provided with adequate context. However, the challenge of extending these capabilities across multiple languages has driven research towards multilingual approaches. For example, Huang et al. [2022] enhanced mBERT with cross-lingual retrieval techniques, improving fact-checking performance in the X-Fact dataset. Hu et al. [2023] further evaluated the factual knowledge of ten different LLMs in 27 languages, revealing insights into the multilingual capabilities of these models. Despite these advances, many studies have grouped non-English languages into a single category without detailed analysis, leaving a gap for users who wish or need to use other under-researched languages.

2.2 Bias in Multilingual Language Models

Wealthier countries often support more LLM research, leading to an uneven distribution of training data favoring their languages [Dong et al., 2024, HAI, 2023]. LLMs also exhibit political and informational biases, emphasizing claims spread by the media in wealthy countries over those in low-income countries. Shafayat et al. [2024] highlighted a significant bias toward Western-centric political information in the factual accuracy of LLMs across nine languages. Moreover, these models tend to produce more factual content in high-resource languages and longer responses in English.

3 Experimental Setup

3.1 Datasets

Our study uses the X-Fact dataset[1] developed by Gupta and Srikumar [2021] as the primary source of claims in selected language families. We systematically source 600 claims for each language family, ensuring a balanced representation of languages within each family and an equal distribution across the dataset's five veracity labels: "True", "Mostly True", "Half True", "Mostly False", and "False". The claims were selected to maintain an even distribution across both languages and veracity labels. This allowed for a diverse corpus encompassing both political and non-political topics. A detailed breakdown of the languages included in each family and the final dataset distribution is provided in Appendix A.1.

[1]X-Fact dataset under MIT License on GitHub (`https://github.com/utahnlp/x-fact`)

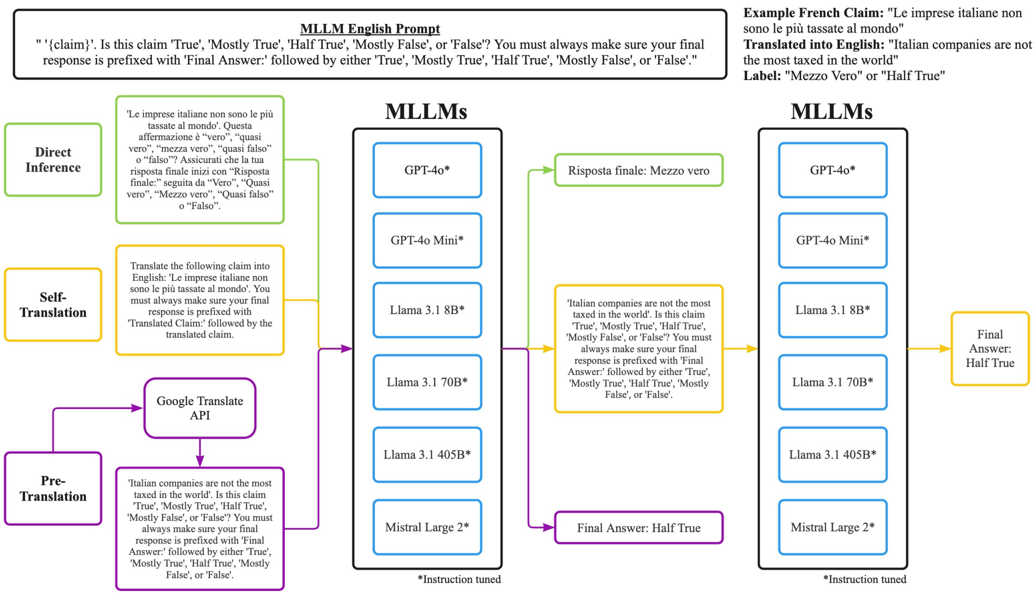

Figure 1: Flowchart illustrating the process for evaluating the claim verification performance of LLMs using Direct Inference, Self-Translation, and Pre-Translation.

3.2 Multilingual Language Models

Each of the LLMs used in our experiments is instruction-tuned. We conduct our experiments on OpenAI's GPT-4o[2] and GPT-4o Mini[3] models, Mistral's Mistral Large 2[4] model with 123B parameters, Meta's Llama 3.1 models with 8B, 70B, and 405B parameters [Dubey et al., 2024], and a fine-tuned version of Google's mBERT multilingual model, following the same training process used by Gupta and Srikumar [2021]. All of the models are pre-trained on multilingual corpora. For each model, we set the temperature to 0 for reproducibility. Each model automatically determined the default token length based on the number of tokens required to complete its output according to its respective context length.

