Community OSCAR: A Community Effort for Multilingual Web Data

Manuel Brack[1,2,3] **Malte Ostendorff**[1,4] **Pedro Ortiz Suarez**[1,5] **Jose Javier Saiz**[6]
Iñaki Lacunza Castilla[6] **Jorge Palomar-Giner**[6] **Alexander Shvets**[6]
Patrick Schramowski[1,2,3,7] **Georg Rehm**[2] **Marta Villegas**[6] **Kristian Kersting**[2,3,7,8]
[1]Occiglot, [2]German Research Center for Artificial Intelligence (DFKI),
[3]Computer Science Department, TU Darmstadt, [4]Deutsche Telekom,
[5]Common Crawl Foundation, [6]Language Technologies Unit, BSC,
[7]Hessian.AI, [8]Centre for Cognitive Science, TU Darmstadt
hello@occiglot.org

Abstract

The development of large language models (LLMs) relies heavily on extensive, high-quality datasets. Publicly available datasets focus predominantly on English, leaving other language communities behind. To address this issue, we introduce Community OSCAR, a multilingual dataset initiative designed to address the gap between English and non-English data availability. Through a collective effort, Community OSCAR covers over 150 languages with 46 billion documents, totaling over 345 TiB of data. Initial results indicate that Community OSCAR provides valuable raw data for training LLMs and enhancing the performance of multilingual models. This work aims to contribute to the ongoing advancements in multilingual NLP and to support a more inclusive AI ecosystem by making high-quality, multilingual data more accessible to those working with low-resource languages.

1 Introduction

The success of large language models (LLMs) hinges on access to vast amounts of high-quality data. The exact composition, procurement, and curation of this data has been one of the more closely guarded secrets of commercial LLMs. Recently, academic and open-source efforts have made significant strides in curating and refining large-scale corpora for English [10, 9, 11, 8]. These data-driven efforts are central to advancing open-source and transparent LLM initiatives.

Nonetheless, a strong disparity remains between the availability of English-language datasets and those for other languages. We argue that access to high-quality data is imperative for ensuring linguistic diversity, academic and economic competitiveness, and AI sovereignty for non-English countries and speakers. However, clean, multilingual datasets like CulturaX [7], for example, can only provide 100B+ tokens for less than ten languages. To bridge this gap, we introduce Community OSCAR, a publicly available multilingual dataset that covers over 150 languages and includes over four times as much data as previous corpora. The creation of Community OSCAR is a collective, community-driven effort, highlighting the importance of collaboration in addressing the challenges of data scarcity for non-English languages[1]. By expanding the availability of non-English data, Community OSCAR seeks to democratize access to resources essential for building inclusive, multilingual AI systems. Our initial results indicate that Community OSCAR provides valuable raw data for downstream LLM training.

Languages	151
Documents	46B
Data size	346 TiB
Crawls	45 (Oct. 2014 - Aug 2024)

Table 1: Community OSCAR dataset statistics. All statistics were calculated on a random subset of 10 releases and extrapolated to the entire dataset.

2 Community OSCAR

As the name suggests, Community OSCAR builds on prior work of the OSCAR corpus [1]. We went ahead and extended these efforts.

OSCAR. The OSCAR project (**O**pen **S**uper-large **C**rawled Aggregated co**R**pus) aims to provide open-source, web-based multilingual resources. Community OSCAR utilizes the high-performance *Ungoliant* data pipeline to process, filter, and annotate data at scale [2]. Most importantly, Ungoliant identifies and splits all documents based on their language [3, 6]. Similar to prior releases of OSCAR, we source our web-crawled data from Common Crawl's (CC) WET files.

[1]Dataset available at https://huggingface.co/datasets/oscar-corpus/community-oscar

Model	German				English			
	T-QA↑	ARC↑	HellaSwag↑	MMLU↑	T-QA↑	ARC↑	HellaSwag↑	MMLU↑
LLama-3-8B	0.476	0.476	0.599	0.537	0.439	0.594	0.821	0.667
LLama-3-8B + DE pre-train	**0.491**○	**0.507**○	**0.654**●	**0.540**●	0.449	0.573	0.804	0.627
LLama-3.1-8B	**0.504**●	0.470	0.608	0.535	0.451	0.577	0.817	0.661
LLama-3.1-8B + DE pre-train	0.483	**0.517**●	**0.650**○	**0.540**●	0.464	0.581	0.802	0.635

Table 2: Multilingual pre-training with Community OSCAR. We report benchmark scores in German and English of Llama-3 models before and after continual pre-training with 80B German tokens from a filtered version of our data.

Dataset Collection & Statistics. Community OSCAR follows the annotation schema established in the OSCAR 23.01 release[2], ensuring consistency and reliability in data quality. Consequently, Community OSCAR contains the raw CC text but includes quality annotations for filtering. In contrast to prior work, we incorporate 45 monthly CC dumps from August 2024 to October 2014. We prioritized more recent data, covering all CC releases from the last four years in addition to hand-selected earlier data. Computation was split over multiple super-computers and high-performance clusters across Europe. Community OSCAR covers 151 different languages and contains over 45B documents for a total of over 345TiB of data.

