Improving Romanian LLM Pretraining Data using Diversity and Quality Filtering

Vlad Negoiță and Mihai Masala and Traian Rebedea
National University of Science and Technology POLITEHNICA Bucharest,
313 Splaiul Independentei, 060042, Bucharest, Romania

Abstract

Large Language Models (LLMs) have recently exploded in popularity, often matching or outperforming human abilities on many tasks. One of the key factors in training LLMs is the availability and curation of high-quality data. Data quality is especially crucial for underrepresented languages, where high-quality corpora are scarce. In this work, we study the characteristics and coverage of Romanian pretraining corpora and we examine how they differ from English data. By training a lightweight multitask model on carefully LLM-annotated Romanian texts, we are able to analyze and perform multi-level filtering (e.g., educational value, topic, format) to generate high-quality pretraining datasets. Our experiments show noteworthy trends in the topics present in Romanian and English data, while also proving the effectiveness of filtering data through improved LLM pretraining performance across multiple benchmarks.

1 Introduction

Recent advances in Artificial Intelligence, especially in Natural Language Processing (NLP), have been driven by Transformer architectures (Vaswani et al., 2017) and Large Language Models (LLMs). These innovations have transformed how machines process language, enabling applications such as conversational AI, intelligent search, machine translation, and content generation.

The success of modern LLMs depends heavily on large, high-quality pretraining datasets (Longpre et al., 2024). With billions of parameters, these models require a large amount of data to capture statistical patterns, semantic nuances, and world knowledge present in human language. Although multilingual datasets (Penedo et al., 2024; Nguyen et al., 2023; Ortiz Suárez et al., 2019) have driven broad NLP progress, language-specific resources are vital for robust performance across diverse languages. Romanian is under-represented in large multilingual corpora, causing models trained on high-resource languages to underperform on Romanian. Dedicated Romanian datasets are therefore essential for training, finetuning, and evaluating language models (Masala et al., 2024).

Recent studies highlight data quality as crucial for strong benchmark performance (Penedo et al., 2024; Ali et al., 2025; Bai et al., 2025). The FineWeb-Edu framework (Lozhkov et al., 2024) proposed an educational-value-based filtering method effective across various tasks, forming the basis for our Romanian adaptation. To understand potential topic bias from this filtering, we also incorporate additional signals to examine diversity issues. Our entire recipe, including code, data, and models, is publicly available[1]. Our contributions can be summarized as follows:

- We build multidimensional (i.e., educational value, topic, format, reader education level) resources: a small human-annotated dataset (100 samples) and a large (1M samples) LLM-annotated dataset.
- We train a lightweight multi-head classifier that enables cost-effective filtering and cross-lingual distribution analysis at scale.
- We build the first high-quality pretraining dataset for Romanian - **FineWeb2-Edu-Ro** - and prove its usefulness by performing continual pretraining. Compared to other approaches, models trained on our filtered dataset exhibit superior performance across a variety of benchmarks.

2 Related Work

Recent efforts in pretraining focus on high-quality data, combining filtering (rule-based or ML-driven) with growing interest in synthetic data for its significant benefits. FineWeb (Penedo et al., 2024),

[1] https://github.com/VladNegoita/FineWeb2-Ro

Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026), pages 140–148

FineWeb2 (Penedo et al., 2024), and FineWeb-Edu (Lozhkov et al., 2024) represent a collection of high-quality web-based datasets for training large language models. The FineWeb initiative offers both English-only datasets (FineWeb - 15T tokens) and multilingual datasets (FineWeb2 - 1000+ languages - 35B Romanian words), cleaned and deduplicated.

