LANGUAGE-DEPENDENT MISCALIBRATION IN MULTILINGUAL LLM EVALUATORS

Ej Zhou, Lucas Resck, Zheng Hui & Anna Korhonen
Language Technology Lab, University of Cambridge

ABSTRACT

Prompted LLM-as-a-Judge systems or trained reward models are typically validated using pairwise accuracy, under the assumption that high accuracy implies reliable and language-invariant evaluation. We demonstrate that multilingual LLM evaluators exhibit large, systematic, and statistically significant language-dependent bias in pointwise scoring. We show that this mismatch has concrete downstream consequences: threshold filtering can result in huge differences in acceptance rates.

1 INTRODUCTION

Large language models (LLMs) are increasingly evaluated not only by humans, but by other language models (Zheng et al., 2023; Kocmi & Federmann, 2023). *LLM-based evaluation* has become central to modern NLP research and deployment (Lou et al., 2025), underpinning model selection, reinforcement learning from human feedback (RLHF) and safety auditing. In practice, such evaluation is carried out by a diverse family of *LLM evaluators*, including prompted LLM-as-a-Judge systems and trained reward models that assign scalar quality scores or preference rankings to model outputs.

As LLMs become multilingual by default, their evaluators are likewise expected to operate reliably across languages. Implicitly there is a strong assumption: that evaluation scores are *language-invariant*, meaning that semantically identical content should receive comparable judgments regardless of the language in which it is expressed. In this work, we show that this assumption does not hold.

We present an empirical study demonstrating that **multilingual LLM evaluators exhibit consistent and statistically significant language-dependent bias in pointwise scoring**. When asked to evaluate semantically identical instruction–response pairs across languages, both prompted LLM judges and trained multilingual reward models assign markedly different absolute scores depending solely on the evaluation language. Crucially, these differences persist across model families, architectures, and training paradigms, indicating that language acts as a latent variable influencing evaluation independently of content quality.

At the same time, this bias is largely *invisible* under standard evaluation practices. A common approach to validating multilingual evaluators is to rely on *pairwise accuracy*, measuring whether accepted responses are ranked above rejected ones. We show that pairwise accuracy remains uniformly high and stable across languages even when pointwise scores differ substantially. As a result, evaluators may appear well-aligned and robust under pairwise metrics while exhibiting severe language-dependent miscalibration in absolute scoring.

This disconnect highlights a fundamental limitation of current evaluation protocols. Pairwise metrics capture relative ordering but are insensitive to systematic shifts in score distributions. Consequently, they fail to detect biases that directly affect downstream uses of evaluators, such as threshold-based filtering, reward shaping, calibration-sensitive training, and cross-language comparison.

2 EXPERIMENTS

2.1 DATASET

Our experiments are conducted on REWARDBENCH (Lambert et al., 2025) and its human-validated multilingual extension M-REWARDBENCH (Gureja et al., 2025). Together, these benchmarks provide

aligned instruction–response pairs across 23 languages and multiple task categories, including *Chat*, *Chat-Hard*, *Reasoning*, and *Safety*. We focus on unchosen responses, which exhibit broader score distributions and avoid ceiling effects that obscure language-dependent variation (Figure 5). Our evaluation spans 23 languages (see Table 1).

2.2 Multilingual Evaluators

All evaluators are applied in a **zero-shot** setting without task-specific fine-tuning. **Prompted LLM Judges (LLM-as-a-Judge):** These evaluators are general-purpose multilingual LLMs prompted with standardized evaluation rubrics to directly assign scalar quality scores. We include Aya Expanse (Dang et al., 2024), Qwen 2.5 (Yang et al., 2025), LLaMA 3.1 (Grattafiori et al., 2024), and M-Prometheus (Pombal et al., 2025). Notably, M-Prometheus is explicitly trained for multilingual evaluation. We use a standardized scoring rubric requesting a **1–5 Likert-scale** rating in each target language. Decoding temperature is fixed to zero. **Multilingual Reward Models:** We additionally evaluate trained multilingual reward models that output scalar scores directly. Specifically, we include URM-LLaMA-3.1-8B (Lou et al., 2025) and BTRM-Qwen-2-7B, Skywork-Reward-Gemma-2-27B, Skywork-Reward-Llama-3.1-8B-v0.2 (Liu et al., 2024). We record *pointwise scalar scores*.