3.3 Evaluation

For each experiment, we record the number of correct, incorrect, and inconclusive responses returned by the model. We express the accuracy score of the LLM as the percentage of correct answers.

3.4 Translation Techniques

We employ the following translation methods when evaluating each model's performance on a language family:

Direct Inference is completing a task in the native language of the prompt without performing any translations. This method is intended to measure the model's ability to understand and generate text in the target language without relying on cross-linguistic skills, thereby isolating its performance on monolingual tasks. Inconclusive outputs in this method occur when the model fails to provide a conclusive answer (e.g., "True," "Mostly True," etc.) as required by the prompt, though the risk of faulty translations is minimized since no external translations are involved.

Self-Translate Etxaniz et al. [2023] involves an LLM performing a translation task itself without relying on external translation services. This technique allows the model to leverage its inherent

[2] `https://openai.com/index/hello-gpt-4o/`

[3] `https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/`

[4] `https://mistral.ai/news/mistral-large-2407/`

multilingual capabilities, effectively using its own understanding of multiple languages to translate text autonomously. For consistent comparisons, we translate into English. This decision reflects the fact that most LLMs are trained predominantly on English data, making it a reasonable default for translation tasks, since translation into less well-trained languages is unlikely to yield better results due to the scarcity of high-quality training data in those languages [Dong et al., 2024, HAI, 2023]. The translation and claim verification steps are conducted in two separate chat sessions, ensuring that context is not preserved between them. This approach allows us to consistently assess the LLM's inherent translation ability, independent of any contextual memory. Inconclusive responses in this method occur if the model fails to properly translate the claim or does not follow the prompt's instructions, resulting in incorrect or incomplete outputs.

Pre-Translate Intrator et al. [2024] involves the use of third-party translation services external to the model, rather than relying on the model's own translation capabilities. Following the approach outlined by Intrator et al. [2024], we use the Google Translate API[5] for this purpose. For consistent comparisons, we translate into English. In this method, inconclusive outputs can arise when there are inaccuracies in translation, which may lead the model to misinterpret the claim and provide an unclear or incorrect answer.

The process for evaluating claim verification performance using Direct Inference, Self-Translation, and Pre-Translation is outlined in the flowchart shown in Figure 1.

3.5 Translation Bias

We assess translation bias using the COMETKIWI model from Rei et al. [2022], which allows for the evaluation of machine translations without requiring reference translations. A reference translation is a pre-existing human translation of a source text that serves as a benchmark for evaluating the accuracy and quality of a machine translation.

The Translation Bias (TB) quantifies the overall quality of machine translations by leveraging the scores from the COMETKIWI model. Given a set of M COMET scores, scores $= \{s_1, s_2, \ldots, s_M\}$, the Translation Bias is calculated as:

$$\text{TB} = 1 - \frac{1}{M}\sum_{j=1}^{M} s_j$$

4 Results and Discussion

4.1 Language-Specific Trends

4.1.1 High- and Low-Resource Languages

Direct inference demonstrated significantly higher accuracy in the Romance, Slavic, and Turkic language families compared to other translation techniques. These families generally consist of high- or moderately high-resource languages, which have abundant data and represent a larger portion of the training data for the models. In contrast, for the Kartvelian and Indo-Aryan language families—mostly low-resource languages—the performance of direct inference was consistently equal to or worse than other translation methods. This suggests that direct inference may be less effective for low-resource languages due to limited training data, resulting in poorer model understanding and higher error rates.

4.1.2 Performance of English

Despite being the most represented language in the training data, English was sometimes outperformed by other languages, possibly because the English claims in the evaluation dataset are more niche and complex—often including a higher proportion of political claims—which may lower accuracy as models struggle with more intricate statements. For instance, Llama 3.1 405B showed higher accuracy for Slavic languages (36.00%) compared to English (33.50%), even though English is typically better resourced.