Quality Filtering. FineWeb-Edu employs classifier-based quality (educational content) filtering using Llama-3-70B (AI@Meta, 2024) annotations and Snowflake-Arctic-Embed (Merrick et al., 2024) to train a lightweight regressor for educational value. JQL (Ali et al., 2025) introduces a language-agnostic setup for annotating educational value in text, using manual labeling and automatic translation to build a multilingual dataset. They benchmark several LLMs and select the top three based on Spearman correlation with human labels, then train lightweight models using Snowflake-Arctic-Embed. Instead of relying on a single metric for data quality, MetaRater (Zhuang et al., 2025) uses classifiers across four key criteria: professionalism, readability, reasoning, and cleanliness. Reliable identification of lower-quality documents allows alternative approaches such as document rewriting to increase the overall quality of datasets (Nguyen et al., 2025).

Romanian Datasets. Besides FineWeb2, we identify only three other important datasets for Romanian, namely CulturaX (Nguyen et al., 2023) (40B tokens in Romanian), FuLG (Bădoiu et al., 2024) (150B tokens in Romanian) and HPLT (Oepen et al., 2025) (100B tokens in Romanian). All three datasets stem from CommonCrawl, with different numbers of snapshots used and different rules for processing and filtering, with HPLT also containing data derived from Internet Archive. Crucially, all datasets employ rather standard rules based on n-gram frequency, stop word ratio or text length, and thus lack a more high-level quality-based filtering. Complementary to these raw pretraining corpora, the **OpenLLM-Ro** initiative (Masala et al., 2024) focuses on downstream performance, providing a suite of instruction-tuning datasets and foundational models (e.g., RoLlama, RoMistral) adapted for Romanian.

Evaluation. Several pretraining efforts evaluate performance across different model sizes (1B–7B) using metrics like perplexity and accuracy on benchmarks such as MMLU (Hendrycks et al., 2021), ARC (Clark et al., 2018), HellaSwag (Zellers et al., 2019), or OpenBookQA (Mihaylov et al., 2018). Crucially for our work, Masala et al. (2024) have introduced translated versions of these benchmarks, enabling native evaluation for Romanian LLMs. FineWeb-Edu trains a 1.71B model on up to 350B tokens reporting improvements over unfiltered datasets on MMLU, ARC, and OpenBookQA. Similarly, JQL pretrains a 2B Llama-based model, outperforming FineWeb2 on various benchmarks. FuLG trains a 1B decoder-only OLMo (Groeneveld et al., 2024) model and reports perplexity gains over mC4 and OSCAR (Ortiz Suárez et al., 2019).

Our approach builds upon recent advancements in data curation for LLMs, but distinguishes itself by jointly predicting multiple signals for a given text in Romanian, as educational value is insufficient for efficient pretraining (Bai et al., 2025).

3 Taxonomy Definition

For educational quality, we utilize the validated 5-point grading scale from FineWeb-Edu (described in Table 1) for its effectiveness and to enable comparison between Romanian and English distributions. Regarding additional signals, we follow the taxonomy of WebOrganizer (Wettig et al., 2025), which developed topic[2] (e.g., Health, Politics) and format[3] (e.g., News Article, Creative Writing) classifiers to analyze large-scale English web distributions. Finally, we also extract the required educational level per text (described in Table 2), which enables progressive learning or curriculum learning for LLMs (Mukherjee et al., 2023) and could enhance quality prediction. Full taxonomies are provided in Appendix A.

4 Approach

We start by manually annotating a subset of 50 organic Romanian texts from FineWeb2. Furthermore, we add another 50 texts from FineWeb-Edu (translated into Romanian and retaining their classifier's educational scores) to enable score alignment and provide a less skewed distribution.