2.3 Results

2.3.1 Language-Dependent Scoring Bias

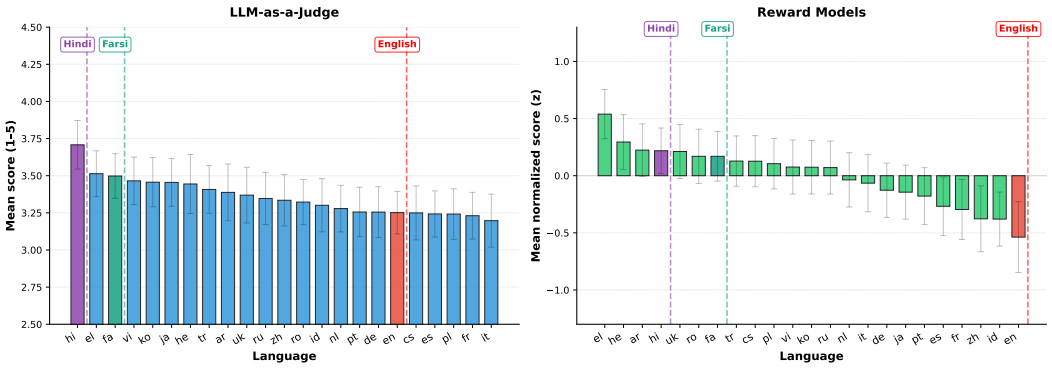

Figure 1: Left: mean pointwise scores assigned by prompted LLM-as-a-Judge models (1–5 scale), averaged across evaluators; right: mean z-normalized scores from trained multilingual reward models. Error bars indicate inter-model variability.

Figure 1 provides a consolidated view of language-dependent scoring behavior across multilingual evaluators. Despite evaluating *identical semantic content*, both prompted LLM-as-a-Judge systems (left) and trained multilingual reward models (right) assign systematically different scores depending solely on the evaluation language.

For prompted LLM judges, we observe substantial variation in absolute pointwise scores across languages. Mean scores span a wide range on the 1–5 scale, with languages exhibiting a highly stable *global ordering* across evaluators and task categories - Hindi, Greek, and Hebrew consistently receiving the highest evaluations, while several Western European languages—including Italian, French, and Spanish—occupy the lower end of the distribution. Notably, these differences are large relative to the effective score range used by the judges, corresponding to shifts of approximately 0.4–0.5 points for semantically identical responses. Detailed per-language and per-model averages are reported in Table 2 (Appendix A).

Crucially, the relative ordering of languages is highly stable across evaluator families. As shown in Figure 1, the language ranking induced by prompted LLM judges closely mirrors that produced by trained reward models, even after z-normalization removes differences in score scale. Languages that receive higher pointwise scores under LLM-as-a-Judge evaluation also tend to receive positive normalized scores under reward models, while lower-scoring languages remain systematically penalized. Full normalized statistics for reward models are provided in Table 3 (Appendix A). English

does not occupy a privileged or neutral position in either setting. Under prompted evaluation, English lies below the median score across languages, and under reward models it is associated with a clearly negative normalized score.

All observed language effects are statistically significant for every prompted LLM judge and reward model tested (one-way ANOVA, $p < 0.001$ in all cases; see Table 4 for full results).

2.3.2 Cross-Model Consistency

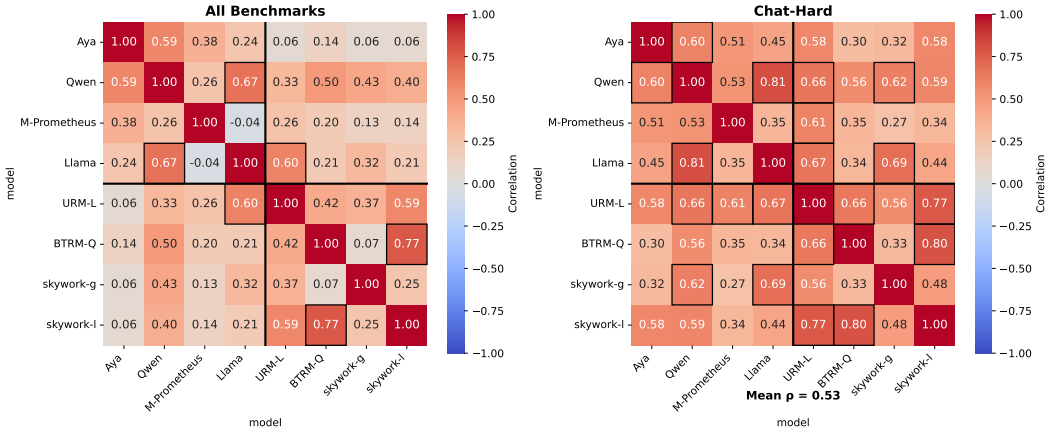

Figure 2: Cross-model alignment of language-dependent scoring patterns. Each cell shows the Pearson correlation between per-language mean scores assigned by two evaluators. Left: correlations aggregated across all benchmarks. Right: correlations computed on the Chat-Hard subset only.

To determine whether language-dependent bias is evaluator-specific or systematic, we compute pairwise Pearson correlations between per-language mean scores across all evaluators. As shown in Figure 6, correlations range from moderate to strong (approximately 0.36–0.88), both when aggregating across all benchmarks and when analyzing individual task categories.

These strong correlations indicate that evaluators largely agree on which languages are scored higher or lower, despite differences in architecture, training objective, and score scale. Together, these results suggest that language-dependent scoring bias reflects a *shared inductive bias across multilingual LLM evaluators*, rather than idiosyncratic behavior of individual models.

Taken together, these results demonstrate that multilingual LLM evaluators encode language-conditioned scoring behavior that is systematic, statistically robust, and shared across models, yet remains undetectable under standard pairwise validation protocols.