[5] `https://py-googletrans.readthedocs.io/en/latest/`

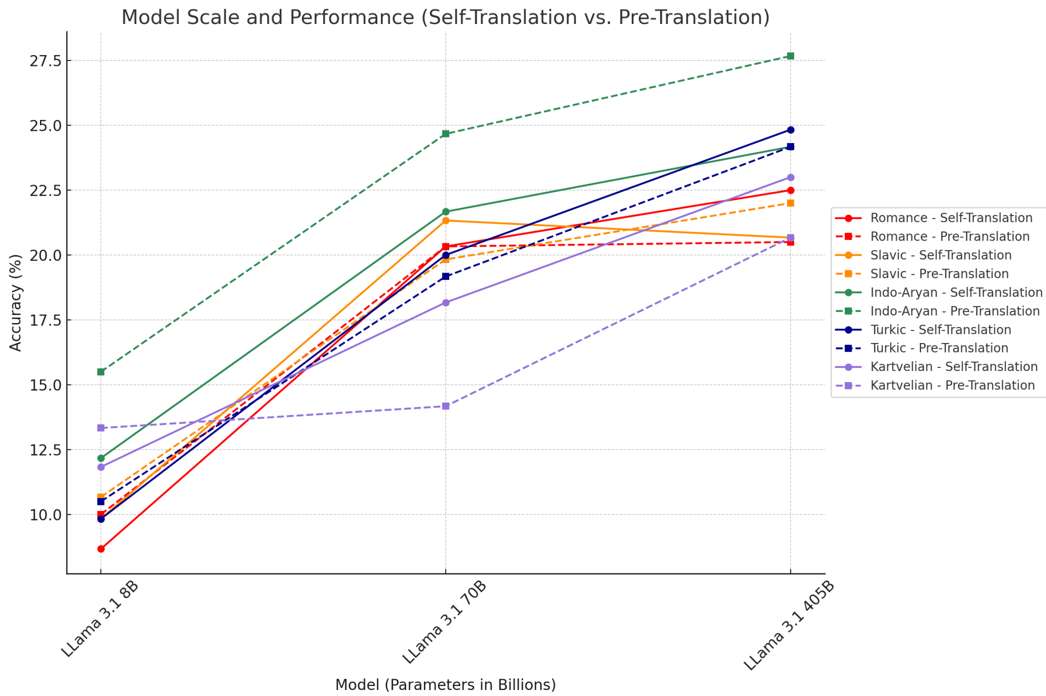

Figure 2: Accuracy performance of Llama 3.1 models across different language families using Self-Translation and Pre-Translation techniques.

4.2 Translation Techniques

While self-translation and pre-translation techniques generally yielded lower accuracy compared to direct inference, they reduced the number of inconclusive results by enhancing LLM comprehension and likely reducing misinterpretations, particularly for complex or nuanced claims. Nonetheless, the accuracy of both translation methods remained lower than that of direct inference.

4.2.1 Self-Translation vs. Pre-Translation

Self-translation performs slightly better than pre-translation which we believe is attributed to the model maintaining internal consistency between generating and verifying translations. When the LLM handles both tasks, its linguistic patterns are more likely to align, reducing interpretation errors. Pre-translation, however, relies on external services that can introduce inconsistencies, leading to more misinterpretations during the verification phase. As a result, pre-translation produced more inconclusive outputs and had lower accuracy than self-translation.

4.3 Model Scale

Looking at Figure 2, smaller models like Llama 3.1 8B perform poorly for both self-translation and pre-translation across all language families, with self-translation slightly outperforming pre-translation. However, as model size increased, the accuracy of self-translation improved significantly. For instance, Llama 3.1 405B demonstrated improved performance across Romance, Slavic, Turkic, and Indo-Aryan languages, surpassing pre-translation in all cases.

Interestingly, although self-translation performed better with larger models, the translation bias scores remained relatively stable, suggesting that increased model size improves accuracy but not fairness across languages. For example, Llama 3.1 405B maintained similar bias scores to smaller models like Llama 3.1 8B, indicating that the increased size of the model improves accuracy but not fairness in translation. A detailed breakdown of the translation bias scores for each method, model, and language family is provided in Appendix A.3.

5 Conclusion

This study examines the translation bias and accuracy of multilingual Large Language Models (LLMs) in cross-language claim verification tasks across five language families. Our findings demonstrate that direct inference performs better in high-resource languages, while self-translation and pre-translation techniques handle low-resource languages more effectively, though with reduced accuracy. Furthermore, as model size increases, the accuracy of self-translation improves, yet translation bias remains consistent across all models, showing that larger models do not necessarily ensure fairness across languages. These results highlight the persistent challenges in achieving equitable multilingual capabilities in LLMs. By identifying specific areas where translation biases occur, we lay the groundwork for developing more balanced and fair language technologies.

