Isolating LLM Performance Gains in Pre-training versus Instruction-tuning for Mid-resource Languages: The Ukrainian Benchmark Study

Yurii Paniv
Ukrainian Catholic University
vul. Kozelnytska 2
Lviv, 79026, Ukraine
paniv@ucu.edu.ua

Abstract

This paper evaluates language model performance on Ukrainian language tasks across multiple downstream benchmarks, including summarization, closed and open question answering, and translation at both sentence and paragraph levels. We also introduce LongFlores, an extension of the FLORES benchmark designed specifically to assess paragraph-level translation capabilities. In our experiments, we compare the performance of base models against their instruction-tuned counterparts to isolate and quantify the source of performance improvements for Ukrainian language tasks. Our findings reveal that for popular open source models, base models are stronger in the few-shot setting for the task than their instruction-tuned counterparts in the zero-shot setting. This suggests lower attention paid to Ukrainian during the instruction-tuning phase, providing valuable insights for future model development and optimization for Ukrainian and potentially other lower-resourced languages.

1 Introduction

Large Language Models (LLMs) have demonstrated SOTA performance across various tasks, yet these capabilities have been predominantly studied within English-language contexts, both in training and evaluation. While recent years have witnessed a surge in multilingual LLMs, a critical question remains unexplored: At which development stage do performance improvements for non-English languages emerge? Is it during pretraining or instruction-tuning? By investigating this question, we can gain valuable insights into the distribution of multilingual data across different stages of model development.

To address this gap, we evaluate LLM performance across diverse Ukrainian language tasks, including summarization, extractive and general question answering, and translation between Ukrainian and eight other languages.

Additionally, we introduce a novel benchmark for long-context translation based on FLORES (Goyal et al., 2021). Current translation benchmarks primarily focus on sentence-level evaluation, with few resources dedicated to assessing long-context translation capabilities. Yet, paragraph-level translation more closely resembles real-world applications and could prove invaluable for developing LLMs for languages with fewer resources than English. With effective translation tools, English instructions can be adapted to lower-resource languages, transferring English capabilities to those languages.

As for our main contribution, we present a thorough evaluation of LLM performance for Ukrainian across multiple downstream tasks, including XLSUM-based summarization, MMLU-based general question answering, Belebele-based extractive question answering, a translated version of SQUAD, and sentence-level translation using the FLORES benchmark across nine languages.

Second, we introduce LongFlores, a long-context setting of the FLORES benchmark designed to evaluate paragraph-level translation.

Third, we investigate performance differences between pretrained and instruction-tuned versions of models (where available) to isolate the development stage at which performance gains occur for Ukrainian language tasks. Our findings, particularly regarding the source of performance improvements, will benefit other mid-resource languages by helping researchers and developers focus on the most impactful development stages for achieving state-of-the-art results in specific lower-resourced languages.

All evaluation scripts, our LongFlores benchmark, eval results, and leaderboard are published here https://github.com/robinhad/

Proceedings of Recent Advances in Natural Language Processing,pages 876–883
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-100

ukrainian-llm-leadeboard to facilitate further research in this direction.

2 Related work

2.1 Performance gains on pretraining vs finetuning

Overall, there is an extensive literature for evaluating LLMs generally, with models like LLama (Grattafiori et al., 2024) publishing general performance gains. Still, models like LLama are trained on majority of English texts, where 50% of data is general knowledge English text, 25% of mathematical and reasoning tokens, 17% code tokens, and 8% multilingual tokens (Grattafiori et al., 2024). Unfortunately, there is no information about how much multilingual data is contained during instruction-tuning phase.

To the best of our knowledge, there are few to no papers that study this specific gap, but there are a couple of works that explore the source of performance gaps in general.

Gao et al. (2024a) explore the effect of multilingual pretraining and instruction tuning on the cross-lingual knowledge alignment mechanism. Researchers measured cross-lingual alignment specifically and found that instruction-tuning improves downstream task performance much more than pretraining. However, they use translated data from English to measure performance in other languages. Also, they explore performance in a zero-shot setting without a few-shot evaluation and, subsequently, don't explore base models.

Jindal et al. (2024) explore how to expand LLM knowledge and how that knowledge then affects downstream performance on different benchmarks. As claimed by that paper, most of the LLM knowledge could come from pretraining, which then can be relatively effortlessly transferred to instruction-tuning capabilities. Nevertheless, they don't explore a few-shot setting for base models to see the origin of LLM performance.

2.2 Benchmarks for Ukrainian

There are numerous benchmarks introduced for Ukrainian language, such as UA-CBT (Hamotskyi et al., 2024), Winograd schema challenge (Kuchmiichuk, 2023), FLORES for multilingual translation across 200 languages (Goyal et al., 2021), XL-SUM for summarization (Hasan et al., 2021), Global MMMU (Singh et al., 2024), which is a human-validated MMLU questions benchmark and national exam, ZNO from UNLP 2024 Shared task (Romanyshyn et al., 2024).