2.4 Resource Level and Writing Script

A natural hypothesis is that language-dependent scoring bias reflects differences in training data availability or writing systems. To examine this, Figure 3 relates per-language evaluation scores to estimated language resource levels, using the number of Common Crawl pages as a proxy. For prompted LLM judges (Figure 3, left), we observe only a weak association between resource availability and mean score. Trained reward models exhibit a markedly stronger and more consistent relationship between resource level and score (Figure 3, right). Here, both Pearson and Spearman correlations are substantially larger (Pearson $r = -0.58$, Spearman $\rho = -0.81$), indicating a strong monotonic decrease in reward scores as resource availability increases. Under reward-model evaluation, high-resource languages are systematically assigned lower normalized scores, while lower-resource languages are scored more favorably.

2.5 Pairwise Accuracy Masks Language-Dependent Decision Bias

Pairwise accuracy is the dominant validation metric for multilingual reward models and LLM-based evaluators, measuring whether preferred responses are ranked above rejected ones. As shown in

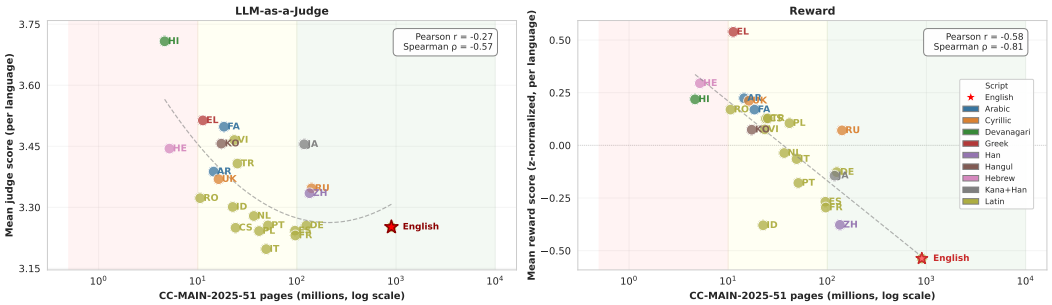

Figure 3: Relationship between language evaluation scores and training data availability. Each point represents a language, plotted by its mean evaluation score against the estimated number of `CC-MAIN-2025-51` pages (log scale). Background bands indicate coarse resource regimes, and colors denote writing script.

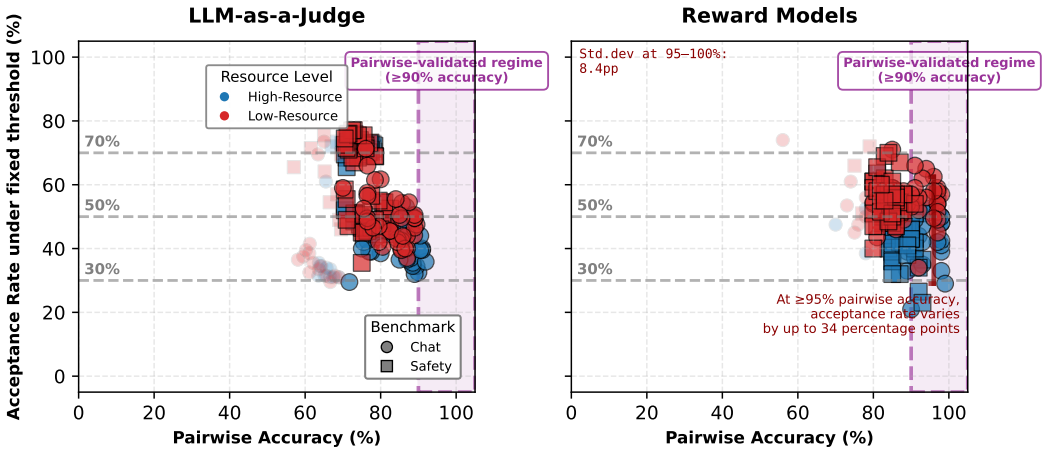

Figure 4: pairwise accuracy does not accurately reflect reward model ability

Table 5, pairwise accuracy is uniformly high across languages and models for both prompted LLM-AS-A-JUDGE evaluators and trained reward models. Under this metric alone, evaluators appear stable, consistent, and largely language-agnostic. However, pairwise accuracy captures only relative ordering and is insensitive to systematic shifts in score distributions. Many practical uses of evaluators—including threshold-based filtering, reward shaping in RLHF, and safety enforcement—depend not on relative rankings but on absolute scores. To examine whether high pairwise accuracy implies comparable downstream behavior across languages, we analyze acceptance rates under a fixed global decision threshold.

Specifically, we calibrate a single threshold on the full evaluation set and apply it uniformly across all languages. For each language, we compute the acceptance rate: the proportion of responses whose score exceeds this threshold. Figure 4 plots per-language acceptance rate against pairwise accuracy for both evaluation paradigms.

Despite pairwise accuracy remaining near saturation (often exceeding 90–95%), acceptance rates diverge dramatically across languages. Even for trained reward models—which exhibit lower overall variance—acceptance rates still vary by up to 34 percentage points at comparable pairwise accuracy. These discrepancies occur squarely within the regime typically considered "validated" by pairwise metrics. Crucially, this divergence is not visible in pairwise accuracy itself. Languages that appear equally well-evaluated under standard pairwise validation can experience substantially different decision outcomes under identical thresholds.