Limitations

Our study of language and translation biases in LLMs for cross-lingual claim verification has several limitations. We used the 2021 X-Fact dataset, which may not reflect the most recent language trends or advancements in model capabilities as of 2024. Additionally, the LLMs tested may have been trained on datasets overlapping with X-Fact, potentially inflating performance metrics. While we focused on 15 languages from diverse families, this selection might not fully represent the linguistic diversity needed to capture trends in low-resource languages. Our evaluation was limited to translations from non-English languages into English, and while examining other language pairs might provide valuable insights, it is unlikely that these pairings would outperform English due to the prevalent training bias toward English data in most LLMs. We used a lighter, older version of the COMETKIWI model to assess translation bias due to computational limitations, which may affect the robustness of our bias measurements. Moreover, we did not compare baseline models with instruction-tuned versions, which could have reduced inconclusive translations and offered further insights into model performance. We also did not incorporate reference translations or employ evidence retrieval, which could have provided a more holistic evaluation of translation quality. Future work should expand to include more recent datasets, evaluate other language pairs, and refine the methods to enhance bias detection and accuracy.

Ethics Statement

This study investigates translation bias in multilingual Large Language Models (LLMs), focusing on disparities across high- and low-resource languages. Our findings highlight that these biases disproportionately affect low-resource languages, potentially leading to misinformation propagation in underrepresented linguistic communities. We acknowledge the potential ethical risks associated with the reliance on LLMs for cross-language claim verification, particularly the unequal access to accurate information. Future work should focus on more balanced model training to mitigate these risks, ensuring fairer outcomes for all language speakers. Additionally, we emphasize the need for collaboration with native speakers and ethical oversight in model development to ensure inclusivity in global language technologies.

References

Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent Ochieng, Rishav Hada, Prachi Jain, Maxamed Axmed, Kalika Bali, and Sunayana Sitaram. Megaverse: Benchmarking large language models across languages, modalities, models and tasks, 2024. URL https://arxiv.org/abs/2311.07463.

Rrubaa Panchendrarajan and Arkaitz Zubiaga. Claim detection for automated fact-checking: A survey on monolingual, multilingual and cross-lingual research. *Natural Language Processing Journal*, 7:100066, June 2024. ISSN 2949-7191. doi: 10.1016/j.nlp.2024.100066. URL http://dx.doi.org/10.1016/j.nlp.2024.100066.

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Yuqi Ye, and Hanwen Gu. A survey on multilingual large language models: Corpora, alignment, and bias, 2024. URL https://arxiv.org/abs/2404.00929.

Kaiyu Huang, Fengran Mo, Hongliang Li, You Li, Yuanchi Zhang, Weijian Yi, Yulong Mao, Jinchen Liu, Yuzhuang Xu, Jinan Xu, Jian-Yun Nie, and Yang Liu. A survey on large language models with multilingualism: Recent advances and new frontiers, 2024. URL https://arxiv.org/abs/2405.10936.

Megha Sundriyal, Tanmoy Chakraborty, and Preslav Nakov. From chaos to clarity: Claim normalization to empower fact-checking. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6594–6609, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.439. URL https://aclanthology.org/2023.findings-emnlp.439.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. ChatGPT MT: Competitive for high- (but not low-) resource languages. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.40. URL https://aclanthology.org/2023.wmt-1.40.

Rachel Bawden and François Yvon. Investigating the translation performance of a large multilingual language model: the case of BLOOM. In Mary Nurminen, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, Eva Vanmassenhove, Sergi Alvarez Vidal, Nora Aranberri, Mara Nunziatini, Carla Parra Escartín, Mikel Forcada, Maja Popovic, Carolina Scarton, and Helena Moniz, editors, *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 157–170, Tampere, Finland, June 2023. European Association for Machine Translation. URL https://aclanthology.org/2023.eamt-1.16.

Dorian Quelle and Alexandre Bovet. The perils and promises of fact-checking with large language models. *Frontiers in Artificial Intelligence*, 7, February 2024. ISSN 2624-8212. doi: 10.3389/frai.2024.1341697. URL http://dx.doi.org/10.3389/frai.2024.1341697.

Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ali Payani, Ninghao Liu, and Mengnan Du. Quantifying multilingual performance of large language models across languages, 2024. URL https://arxiv.org/abs/2404.11553.

James Thorne and Andreas Vlachos. Automated fact checking: Task formulations, methods and future directions. In Emily M. Bender, Leon Derczynski, and Pierre Isabelle, editors, *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3346–3359, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL https://aclanthology.org/C18-1283.

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 Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, 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, 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. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Caiqi Zhang, Zhijiang Guo, and Andreas Vlachos. Do we need language-specific fact-checking models? the case of chinese, 2024. URL https://arxiv.org/abs/2401.15498.

Ashim Gupta and Vivek Srikumar. X-fact: A new benchmark dataset for multilingual fact checking, 2021. URL https://arxiv.org/abs/2106.09248.