Last, but not least, researchers from INSAIT presented a set of classic benchmarks as part of release of their model specifically for Ukrainian language (Yukhymenko et al., 2025), where they introduced adapted versions of Winogrande challenge, Hellaswag, ARC Easy/Challenge, TriviaQA, GSM-8k, MMLU, IFEval and ZNO, testing knowledge of the Ukrainian high school curriculum in Ukrainian language & literature, history, mathematics and geography. To the best of our knowledge, there is no comprehensive evaluation of the Ukrainian language across downstream benchmarks and measuring gains from particular training stages.

3 Benchmarks

We follow the same methodology as IberoBench (Baucells et al., 2025), by evaluating models across a set of classic downstream benchmarks.

Summarization. For the summarization task, we use XLSUM benchmarks based on BBC news articles and professionally annotated summaries. We use only Ukrainian split.

Extractive question answering We test Ukrainian version of SQUAD dataset (Ivanyuk-Skulskiy et al., 2021), which was translated and annotated by students. We also use Ukrainian Belebele split (Bandarkar et al., 2024) for this task.

Option question answering As for general knowledge testing, we use Global MMLU benchmark (Singh et al., 2024) from Cohere, which contains adapted MMLU question across various subjects from STEM to Humanities into different languages that are human-annotated and validated after translation, making them usable for evaluation.

Translation We test on the classic FLORES benchmark (Goyal et al., 2021) for sentence-level translation. Besides that, we introduce a paragraph-level version of this benchmark called LongFLORES. We test this across 9 languages in both directions from and to Ukrainian. The languages are: English, Crimean Tatar, Polish, Russian, Romanian, German, Czech, Hungarian, Slovak. We selected those languages based on several criteria: 1) languages of minorities in Ukraine, 2) languages of neighboring countries 3) languages of countries with considerable Ukrainian diaspora. FLORES benchmarks contain sentence positions and paragraph metadata, which enable researchers to reconstruct source paragraphs. This helps us to create

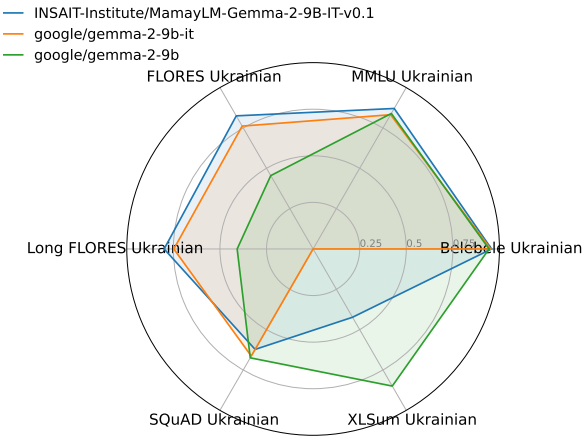

Figure 1: Comparison across gemma-2-9b-it and instruction-tuned versions of that model. We can see that MamayLM provides performance improvement both over the instruction-tuned version and the base version, indicating a benefit of extensive instruction-tuning for Ukrainian.

a paragraph-level benchmark. A similar setting is introduced for Finno-Ugric languages (Pashchenko et al., 2025), but is subsequently translated and validated due to source sentence problems. As a result, the reconstructed set contains 281 items in a dev set that we use for this benchmark. We found sentence-level problems in the FLORES benchmark, which are subsequently introduced in the paragraph-level setting. Despite that, we believe this benchmark would be helpful to get a rough estimate of translation performance for both sentence-level and paragraph-level settings, but better benchmarks are needed.

4 Experimental Setup

Due to budgetary constraints and practicality concerns, we tested popular open source models with parameters ranging from 4 to 32 billion. We use lm-evaluation-harness (Gao et al., 2024b) and evaluate models through the VLLM (Kwon et al., 2023) framework. For most models, we use a single node with 2x RTX A6000 Ada and approximately 2 weeks of GPU hours. As for model selection, we selected models based on popularity and claims about training on Ukrainian data. We run each model on standardized prompt sets and adapt generation parameters from existing reference implementations in lm-evaluation-harness from other languages. As for metrics, for most benchmarks we use BLEU (Papineni et al., 2002) (including paragraph-level translation as suggested by Deutsch et al. (2023)). with the exception of Global MMLU for which we use accuracy score. We evaluate all models in 0-shot and 3-shot settings (with the exception of base models evaluated only in 3-shot). For model comparison, we record each model's ranking on each task and then calculate its average rank.