USE OF LARGE LANGUAGE MODELS

LLMs were used as auxiliary tools to assist with code generation and debugging, and to polish the writing and presentation of the manuscript. All scientific contributions, experimental decisions, and interpretations were made by the authors.

ACKNOWLEDGEMENTS

Ej Zhou and Lucas Resck gratefully acknowledge funding from the Cambridge Commonwealth, European and International Trust through PhD scholarships.

REFERENCES

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. Aya expanse: Combining research breakthroughs for a new multilingual frontier, 2024. URL https://arxiv.org/abs/2412.04261.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, 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, Danny Wyatt, 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, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, 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, Jack Zhang, 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, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, 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, Maria Tsimpoukelli, 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, Ning Zhang, 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, Rohan Maheswari, 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,

Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, 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, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, 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, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, 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, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, 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, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, 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, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, 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, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, 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, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, 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, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, 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, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Srishti Gureja, Lester James Validad Miranda, Shayekh Bin Islam, Rishabh Maheshwary, Drishti Sharma, Gusti Triandi Winata, Nathan Lambert, Sebastian Ruder, Sara Hooker, and Marzieh Fadaee. M-RewardBench: Evaluating reward models in multilingual settings. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 43–58, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.3. URL https://aclanthology.org/2025.acl-long.3/.

Tom Kocmi and Christian Federmann. Large language models are state-of-the-art evaluators of translation quality. 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 (eds.), *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pp. 193–203, Tampere, Finland, June 2023. European Association for Machine Translation. URL https://aclanthology.org/2023.eamt-1.19/.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. RewardBench: Evaluating reward models for language modeling. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1755–1797, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.96. URL https://aclanthology.org/2025.findings-naacl.96/.

Chris Yuhao Liu, Liang Zeng, Jiacai Liu, Rui Yan, Jujie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. Skywork-reward: Bag of tricks for reward modeling in llms, 2024. URL https://arxiv.org/abs/2410.18451.

Xingzhou Lou, Dong Yan, Wei Shen, Yuzi Yan, Jian Xie, and Junge Zhang. Uncertainty-aware reward model: Teaching reward models to know what is unknown, 2025. URL https://arxiv.org/abs/2410.00847.

José Pombal, Dongkeun Yoon, Patrick Fernandes, Ian Wu, Seungone Kim, Ricardo Rei, Graham Neubig, and André F. T. Martins. M-prometheus: A suite of open multilingual llm judges, 2025. URL https://arxiv.org/abs/2504.04953.

Qwen : An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf.

A Experimental Details & Dataset

A.1 Languages and Linguistic Properties

Code	Language	Script	Family	Resource Class
ar	Arabic	Arabic	Afro-Asiatic	High
cs	Czech	Latin	Indo-European	High
de	German	Latin	Indo-European	High
el	Greek	Greek	Indo-European	Mid
en	English	Latin	Indo-European	High
es	Spanish	Latin	Indo-European	High
fa	Persian	Arabic	Indo-European	High
fr	French	Latin	Indo-European	High
he	Hebrew	Hebrew	Afro-Asiatic	Mid
hi	Hindi	Devanagari	Indo-European	High
id	Indonesian	Latin	Austronesian	Mid
it	Italian	Latin	Indo-European	High
ja	Japanese	Japanese	Japonic	High
ko	Korean	Hangul	Koreanic	Mid
nl	Dutch	Latin	Indo-European	High
pl	Polish	Latin	Indo-European	High
pt	Portuguese	Latin	Indo-European	High
ro	Romanian	Latin	Indo-European	Mid
ru	Russian	Cyrillic	Indo-European	High
tr	Turkish	Latin	Turkic	High
uk	Ukrainian	Cyrillic	Indo-European	Mid
vi	Vietnamese	Latin	Austroasiatic	High
zh	Chinese	Han / Hant	Sino-Tibetan	High

Table 1: Languages in RewardBench, M-RewardBench and their linguistic properties. Data from Gureja et al. (2025).

A.2 Benchmark Composition and Task Breakdown

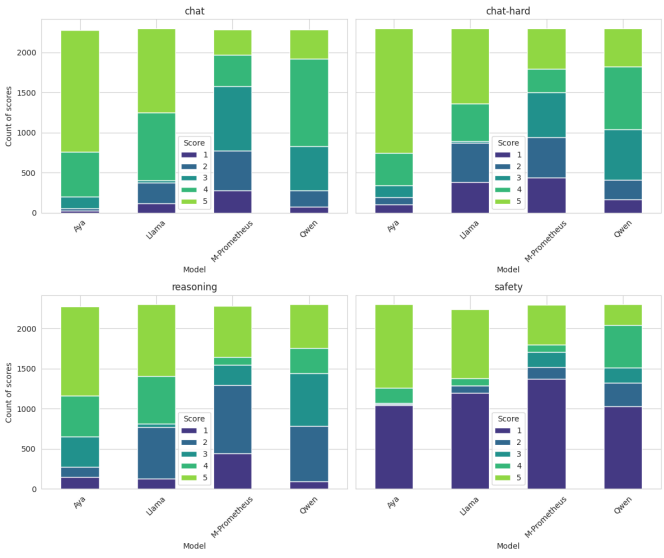

Figure 5: Distribution of reward scores for *unchosen* responses across benchmark domains. Stacked bars show the count of instances assigned to each discrete score level by different reward models. Unchosen responses exhibit wider score dispersion than chosen responses, reducing ceiling effects and revealing language-dependent variation.