Kung-Hsiang Huang, ChengXiang Zhai, and Heng Ji. CONCRETE: Improving cross-lingual fact-checking with cross-lingual retrieval. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1024–1035, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.86.

Xuming Hu, Junzhe Chen, Xiaochuan Li, Yufei Guo, Lijie Wen, Philip S. Yu, and Zhijiang Guo. Do large language models know about facts?, 2023. URL https://arxiv.org/abs/2310.05177.

Guoliang Dong, Haoyu Wang, Jun Sun, and Xinyu Wang. Evaluating and mitigating linguistic discrimination in large language models, 2024. URL https://arxiv.org/abs/2404.18534.

Stanford HAI. Improving equity and access with non-english large language models, 2023. URL https://hai.stanford.edu/news/improving-equity-and-access-non-english-large-language-models. Accessed: September 24, 2024.

Sheikh Shafayat, Eunsu Kim, Juhyun Oh, and Alice Oh. Multi-fact: Assessing multilingual llms' multi-regional knowledge using factscore, 2024. URL https://arxiv.org/abs/2402.18045.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. Do multilingual language models think better in english?, 2023. URL https://arxiv.org/abs/2308.01223.

Yotam Intrator, Matan Halfon, Roman Goldenberg, Reut Tsarfaty, Matan Eyal, Ehud Rivlin, Yossi Matias, and Natalia Aizenberg. Breaking the language barrier: Can direct inference outperform pre-translation in multilingual llm applications?, 2024. URL https://arxiv.org/abs/2403.04792.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.60.

A Appendix

A.1 Dataset Distribution

Table 1: Distribution of Romance language family claims.

Language	False	Half True	Mostly False	True	Total Claims
French (fr)	30	30	0	21	109
Italian (it)	34	36	0	37	139
Spanish (es)	34	34	0	37	136
Portuguese (pt)	80	43	1	50	216

Table 2: Distribution of Slavic language family claims.

Language	False	Half True	Mostly False	True	Total Claims
Serbian (sr)	70	42	44	45	234
Russian (ru)	50	51	1	42	153
Polish (pl)	64	50	0	99	213

Table 3: Distribution of Turkic language family claims.

Language	False	Half True	Mostly False	True	Total Claims
Turkish (tr)	60	63	82	96	407
Azerbaijani (az)	60	57	38	24	193

Table 4: Distribution of Indo-Aryan language family claims.

Language	False	Half True	Mostly False	True	Total Claims
Bengali (bn)	36	35	91	1	163
Hindi (hi)	89	57	118	0	264
Marathi (mr)	26	26	0	0	52
Punjabi (pa)	25	40	0	0	65
Gujarati (gu)	27	29	0	0	56

Table 5: Distribution of Kartvelian language family claims.

Language	False	Half True	Mostly False	True	Total Claims
Georgian (ka)	120	120	120	120	600

A.2 Model Performance Across Language Families

Table 6: Performance distribution of LLMs using direct inference on English claims.

Model	Total Correct	Total Incorrect	Total Inconclusive	Accuracy
GPT-4o	215	377	8	**35.83%**
GPT-4o Mini	185	413	2	30.83%
Mistral Large 2	183	403	14	30.50%
Llama 3.1 8B	95	278	227	15.83%
Llama 3.1 70B	166	353	81	27.67%
Llama 3.1 405B	201	386	13	33.50%
mBERT	95	340	165	15.83%

Table 7: Performance distribution of LLMs using direct inference, self-translate, and pre-translate on Romance claims.

Model	Technique	Total Correct	Total Incorrect	Total Inconclusive	Accuracy
GPT-4o	Direct Inference	185	381	34	30.83%
GPT-4o	Self-Translation	174	396	30	29.00%
GPT-4o	Pre-Translation	150	413	37	25.00%
GPT-4o Mini	Direct Inference	197	388	15	**32.83%**
GPT-4o Mini	Self-Translation	165	434	1	**27.50%**
GPT-4o Mini	Pre-Translation	154	445	1	25.67%
Mistral Large 2	Direct Inference	155	405	40	25.83%
Mistral Large 2	Self-Translation	123	422	55	20.50%
Mistral Large 2	Pre-Translation	97	386	117	16.17%
Llama 3.1 8B	Direct Inference	126	389	85	21.00%
Llama 3.1 8B	Self-Translation	52	236	312	8.67%
Llama 3.1 8B	Pre-Translation	60	296	244	10.00%
Llama 3.1 70B	Direct Inference	172	398	30	28.67%
Llama 3.1 70B	Self-Translation	122	303	175	20.33%
Llama 3.1 70B	Pre-Translation	122	301	177	20.33%
Llama 3.1 405B	Direct Inference	191	404	5	**31.83%**
Llama 3.1 405B	Self-Translation	135	422	43	22.50%
Llama 3.1 405B	Pre-Translation	123	427	50	20.50%
mBERT	Direct Inference	166	255	179	27.67%
mBERT	Pre-Translation	106	444	50	17.67%

Table 8: Performance distribution of LLMs using direct inference, self-translate, and pre-translate on Slavic claims.