5 Results & Discussion

Our evaluation reveals a performance gap between base and instruction-tuned models when processing Ukrainian language tasks. The results demonstrate that pre-training appears to be the primary source of Ukrainian language capabilities across most model architectures, with instruction-tuning lacking performance on Ukrainian-specific tasks. We show an overall breakdown in **??** and a detailed evaluation by language pair in **??**.

Base models consistently outperform their instruction-tuned counterparts across multiple model families when evaluated with few-shot prompting as shown on Figure 2 and in Table 1. For example, Qwen3-14B-Base (Yang et al., 2025) achieving an average rank of 11.00 compared to 18.00 for the instruction-tuned variant, and Qwen3-8B-Base reaching 16.00 versus 20.67 for its instruction-tuned version. Similarly, Llama 3.1-8B (Grattafiori et al., 2024) base model (rank 24.17) substantially outperforms the instruct version (34.17), while Mistral-7B-v0.3 (Jiang et al., 2023) base (26.83) exceeds the performance of its instruction-tuned counterpart (30.83). EuroLLM (Martins et al., 2024) follows the same trend, with

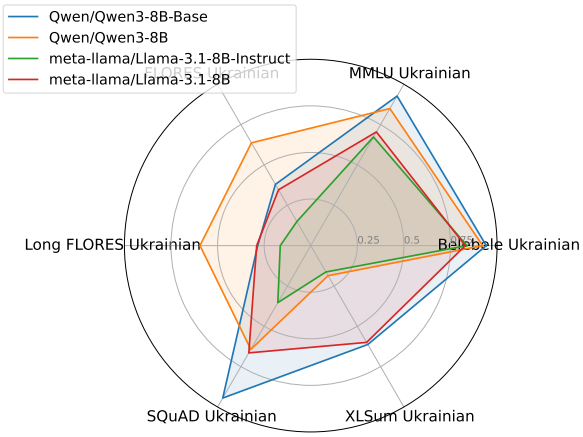

Figure 2: Comparison of Qwen3 models and Llama-3.1 models. We can see that most of the performance is contained in a pretraining stage of those models.

Model	Belebele Ukrainian	MMLU Ukrainian	FLORES Ukrainian	Long FLORES Ukrainian	SQuAD Ukrainian	XLSum Ukrainian	Average Rank
(0) INSAIT-Institute/MamayLM-Gemma-2-9B-IT-v0.1	88.00	63.18	**18.98**	**19.26**	34.13	2.59	**14.17**
(0) google/gemma-2-9b-it	87.56	60.38	17.52	18.01	36.39	0.00	18.83
(3) google/gemma-2-9b	86.78	60.94	10.46	9.82	36.99	5.20	17.00
(0) Qwen/Qwen3-14B	87.22	65.35	17.03	16.03	21.66	1.61	18.00
(3) Qwen/Qwen3-14B-Base	**90.56**	**70.64**	11.02	8.29	50.91	**5.36**	11.00
(0) Qwen/Qwen3-8B	84.78	61.63	14.65	14.32	35.26	1.14	20.67
(3) Qwen/Qwen3-8B-Base	86.56	67.26	8.73	6.86	**51.80**	3.75	16.00
(0) meta-llama/Llama-3.1-8B-Instruct	77.00	48.86	3.43	3.94	19.38	1.00	34.17
(3) meta-llama/Llama-3.1-8B	76.22	51.13	7.96	6.99	36.45	3.67	24.17
(0) utter-project/EuroLLM-9B-Instruct	69.44	50.86	16.00	13.85	27.57	1.55	25.33
(3) utter-project/EuroLLM-9B	72.44	52.56	16.38	15.10	42.91	4.69	17.17
(0) mistralai/Mistral-7B-Instruct-v0.3	60.00	44.54	9.90	9.62	19.83	1.86	30.83
(3) mistralai/Mistral-7B-v0.3	71.89	48.43	7.76	6.37	35.22	4.27	26.83

Table 1: Side-by-side comparison between base models in 3-shot setting and instruction-tuned versions in 0-shot. Across most families, base models demonstrate much better performance for Ukrainian than their instruction-tuned counterparts, with an exception of gemma-2-9b-it and MamayLM model, tuned specifically for Ukrainian language understanding

the base model achieving a rank of 17.17 compared to 25.33 for the instruction-tuned version.

The sole exception to this pattern emerges with the Gemma-2-9B (Team et al., 2024) family as shown in Figure 1, where the instruction-tuned model achieves a better average rank (18.83) than the base model with 3-shot prompting (17.00). However, the Ukrainian fine-tuned MamayLM-Gemma-2-9B-IT model, achieves the best overall performance with an average rank of 14.17, suggesting that domain-specific instruction-tuning can be beneficial when properly executed.