By offering this extensive corpus, we hope to contribute to the ongoing efforts to improve multilingual NLP. Further, Community OSCAR aims to ensure these advancements are accessible to a broader audience, including researchers and developers working with low-resource languages.

3 Outlook

The release of Community OSCAR now enables further progress in multilingual language modeling. We are actively working on extending the dataset to at least all available CC dumps, curating a high-quality subset from the raw data, and training LLMs on that data. All three steps yield good initial results, which we will discuss in the following section. Specifically, we conducted initial experiments with subsets of the data and plan to extend our insights to the rest of the dataset.

Extending Community OSCAR. Despite its size, this initial release of Community OSCAR still leaves room for more data to be included. We aim to provide continuous support for the dataset, processing and adding any upcoming CC dumps whenever they become available. Further, out of 100 current CC releases, we only cover 45%. We are continuing the Community OSCAR effort to incorporate every existing CC dump since 2014. We globally deduplicated a subset of Community OSCAR for over ten languages and found that consecutive crawls contain significant numbers of unique documents. Especially for very low-resource languages, that additional data can be crucial in enabling LLM training at scale.

Data Curation. The raw Community OSCAR data should be processed further before being used for LLM training. To begin with, different crawls contain large amounts of duplicate documents. Additionally, the raw data from CC consists of different quality levels concerning syntactical and grammatical correctness, factual accuracy, quality of HTML parsing, unsafe content, etc. We want to identify the high-quality subset of all documents for training and remove duplicates. Community OSCAR has already been annotated with important information to enable curation efforts. Additionally, we have begun implementing a more sophisticated curation pipeline building on fineweb [8]. We identified several steps in the fineweb filtering that must be adjusted for the specific target language. We have already made an initial cleaned and deduplicated subset of the data available online for 10 languages[3].

LLM Training. Lastly, we filtered and deduplicated the German data from 20 Community OSCAR dumps to assess its potential for LLM training. We follow existing approaches for the multilingual extension of pre-trained LLMs [5] and performed continual pre-training on LLama-3.x-8B checkpoints [4]. Specifically, we further trained the Llama-3 and LLama-3.1 checkpoints on roughly 80B German tokens interleaved with 5% English replay from fineweb-edu.

The evaluation results are depicted in Tab. 2. We can clearly see that continual pre-training on our German data significantly improves the model's

[2]Annotation scheme documented at: https://oscar-project.github.io/documentation/versions/oscar-2301/

[3]fineweb dataset at: https://huggingface.co/datasets/occiglot/occiglot-fineweb-v0.5

German performance. Crucially, that observation also holds for Llama-3.1 which is already a multi-lingual model with German capabilities.

Community OSCAR's ongoing work contributes to multilingual NLP and aims to make advancements accessible to a broader audience.

Acknowledgements

This release is supported by and was enabled by contributions from the OSCAR team at Inria (project-team ALMAnaCH), specially by Julien Abadji, Rua Ismail, and Benoit Sagot, the Common Crawl Foundation, the SLT and SAINT teams at DFKI, TU Darmstadt, the LangTech unit at the Barcelona Supercomputing Center, the 42 supercomputer and Hessian AI, the OpenGPT-X project, Fraunhofer, Jülich Supercomputing Centre, TU Dresden, Deutsche Telekom, as well as by members of the OSCAR community, in particular Sotaro Takeshita, Sebastian Nagel.

References

[1] Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a cleaner document-oriented multilingual crawled corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4344–4355, Marseille, France. European Language Resources Association.

[2] Julien Abadji, Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2021. Ungoliant: An optimized pipeline for the generation of a very large-scale multilingual web corpus. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-9) 2021. Limerick, 12 July 2021 (Online-Event), pages 1 – 9, Mannheim. Leibniz-Institut für Deutsche Sprache.

[3] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

[4] 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, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, 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, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. 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, and et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

[5] Adam Ibrahim, Benjamin Thérien, Kshitij Gupta, Mats L. Richter, Quentin Gregory Anthony, Eugene Belilovsky, Timothée Lesort, and Irina Rish. 2024. Simple and scalable strategies to continually pre-train large language models. *Trans. Mach. Learn. Res.*

[6] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics.

[7] Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2023. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. *Preprint*, arXiv:2309.09400.

[8] Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. *arXiv preprint arXiv:2406.17557*.

[9] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon LLM: outperforming curated corpora with web data only. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

[10] Daria Soboleva, Faisal Al-Khateeb, Joel Hestness, Nolan Dey, Robert Myers, and Jacob Robert Steeves. 2023. Slimpajama: A 627b token, cleaned and deduplicated version of redpajama.

[11] Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur,

Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Evan Walsh, Luke Zettlemoyer, Noah Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. 2024. Dolma: an open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.