Model	Technique	Total Correct	Total Incorrect	Total Inconclusive	Accuracy
GPT-4o	Direct Inference	199	315	86	33.17%
GPT-4o	Self-Translation	195	384	21	**32.50%**
GPT-4o	Pre-Translation	161	404	35	26.83%
GPT-4o Mini	Direct Inference	206	334	60	**34.33%**
GPT-4o Mini	Self-Translation	135	465	0	22.50%
GPT-4o Mini	Pre-Translation	139	461	0	23.17%
Mistral Large 2	Direct Inference	177	298	125	29.50%
Mistral Large 2	Self-Translation	123	439	38	20.50%
Mistral Large 2	Pre-Translation	102	423	75	17.00%
Llama 3.1 8B	Direct Inference	121	250	229	20.17%
Llama 3.1 8B	Self-Translation	59	253	288	9.83%
Llama 3.1 8B	Pre-Translation	64	280	256	10.67%
Llama 3.1 70B	Direct Inference	177	290	133	29.50%
Llama 3.1 70B	Self-Translation	128	357	115	21.33%
Llama 3.1 70B	Pre-Translation	119	388	93	19.83%
Llama 3.1 405B	Direct Inference	216	353	31	36.00%
Llama 3.1 405B	Self-Translation	124	468	8	20.67%
Llama 3.1 405B	Pre-Translation	132	454	14	**22.00%**
mBERT	Direct Inference	79	251	270	13.17%
mBERT	Pre-Translation	130	411	59	21.67%

Table 9: Performance distribution of LLMs using direct inference, self-translate, and pre-translate on Indo-Aryan claims.

Model	Technique	Total Correct	Total Incorrect	Total Inconclusive	Accuracy
GPT-4o	Direct Inference	150	425	25	25.00%
GPT-4o	Self-Translation	180	360	60	**30.00%**
GPT-4o	Pre-Translation	157	346	97	26.17%
GPT-4o Mini	Direct Inference	190	431	0	**28.17%**
GPT-4o Mini	Self-Translation	144	434	0	27.67%
GPT-4o Mini	Pre-Translation	171	429	0	**28.50%**
Mistral Large 2	Direct Inference	85	281	234	14.17%
Mistral Large 2	Self-Translation	173	364	63	28.83%
Mistral Large 2	Pre-Translation	146	300	154	24.33%
Llama 3.1 8B	Direct Inference	95	278	227	15.83%
Llama 3.1 8B	Self-Translation	73	192	335	12.17%
Llama 3.1 8B	Pre-Translation	93	222	285	15.50%
Llama 3.1 70B	Direct Inference	127	426	47	21.17%
Llama 3.1 70B	Self-Translation	130	344	126	21.67%
Llama 3.1 70B	Pre-Translation	148	321	131	24.67%
Llama 3.1 405B	Direct Inference	166	358	76	27.67%
Llama 3.1 405B	Self-Translation	143	379	76	24.17%
Llama 3.1 405B	Pre-Translation	166	358	76	27.67%
mBERT	Direct Inference	81	279	240	13.50%
mBERT	Pre-Translation	83	281	216	17.17%

Table 10: Performance distribution of LLMs using direct inference, self-translate, and pre-translate on Turkic claims.