We use the resulting dataset (100 samples) to select the best performing LLM (e.g., model, prompt, prompt language) that we will use for annotating

[2] https://huggingface.co/WebOrganizer/TopicClassifier-NoURL

[3] https://huggingface.co/WebOrganizer/FormatClassifier-NoURL

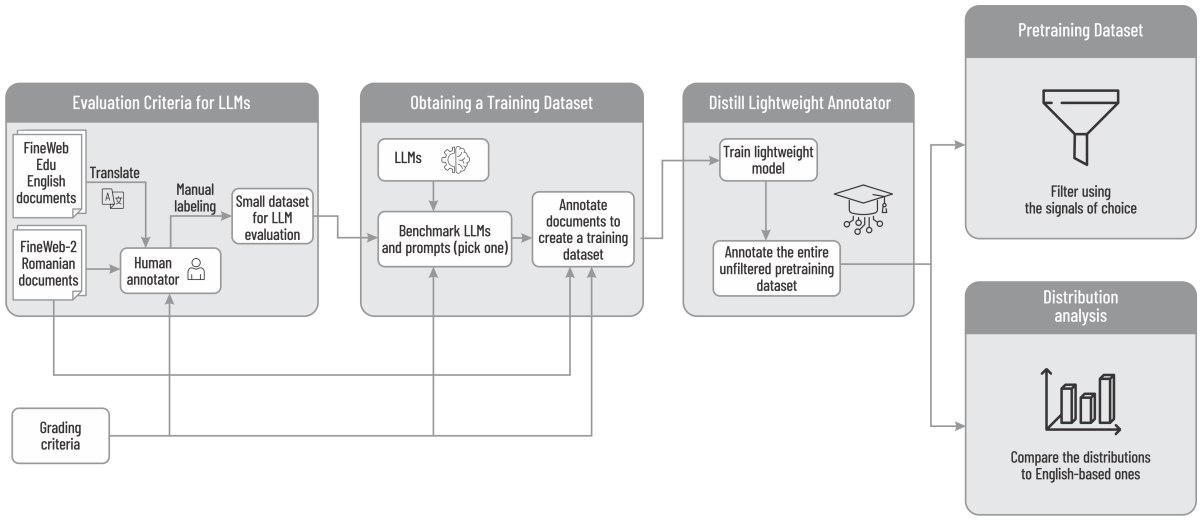

Figure 1: Flowchart detailing the multi-stage pipeline for building an educational Romanian pretraining dataset. The process includes initial human annotation and LLM benchmarking, distillation of a lightweight annotator for full-scale labeling, and final filtering with a subsequent cross-lingual distribution analysis.

Score	Criteria
1 Point	Provides basic information relevant to education but includes irrelevant or promotional material.
2 Points	Addresses educational elements but is superficial, disorganized, incoherent, or lacks alignment with standards.
3 Points	Appropriate and coherent (introductory textbook style) but not comprehensive or slightly too complex.
4 Points	Highly relevant and beneficial (textbook chapter style) with exercises and minimal irrelevant info.
5 Points	Outstanding value; detailed reasoning, profound insights, and perfectly suited for primary/grade school teaching.

Table 1: Summary of the additive 5-point educational value system.

Level	Primary Focus
Preschool	Social, emotional, cognitive, and physical development through play.
Elementary School	Literacy, basic math, science intro, and fundamental skills.
Middle School	Deepening subjects and introducing new disciplines (e.g., sciences).
High School	Various tracks (science, tech) preparing for higher education or workforce.
Bachelor's Degree	Foundational knowledge and practical skills in a specific discipline.
Graduate Degree	Advanced specialization (Master's, PhD) and rigorous training.

Table 2: Summary of the educational taxonomy levels.

a larger dataset, which will be used for training a lightweight classifier. Figure 1 illustrates the pretraining dataset construction process.

4.1 Training Dataset

Following manual annotation, we evaluated a wide range of models, including Llama-3, 3.1, and 3.3, Gemma-2 (Team et al., 2024) and Gemma-3 (Team et al., 2025), Cohere-Aya (Aryabumi et al., 2024), Qwen-2.5 (Qwen et al., 2025), and Mistral-Small (Mistral AI, 2025). We also tested multiple prompting strategies: chain-of-thought and few-shot prompting in both Romanian and English. When prompted in Romanian with a chain-of-thought approach, Gemma-3 performed exceptionally well for a model of its size (full list of models and results are included in Appendix B).