B Full Pointwise Score Statistics

These tables provide the complete per-language statistics underlying Figures 1 and 4 in the main text.

Lang	Model Scores																Task Averages				
	Chat				Chat-Hard				Reasoning				Safety				C	H	R	S	O
	Aya	Q	M	LLaMA	Aya	Q	M	LLaMA	Aya	Q	M	LLaMA	Aya	Q	M	LLaMA	Avg	Avg	Avg	Avg	Avg
hi	4.64	4.01	3.75	4.15	4.74	3.98	3.56	4.04	4.50	3.55	2.83	3.93	3.41	2.72	2.57	2.94	4.14	4.08	3.70	2.91	**3.71**
el	4.62	3.75	3.45	4.20	4.49	3.62	3.44	3.72	4.14	3.15	2.86	3.65	3.07	2.48	2.64	2.95	4.00	3.82	3.45	2.79	3.51
fa	4.45	3.72	3.10	4.06	4.46	3.76	3.21	3.88	4.25	3.27	2.95	3.65	3.20	2.61	2.52	2.88	3.83	3.83	3.53	2.80	3.50
vi	4.71	3.78	3.04	4.08	4.46	3.68	3.29	3.48	4.05	3.29	2.90	3.63	3.27	2.54	2.29	2.96	3.90	3.73	3.47	2.76	3.46
ko	4.37	3.73	2.95	4.32	4.27	3.68	2.84	4.15	4.05	3.37	2.91	3.59	3.42	2.63	2.07	2.95	3.84	3.73	3.48	2.77	3.46
ja	4.47	3.77	3.02	4.36	4.20	3.75	2.86	3.79	4.19	3.42	3.02	3.64	3.12	2.60	2.30	2.77	3.90	3.65	3.57	2.70	3.46
he	4.66	3.70	3.13	4.50	4.69	3.71	3.22	4.08	3.40	3.25	2.71	3.89	3.25	2.29	1.76	2.88	4.00	3.92	3.31	2.54	3.44
tr	4.46	3.76	2.86	4.11	4.42	3.68	3.06	3.72	3.98	3.33	2.72	3.64	3.10	2.53	2.29	2.86	3.80	3.72	3.42	2.69	3.41
ar	4.60	3.78	2.95	4.11	4.58	3.73	2.88	3.72	4.03	3.32	2.78	3.73	3.00	2.49	1.67	2.84	3.86	3.73	3.46	2.50	3.39
uk	4.74	3.64	3.02	4.05	4.58	3.62	3.20	3.60	4.02	3.20	2.77	3.75	2.94	2.39	2.19	2.20	3.86	3.75	3.44	2.43	3.37
ru	4.69	3.60	2.99	3.96	4.62	3.34	2.99	3.40	4.09	3.23	2.86	3.59	2.97	2.39	2.25	2.57	3.81	3.59	3.44	2.55	3.35
zh	4.61	3.60	2.96	4.02	4.43	3.40	2.66	3.43	4.07	3.16	2.98	3.68	3.16	2.40	2.08	2.72	3.80	3.48	3.47	2.59	3.33
ro	4.51	3.64	3.05	4.06	4.31	3.41	3.04	3.36	3.63	3.17	2.93	3.65	3.00	2.34	2.45	2.60	3.82	3.53	3.35	2.60	3.32
id	4.47	3.56	2.86	4.01	4.36	3.46	2.71	3.24	4.19	3.25	2.76	3.69	3.20	2.37	1.79	2.89	3.73	3.44	3.47	2.56	3.30
nl	4.48	3.50	3.04	3.99	4.24	3.27	3.11	3.11	3.99	3.13	2.90	3.48	3.12	2.36	2.24	2.50	3.75	3.43	3.38	2.56	3.28
pt	4.54	3.57	2.73	4.04	4.27	3.32	2.77	3.25	3.94	3.16	2.68	3.56	3.01	2.38	2.19	2.69	3.72	3.40	3.33	2.57	3.26
de	4.60	3.57	2.94	3.89	4.32	3.23	2.70	3.10	4.18	3.15	2.89	3.48	2.98	2.33	2.26	2.46	3.75	3.34	3.42	2.51	3.25
en	4.42	3.40	3.18	3.87	4.16	3.14	3.01	3.01	3.83	3.08	3.04	3.53	2.94	2.26	2.43	2.72	3.72	3.33	3.37	2.59	3.25
cs	4.54	3.62	2.76	3.95	4.28	3.46	2.66	3.44	4.25	3.11	2.57	3.61	2.56	2.45	2.12	2.61	3.72	3.46	3.38	2.44	3.25
es	4.38	3.53	2.76	3.98	4.24	3.34	2.75	3.08	3.80	3.18	2.85	3.68	3.04	2.33	2.38	2.56	3.66	3.35	3.38	2.58	3.24
pl	4.57	3.66	2.89	3.89	4.38	3.47	2.95	3.19	3.61	3.12	2.90	3.59	2.88	2.36	2.06	2.35	3.75	3.50	3.31	2.41	3.24
fr	4.43	3.51	2.70	4.00	4.22	3.25	2.69	3.06	3.88	3.14	2.75	3.59	3.07	2.38	2.37	2.65	3.66	3.31	3.34	2.62	3.23
it	4.53	3.41	2.59	3.92	4.40	3.28	2.58	3.01	4.08	3.18	2.78	3.50	2.97	2.32	2.06	2.56	3.61	3.32	3.38	2.48	**3.20**