Model	Technique	Total Correct	Total Incorrect	Total Inconclusive	Accuracy
GPT-4o	Direct Inference	159	437	4	26.50%
GPT-4o	Self-Translation	150	416	34	25.00%
GPT-4o	Pre-Translation	141	427	32	23.50%
GPT-4o Mini	Direct Inference	130	469	1	21.67%
GPT-4o Mini	Self-Translation	147	452	1	**24.50%**
GPT-4o Mini	Pre-Translation	135	462	3	22.50%
Mistral Large 2	Direct Inference	129	469	2	21.50%
Mistral Large 2	Self-Translation	123	418	59	20.50%
Mistral Large 2	Pre-Translation	111	396	93	18.50%
Llama 3.1 8B	Direct Inference	106	454	40	17.67%
Llama 3.1 8B	Self-Translation	59	247	294	9.83%
Llama 3.1 8B	Pre-Translation	63	307	230	10.50%
Llama 3.1 70B	Direct Inference	131	443	26	21.83%
Llama 3.1 70B	Self-Translation	120	359	121	20.00%
Llama 3.1 70B	Pre-Translation	115	379	106	19.17%
Llama 3.1 405B	Direct Inference	154	445	1	**25.67%**
Llama 3.1 405B	Self-Translation	149	432	19	24.83%
Llama 3.1 405B	Pre-Translation	145	439	16	24.17%
mBERT	Direct Inference	98	331	171	16.33%
mBERT	Pre-Translation	109	478	13	**18.17%**

Table 11: Performance distribution of LLMs using direct inference, self-translate, and pre-translate on Kartvelian claims.

Model	Technique	Total Correct	Total Incorrect	Total Inconclusive	Accuracy
GPT-4o	Direct Inference	28	503	69	4.67%
GPT-4o	Self-Translation	131	423	46	21.83%
GPT-4o	Pre-Translation	127	442	31	21.17%
GPT-4o Mini	Direct Inference	38	559	3	6.33%
GPT-4o Mini	Self-Translation	138	459	3	**23.00%**
GPT-4o Mini	Pre-Translation	132	465	3	**22.00%**
Mistral Large 2	Direct Inference	42	303	255	7.00%
Mistral Large 2	Self-Translation	118	404	78	19.67%
Mistral Large 2	Pre-Translation	107	386	107	17.83%
Llama 3.1 8B	Direct Inference	29	135	436	4.83%
Llama 3.1 8B	Self-Translation	71	236	293	11.83%
Llama 3.1 8B	Pre-Translation	80	267	253	13.33%
Llama 3.1 70B	Direct Inference	55	511	34	9.17%
Llama 3.1 70B	Self-Translation	109	336	155	18.17%
Llama 3.1 70B	Pre-Translation	85	313	202	14.17%
Llama 3.1 405B	Direct Inference	0	598	0	0.00%
Llama 3.1 405B	Self-Translation	138	435	27	**23.00%**
Llama 3.1 405B	Pre-Translation	124	439	37	20.67%
mBERT	Direct Inference	133	463	4	**22.17%**
mBERT	Pre-Translation	99	398	103	16.50%

A.3 Translation Bias Scores

Table 12: Translation bias scores across Romance, Slavic, Turkic, Indo-Aryan, and Kartvelian language families using the pre-translation technique (Google Translate API).

Language Family	Translation Bias Score
Romance	0.33
Slavic	0.35
Turkic	0.11
Indo-Aryan	0.22
Kartvelian	0.22

Table 13: Translation bias scores for LLMs across Romance, Slavic, Turkic, Indo-Aryan, and Kartvelian language families using the self-translation technique.

Model	Language Family	Translation Bias Score
GPT-4o	Romance	0.16
GPT-4o	Slavic	0.16
GPT-4o	Turkic	0.17
GPT-4o	Indo-Aryan	0.16
GPT-4o	Kartvelian	0.16
GPT-4o Mini	Romance	0.16
GPT-4o Mini	Slavic	0.16
GPT-4o Mini	Turkic	0.17
GPT-4o Mini	Indo-Aryan	0.17
GPT-4o Mini	Kartvelian	0.16
Mistral Large 2	Romance	0.16
Mistral Large 2	Slavic	0.16
Mistral Large 2	Turkic	0.17
Mistral Large 2	Indo-Aryan	0.17
Mistral Large 2	Kartvelian	0.17
Llama 3.1 8B	Romance	0.18
Llama 3.1 8B	Slavic	0.19
Llama 3.1 8B	Turkic	0.20
Llama 3.1 8B	Indo-Aryan	0.19
Llama 3.1 8B	Kartvelian	0.21
Llama 3.1 70B	Romance	0.16
Llama 3.1 70B	Slavic	0.17
Llama 3.1 70B	Turkic	0.18
Llama 3.1 70B	Indo-Aryan	0.18
Llama 3.1 70B	Kartvelian	0.19
Llama 3.1 405B	Romance	0.16
Llama 3.1 405B	Slavic	0.16
Llama 3.1 405B	Turkic	0.17
Llama 3.1 405B	Indo-Aryan	0.16
Llama 3.1 405B	Kartvelian	0.16