Thus, we used Gemma-3-12B[4] to annotate over 1M samples from the Romanian split of FineWeb2. These examples were partitioned, allocating 10,400 for validation and 20,800 for the test set, with the substantial remainder reserved for the training corpus.

4.2 Model Distillation

Using the validation split (10,400 examples) and focusing solely on educational value prediction (as a regression task), we selected the encoder architecture, various hyperparameters, and the training set size. Table 3 presents the performance of the evaluated encoders, which included Romanian-specific architectures—such as RoBERT-small, RoBERT-base, RoBERT-large (Masala et al., 2020), and bert-base-romanian-uncased-v1 (Dumitrescu et al., 2020)—alongside the multilingual BERT base model (Devlin et al., 2018). For further experiments we select RoBERT-base as our encoder.

Figure 2 presents how the three considered metrics (Pearson correlation, Spearman correlation, and R^2) improve on the validation set as the training dataset size increases, also confirming that the training data size was sufficient for effective learning.

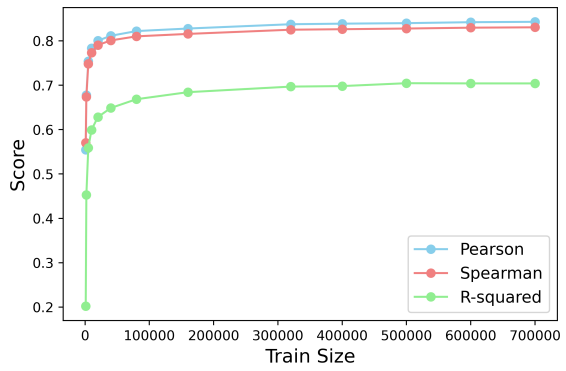

Figure 2: Educational value scores against training size.

We train a multitask model by adding four heads on top of the encoder: three for classification (topic, format, and educational level) and one for regression (educational value). The model is optimized with a composite loss function, which is a weighted sum of the individual losses from each head. The educational value loss has a weight of 1, while the three classification losses are weighted by a hyperparameter, α. Our experiments showed that this weighting factor did not significantly impact the model's final performance. The final lightweight model configuration is presented in Appendix C.

[4] https://huggingface.co/google/gemma-3-12b-it

Model	RMSE	MAE	Spearman
RoBERT-small	0.754	0.594	0.750
RoBERT-base	**0.729**	**0.574**	**0.767**
RoBERT-large	0.830	0.631	0.689
bert-multilingual	0.863	0.685	0.671
bert-romanian-v1	0.765	0.601	0.739

Table 3: Validation metrics for encoders including RMSE, MAE, and correlation coefficients in finetuning setup.

5 Results and Discussion

We compare the performance of a Llama-2-7B base model (Touvron et al., 2023) continually pretrained on filtered data versus unfiltered data. We specifically selected Llama-2 due to its predominantly English training data and consequently limited exposure to the Romanian language. This characteristic makes the model highly sensitive to the quality of the continual pretraining data, allowing us to effectively measure the impact of our corpus. Given that our dataset size (≈6B tokens) is rather small compared to standard trillion-token pretraining budgets, using a model with robust prior multilingual knowledge might obscure the improvements gained from higher quality data.

As we propose a method for obtaining higher-quality pretraining datasets for Romanian, we compare our approach with JQL, and we pick FineWeb2 as the underlying dataset to facilitate comparisons with FineWeb-Edu, in terms of topic distribution. To create the FineWeb2-Edu-Ro dataset, we augment FineWeb2 with our additional signals, filter for texts with an educational value of 3.5 or higher, and truncate them to a maximum of 4096 tokens. Table 4 lists the token and sample counts for the FineWeb2-Edu-Ro and the JQL dataset at the chosen filtering thresholds (selected as to ensure a fair comparison). Appendix D presents the impact of various thresholds for our dataset. To ensure fairness, we also matched the number of non-padding tokens, accounting for the longer average length of filtered texts.