Table 2: Average evaluation scores (**1-5**) across languages for four task categories (Chat, Chat-Hard, Reasoning, Safety) and four models (Aya, Qwen, Unbabel, LLaMA). Task-level and overall averages are reported for each language.

Lang	Model Scores																Task Averages				
	Chat				Chat-Hard				Reasoning				Safety				C	H	R	S	O
	URM	BTRM	SkyG	SkyL	URM	BTRM	SkyG	SkyL	URM	BTRM	SkyG	SkyL	URM	BTRM	SkyG	SkyL	Avg	Avg	Avg	Avg	Avg
el	1.58	1.14	0.62	1.59	0.46	0.97	0.92	0.84	1.78	−0.11	0.35	1.22	−0.28	−0.70	−1.14	−0.62	1.23	0.80	0.81	−0.69	**0.54**
he	1.56	0.66	0.67	1.14	0.10	0.63	0.66	0.58	1.79	−0.11	0.36	0.75	0.26	−1.50	−1.59	−1.24	1.01	0.49	0.70	−1.02	0.29
ar	0.95	1.21	0.61	1.09	−0.35	0.95	0.50	0.45	1.39	0.18	0.31	1.00	−0.95	−1.20	−1.70	−0.84	0.96	0.39	0.72	−1.17	0.22
hi	0.82	0.99	0.73	0.92	−0.15	0.96	0.80	0.33	0.96	0.11	0.38	0.75	−0.65	−0.87	−1.46	−1.12	0.86	0.48	0.55	−1.02	0.22
uk	0.42	0.97	0.86	1.36	−0.39	0.84	0.75	0.77	1.27	−0.16	0.38	1.17	−1.23	−1.43	−1.43	−0.75	0.90	0.49	0.66	−1.21	0.21
ro	1.01	0.86	0.63	1.08	−0.39	0.79	0.71	0.45	1.37	−0.08	0.35	1.09	−1.04	−1.40	−1.66	−1.05	0.89	0.39	0.68	−1.29	0.17
fa	0.90	1.07	0.36	0.67	−0.10	1.05	0.49	0.25	1.48	0.01	0.14	0.95	−0.80	−0.95	−1.58	−1.22	0.75	0.43	0.64	−1.14	0.17
tr	0.58	0.82	1.07	0.70	−0.66	0.85	0.95	0.18	1.11	−0.06	0.57	0.76	−1.25	−1.18	−1.33	−1.05	0.79	0.33	0.59	−1.20	0.13
cs	0.42	1.04	0.90	1.06	−0.65	0.89	0.66	0.25	1.03	−0.06	0.43	0.90	−0.97	−1.46	−1.47	−0.95	0.85	0.29	0.57	−1.21	0.13
pl	0.08	1.03	0.74	1.04	−0.55	0.79	0.42	0.29	1.10	0.04	0.65	0.98	−1.07	−1.29	−1.46	−1.10	0.72	0.24	0.69	−1.23	0.11
vi	0.66	0.99	0.49	1.06	−0.71	0.76	0.58	0.34	1.27	0.03	0.32	0.88	−1.19	−1.44	−1.73	−1.09	0.80	0.24	0.62	−1.36	0.08
ko	0.84	0.75	1.15	0.30	−0.55	0.67	1.14	−0.47	1.27	−0.07	0.78	0.39	−1.01	−1.41	−0.86	−1.73	0.76	0.20	0.59	−1.25	0.07
ru	0.72	0.98	0.60	1.08	−0.60	0.67	0.45	0.19	1.05	0.04	0.59	0.85	−1.22	−1.48	−1.58	−1.22	0.85	0.18	0.63	−1.37	0.07
nl	0.70	0.90	0.55	0.77	−0.62	0.69	0.09	−0.19	1.04	−0.01	0.68	0.77	−1.03	−1.52	−1.94	−1.45	0.73	−0.01	0.62	−1.49	−0.04
it	0.38	1.09	0.47	0.97	−1.21	0.67	0.31	−0.01	1.01	0.12	0.49	0.83	−1.46	−1.44	−1.96	−1.30	0.72	−0.06	0.61	−1.54	−0.07
de	0.54	0.87	0.68	0.55	−1.04	0.65	0.38	−0.43	0.86	0.00	0.57	0.53	−1.14	−1.62	−1.80	−1.60	0.66	−0.11	0.49	−1.54	−0.13
ja	0.08	0.66	0.72	0.42	−0.86	0.60	0.63	−0.26	1.13	−0.08	0.31	0.60	−1.54	−1.90	−1.12	−1.70	0.47	0.03	0.49	−1.57	−0.14
pt	0.56	0.89	0.45	0.30	−0.94	0.55	0.28	−0.59	1.09	0.10	0.65	0.46	−1.18	−1.62	−2.02	−1.85	0.55	−0.17	0.57	−1.67	−0.18
es	0.11	0.96	0.57	0.54	−1.44	0.61	−0.02	−0.53	0.82	0.12	0.45	0.52	−1.54	−1.57	−2.23	−1.65	0.54	−0.35	0.48	−1.75	−0.27
fr	0.18	0.88	0.49	0.71	−1.52	0.51	0.25	−0.53	0.48	−0.02	0.64	0.55	−1.83	−1.69	−2.09	−1.74	0.57	−0.32	0.41	−1.84	−0.30
zh	0.67	−0.05	0.77	0.09	−1.08	−0.08	0.79	−0.90	1.08	−0.02	0.60	0.19	−1.34	−3.10	−1.49	−2.18	0.37	−0.32	0.47	−2.03	−0.38
id	−0.06	0.61	0.34	0.15	−0.80	0.42	−0.13	−0.67	0.88	−0.02	0.24	0.32	−1.25	−2.10	−2.06	−1.93	0.26	−0.30	0.35	−1.84	−0.38
en	0.71	0.58	0.75	−0.21	−1.49	0.18	0.13	−2.09	0.67	−0.12	0.83	0.05	−1.10	−2.47	−2.29	−2.72	0.46	−0.82	0.36	−2.15	**-0.54**