A.4 Code Repository

The code used in our experiments and for generating the results presented in this paper can be accessed at the following GitHub repository:
`https://github.com/3x-dev/Comparative-Study-of-Bias-and-Accuracy-in-Multilingual-LLMs-for-Cross-Language-Claim-Verification`

A.5 Compute Resources

The experiments were conducted using a combination of MacBook Pros and a dedicated GPU cluster for pre-training the mBERT model. Below are the general specifications for each setup:

GPU Resources: The mBERT pre-training was performed on a GPU cluster equipped with NVIDIA A100 Tensor Core GPUs (40 GB VRAM) for high performance training. Inference and other experiments performed on MacBook Pros did not use GPUs because MacBook Pros do not have discrete GPUs suitable for machine learning tasks.

CPU Resources: Experiments run on MacBook Pros used Apple's **M1 Pro** or **M1 Max** processors (8- to 10-core CPUs), and some collaborators used **Intel Core i9** processors (8-core) in older MacBook Pro models. These CPU configurations were sufficient for smaller experiments and model inference tasks.

Memory: MacBook Pro memory capacity ranged from **16GB to 64GB of unified memory** on Apple Silicon (M1) models to **32GB of DDR4 RAM** on Intel-based MacBook Pros. These configurations were sufficient for model inference, but could limit performance with larger models and datasets.

Storage: Experiments conducted on MacBook Pros used **SSD storage ranging from 512GB to 2TB**. Local storage was used to manage smaller datasets and model checkpoints. For larger datasets and models, external storage or cloud services were used to mitigate local storage limitations.

Pre-training and Inference Times:

- **Pre-training:** Pre-training mBERT on the GPU cluster with NVIDIA A100 GPUs took approximately **12 hours** using 4 GPUs in parallel. This was essential to ensure the mBERT model was fine-tuned for multilingual tasks.
- **Inference:** Inference on the MacBook Pros varied depending on model size. For smaller models like GPT-4 Mini, inference times ranged between **3 to 5 hours** per language family. However, larger models like Llama 3.1 405B were run in a distributed fashion, with inference times extending to **8 to 10 hours** due to limited hardware.

Total Computing Time: The total computation time for all experiments, including pre-training, tuning, and inference, was approximately **150 GPU hours** on the cluster for pre-training and **100 CPU hours** on MacBook Pros for inference and evaluation.

Considerations for Reproducibility: Replicating these results on similar hardware, such as MacBook Pros with M1/M2 chips or Intel processors, should result in longer computation times, especially for larger models. For pre-training or large-scale fine-tuning, access to a GPU cluster or cloud-based GPU services is recommended.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the focus on evaluating language and translation biases in LLMs for cross-lingual claim verification. The claims align with the experiments and analysis conducted, and the scope of the study is appropriately represented. See Abstract and Section 1.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper contains a dedicated limitations section, which outlines key constraints such as dataset recency, reliance on English-centric training data, the exclusion of non-English language pairs, and the use of an older COMETKIWI model for bias evaluation. See the Limitations section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper focuses on empirical results and does not present theoretical results or proofs.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The experimental setup is well documented with details on datasets, models, evaluation techniques, and hyperparameters. The methods section (Section 3) provides enough information to reproduce the experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The paper links to a GitHub repository containing the code and scripts used in the experiments. See Appendix A.4.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper outlines the datasets, models, evaluation techniques, and hyperparameters, allowing a clear understanding of the experimental results. Details on the dataset splits, language families, and claim distribution are provided in Section 3 and Appendix A.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper does not report error bars or confidence intervals. While the results are based on accuracy measurements, the paper does not include formal statistical significance tests.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper specifies the compute resources used, including MacBook Pros with Apple M1 Pro/Max and Intel Core i9 processors for inference, and a GPU cluster with NVIDIA A100 GPUs for pre-training mBERT, along with estimates of compute time (150 GPU-hours for pre-training and 100 CPU-hours for inference). Please see Appendix A.5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper conforms to the NeurIPS Code of Ethics. It addresses the impact of translation bias on low-resource languages and emphasizes the importance of fairer language technologies. The paper's conclusions are in line with ethical guidelines.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper highlights the positive societal impact of reducing bias in multilingual LLMs, which can improve the accessibility of accurate information. It also discusses the risks of spreading misinformation in low-resource languages. See Ethics Statement.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper does not involve the release of potentially harmful models or data, and thus this question does not apply.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

Justification: The paper references all datasets and models used, such as the X-Fact dataset and pre-existing LLMs (e.g., GPT-4, mBERT), with proper citations. See Section 3.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not introduce new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing or human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.