Evaluation on Romanian translated (Masala et al., 2024) versions of MMLU (Hendrycks et al., 2021), ARC (Clark et al., 2018), and HellaSwag (Zellers et al., 2019) shows that filtering yields clear performance gains (see Figure 3).

Data	Threshold	#Tokens	#Samples
Ours	3.5	6.43B	3.9M
JQL	P92	6.23B	2.7M

Table 4: Comparison of two filtering methods for the FineWeb2 dataset, both truncated at 4096 tokens. The threshold of 3.5 for our method was selected to yield a token count comparable to the JQL P92 baseline (≈6B tokens), facilitating a fair comparison of filtering quality under a fixed compute budget.

Both filtering methods improve over the unfiltered FineWeb2 baseline, with our proposed method exhibiting stronger performance on all considered benchmarks. For all metrics, we report the average of multiple few-shot setups: 0, 1, 3, 5 shots for Ro-MMLU, 0, 1, 3, 5, 10, 25 shots for Ro-ARC, and 0, 1, 3, 5, 10 shots for Ro-HellaSwag matching previous work for Romanian (Masala et al., 2024) for a fair comparison.

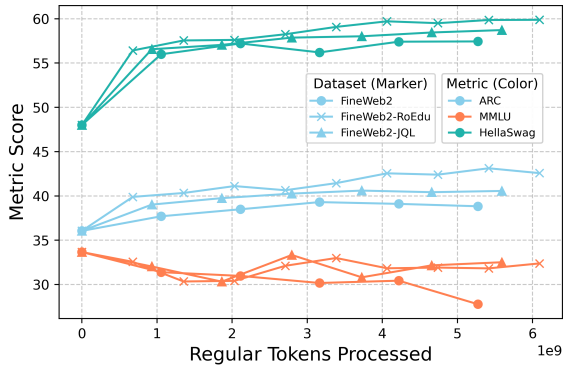

Figure 3: Pretraining results using filtered (RoEdu&JQL) and randomly-selected unfiltered FineWeb2 data. Note the consistent improvement in performance when using filtered data, with the best overall results obtained using our proposed filtering approach.

Figure 4 and Figure 5 show the topic distributions for FineWeb2 (Romanian, annotated with our multitask classifier) and FineWeb (English, annotated using WebOrganizer's *Topic Classifier no-URL* (Wettig et al., 2025)) for unfiltered and filtered data (with our method, using a 3.5 threshold). While overall similar, Romanian texts feature higher proportions of *Finance & Business*, *Health*, and *Politics*, whereas English texts have more *Software*, *Software Development*, and *Education & Jobs* content.

A comparative analysis following the filtering process reveals a disparity between the datasets. The Romanian corpus is notably deficient in *Science, Math & Technology* texts compared to its English counterpart, while exhibiting an overrepresentation of topics such as *Finance & Business* and *Food & Dining*. Interestingly, *Food & Dining* has similar proportions in both unfiltered datasets, but the Romanian classifier amplifies its importance, while the English one diminishes it.

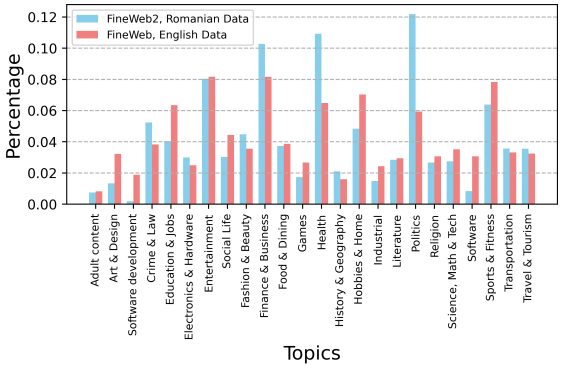

Figure 4: Topic distribution of unfiltered texts.