Table 3: Average evaluation normalized z-scores across languages for four task categories (Chat, Chat-Hard, Reasoning, Safety) and four reward models (URM-LLaMa, BTRM-Qwen, Skywork-Gemma, Skywork-LLaMa). Scores are z-normalized within each model. Task-level and overall averages are reported for each language.

C Statistical Significance of Language-Dependent Effects

This appendix reports the full statistical results for both prompted LLM judges and trained reward models. For all evaluators, we reject the null hypothesis of equal mean scores across languages ($p < 0.001$). Reward models generally exhibit larger F-statistics than prompted judges, suggesting that supervised preference training amplifies language-conditioned score differences.

D Cross-Model Correlation Matrices

Task-specific correlations show similar qualitative trends and are included for completeness.

E Per-Model Pairwise Accuracy Tables

LLM-as-a-Judge	F	p-value	n	**Reward Model**	F	p-value	n
Aya-Expanse-32B	2.31	< 0.001	9148	RM-LLaMA-3.1-8B	7.25	< 0.001	9200
Qwen-2.5-72B	4.45	< 0.001	9187	BTRM-Qwen	8.42	< 0.001	9200
M-Prometheus-14B	5.38	< 0.001	9161	Skywork-Gemma-2-27B	4.10	< 0.001	9200
LLaMA-3.1-70B-Instruct	4.83	< 0.001	9138	Skywork-LLaMA-3.1-8B	29.10	< 0.001	9200

Table 4: One-way ANOVA results testing for differences in mean evaluation scores across languages. All evaluators exhibit statistically significant language effects ($p < 0.001$).

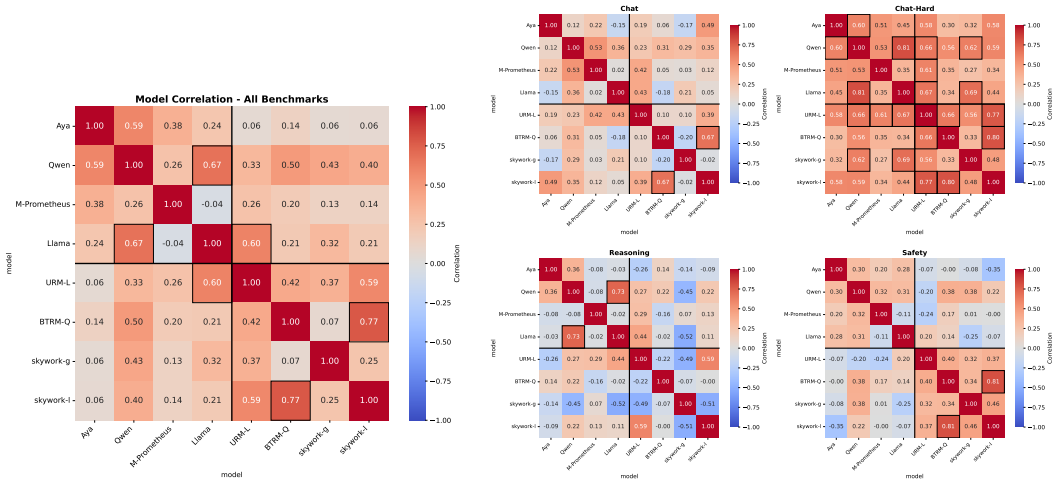

Figure 6: Correlation of language-dependent scoring patterns across judge models, aggregated over all benchmarks. Each cell reports the Pearson correlation between per-language mean scores assigned by two models.