Task-specific analysis reveals that question answering benchmarks like Belebele, Global MMLU, and SQuAD Ukrainian show variable performance patterns between base and instruction-tuned models, indicating that both pre-training knowledge and instruction-following capabilities contribute to success on these tasks. However, base models consistently achieving superior performance on FLORES translation tasks when provided with few-shot examples, while instruction-tuned models frequently struggle with the XLSum summarization task, often scoring near zero.

The superior performance of base models with few-shot prompting suggests that instruction-tuning datasets may contain insufficient Ukrainian examples or that the multilingual instruction-following training process interferes with the models' pre-existing Ukrainian language representations, being optimized primarily for English.

6 Conclusion

In this study, we evaluated popular large language models on Ukrainian language tasks to investigate the relative contributions of pre-training and instruction-tuning to multilingual capabilities. Our findings demonstrate that instruction-tuned models consistently underperform their base counterparts on Ukrainian tasks in zero-shot settings. We attribute this degradation to a lack of Ukrainian instruction data during the instruction-tuning phase, where models appear to lose pre-trained Ukrainian capabilities without gaining equivalent instruction-following proficiency in the language. The superior performance of base models with few-shot prompting suggests that Ukrainian language understanding in popular models primarily occurs during pre-training. These results have important implications for other mid-resource languages. Our study suggests that practitioners should prioritize two key factors: first, ensuring robust pre-training with substantial target language representation, and second, incorporating extensive instruction data in the target language during instruction-tuning. Without adequate instruction data, the instruction-tuning process may diminish rather than enhance multilingual capabilities.

References

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Irene Baucells, Javier Aula-Blasco, Iria de Dios-Flores, Silvia Paniagua Suárez, Naiara Perez, Anna Salles, Susana Sotelo Docio, Júlia Falcão, Jose Javier Saiz, Robiert Sepulveda Torres, Jeremy Barnes, Pablo Gamallo, Aitor Gonzalez-Agirre, German Rigau, and Marta Villegas. 2025. IberoBench: A benchmark for LLM evaluation in Iberian languages. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10491–10519, Abu Dhabi, UAE. Association for Computational Linguistics.

Daniel Deutsch, Juraj Juraska, Mara Finkelstein, and Markus Freitag. 2023. Training and meta-evaluating machine translation evaluation metrics at the paragraph level. In *Proceedings of the Eighth Conference on Machine Translation*, pages 996–1013, Singapore. Association for Computational Linguistics.

Changjiang Gao, Hongda Hu, Peng Hu, Jiajun Chen, Jixing Li, and Shujian Huang. 2024a. Multilingual pretraining and instruction tuning improve cross-lingual knowledge alignment, but only shallowly.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024b. The language model evaluation harness.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation.

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. 2024. The llama 3 herd of models.

Serhii Hamotskyi, Anna-Izabella Levbarg, and Christian Hänig. 2024. Eval-UA-tion 1.0: Benchmark for evaluating Ukrainian (large) language models. In *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024*, pages 109–119, Torino, Italia. ELRA and ICCL.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Bogdan Ivanyuk-Skulskiy, Anton Zaliznyi, Oleksand Reshetar, Oleksiy Protsyk, Bohdan Romanchuk, and Vladyslav Shpihanovych. 2021. ua$_d$$atasets$: $acollectionofukrainianlanguagedatasets$.

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

Ishan Jindal, Chandana Badrinath, Pranjal Bharti, Lakkidi Vinay, and Sachin Dev Sharma. 2024. Balancing continuous pre-training and instruction fine-tuning: Optimizing instruction-following in llms.

Pavlo Kuchmiichuk. 2023. Silver data for coreference resolution in Ukrainian: Translation, alignment, and projection. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 62–72, Dubrovnik, Croatia. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2024. Eurollm: Multilingual language models for europe.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Dmytro Pashchenko, Lisa Yankovskaya, and Mark Fishel. 2025. Paragraph-level machine translation for low-resource Finno-Ugric languages. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 458–469, Tallinn, Estonia. University of Tartu Library.

Mariana Romanyshyn, Oleksiy Syvokon, and Roman Kyslyi. 2024. The UNLP 2024 shared task on fine-tuning large language models for Ukrainian. In *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024*, pages 67–74, Torino, Italia. ELRA and ICCL.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I. Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Wei-Yin Ko, Madeline Smith, Antoine Bosselut, Alice Oh, Andre F. T. Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermis, and Sara Hooker. 2024. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek An-

dreev. 2024. Gemma 2: Improving open language models at a practical size.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Hanna Yukhymenko, Anton Alexandrov, and Martin Vechev. 2025. Mamaylm: An efficient state-of-the-art ukrainian llm. https://huggingface.co/blog/INSAIT-Institute/mamaylm.