Language	**LLM-as-a-Judge**								**Reward Models**							
	Aya		Qwen		M-Prom		Llama 3		BTRM-Q		URM-L		Skywork-g		Skywork-l	
	Chat	Safety	Chat	Safety	Chat	Safety	Chat	Safety	Chat	Safety	Chat	Safety	Chat	Safety	Chat	Safety
English	81.4	83.3	93.0	87.0	91.6	77.6	87.6	83.9	98.0	92.0	97.0	90.0	88.0	93.0	99.0	93.0
Russian	76.1	83.8	91.0	84.7	91.8	78.5	85.6	81.5	97.0	85.0	92.0	88.0	84.0	86.0	94.0	91.0
Italian	79.4	83.7	93.6	83.8	93.9	76.8	85.4	85.2	98.0	85.0	89.0	86.0	85.0	85.0	97.0	91.0
German	78.4	82.7	92.0	85.0	90.8	74.7	86.1	82.8	98.0	89.0	92.0	85.0	83.0	85.0	93.0	91.0
Portuguese	79.1	86.0	91.4	83.9	93.0	75.6	85.5	83.0	98.0	86.0	91.0	86.0	85.0	85.0	93.0	89.0
Romanian	78.2	84.4	90.1	83.6	92.2	74.7	85.1	83.2	96.0	86.0	89.0	81.0	86.0	84.0	91.0	86.0
Indonesian	80.0	80.3	92.3	86.0	92.7	77.0	87.4	80.8	98.0	84.0	87.0	84.0	78.0	88.0	89.0	92.0
Spanish	82.6	81.0	92.3	84.2	93.0	72.7	84.9	83.9	98.0	88.0	90.0	85.0	78.0	87.0	93.0	90.0
Vietnamese	75.7	81.4	87.9	79.7	88.6	74.6	85.5	81.2	97.0	85.0	87.0	82.0	87.0	88.0	92.0	83.0
Chinese	78.2	79.9	89.4	82.7	89.2	76.6	85.0	78.6	90.0	85.0	86.0	83.0	84.0	83.0	91.0	85.0
Turkish	79.9	77.8	92.2	86.7	92.2	76.9	85.9	81.2	97.0	82.0	86.0	80.0	82.0	84.0	87.0	82.0
French	80.0	83.5	92.7	85.1	92.3	74.9	84.2	85.1	98.0	88.0	85.0	85.0	85.0	91.0	91.0	90.0
Korean	78.4	77.1	90.1	82.5	92.2	78.4	79.0	72.6	97.0	85.0	90.0	85.0	81.0	83.0	92.0	80.0
Dutch	80.5	82.5	92.9	86.0	91.1	81.5	85.1	83.2	98.0	88.0	84.0	87.0	83.0	87.0	94.0	90.0
Polish	78.6	84.2	88.7	86.3	89.1	78.2	86.9	83.9	97.0	84.0	83.0	85.0	78.0	87.0	90.0	88.0
Arabic	77.8	80.9	88.3	85.8	92.5	74.2	85.1	82.9	97.0	85.0	83.0	80.0	76.0	86.0	85.0	79.0
Greek	77.3	79.8	92.4	82.3	86.7	79.5	80.4	79.5	94.0	80.0	85.0	81.0	73.0	81.0	85.0	84.0
Czech	75.9	81.8	91.4	80.9	89.9	77.5	86.9	83.0	98.0	87.0	82.0	82.0	80.0	83.0	90.0	85.0
Ukrainian	74.4	78.1	90.1	81.1	88.5	74.1	84.5	83.5	96.0	84.0	86.0	84.0	80.0	77.0	91.0	87.0
Hindi	76.3	81.5	91.4	84.1	84.8	70.1	85.1	79.8	93.0	80.0	88.0	83.0	75.0	82.0	83.0	85.0
Japanese	77.7	82.2	91.8	84.7	91.6	72.2	79.3	80.2	95.0	87.0	86.0	85.0	70.0	81.0	90.0	89.0
Persian	80.9	81.8	89.9	85.4	90.1	76.4	86.7	79.3	96.0	80.0	78.0	85.0	81.0	85.0	88.0	86.0
Hebrew	76.7	77.5	87.1	81.7	86.8	66.6	77.2	78.3	97.0	85.0	56.0	75.0	82.0	82.0	74.0	85.0
Mean	80.0		87.5		83.1		83.1		90.9		84.8		83.0		88.5	
Std Dev	2.7		3.9		8.0		3.1		6.2		5.8		4.4		4.7	

Table 5: Per-language pairwise accuracy (accepted $\geq$ rejected) for the Chat and Safety benchmarks. Despite substantial language-dependent variation in pointwise model scores, pairwise accuracy remains uniformly high and consistent across languages and models, with minimal variation ($\leq 5\%$) within each evaluator.