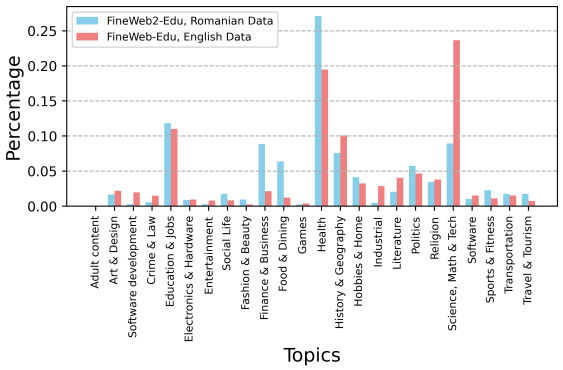

Figure 5: Topic distribution of filtered texts.

6 Conclusions

We surpass standard educational value extraction by augmenting large-scale datasets with rich metadata covering topic, format, and educational level. This enrichment allows for precise control over data filtering and facilitates meaningful cross-lingual comparisons. When applied to FineWeb2, this strategy drove substantial improvements in Llama-2 pretraining performance, demonstrating its effectiveness.

Finally, we release valuable resources, including datasets and models, with the aim of fostering further research on efficient LLM training for low-resource languages.

Limitations

Filtering reduces dataset size below what is needed for optimal large model training per the Chinchilla scaling law (Hoffmann et al., 2022). Our study specifically employs a strict quality threshold (3.5) to enable controlled comparisons with existing high-quality baselines. While this ensures high data density, it discards a substantial volume of text. Future work should explore the trade-off between data quantity and quality by evaluating performance curves across less restrictive thresholds, which may be beneficial for maximizing token volume in low-resource settings where larger datasets are required.

Additionally, limited format diversity in the current training data affects classifier accuracy on out-of-distribution texts. Future work will focus on expanding data diversity to improve robust generalization across different text types.

Acknowledgments

This research is supported by the project "Romanian Hub for Artificial Intelligence - HRIA", Smart Growth, Digitization and Financial Instruments Program, 2021-2027, MySMIS no. 334906.

References

AI@Meta. 2024. Llama 3 model card.

Mehdi Ali, Manuel Brack, Max Lübbering, Elias Wendt, Abbas Goher Khan, Richard Rutmann, Alex Jude, Maurice Kraus, Alexander Arno Weber, Felix Stollenwerk, David Kaczér, Florian Mai, Lucie Flek, Rafet Sifa, Nicolas Flores-Herr, Joachim Köhler, Patrick Schramowski, Michael Fromm, and Kristian Kersting. 2025. Judging quality across languages: A multilingual approach to pretraining data filtering with language modelss. *arXiv preprint arXiv:2505.22232*.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Kelly Marchisio, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress. *Preprint*, arXiv:2405.15032.

Tianyi Bai, Ling Yang, Zhen Hao Wong, Fupeng Sun, Xinlin Zhuang, Jiahui Peng, Chi Zhang, Lijun Wu, Qiu Jiantao, Wentao Zhang, and 1 others. 2025. Efficient pretraining data selection for language models via multi-actor collaboration. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9465–9491.

Vlad-Andrei Bădoiu, Mihai-Valentin Dumitru, Alexandru M. Gherghescu, Alexandru Agache, and Costin Raiciu. 2024. Fulg: 150b romanian corpus for language model pretraining. *Preprint*, arXiv:2407.13657.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *Preprint*, arXiv:1803.05457.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Stefan Dumitrescu, Andrei-Marius Avram, and Sampo Pyysalo. 2020. The birth of Romanian BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4324–4328, Online. Association for Computational Linguistics.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, and 24 others. 2024. Olmo: Accelerating the science of language models. *Preprint*, arXiv:2402.00838.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, and 3 others. 2022. Training compute-optimal large language models. *Preprint*, arXiv:2203.15556.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ippolito. 2024. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3245–3276, Mexico City, Mexico. Association for Computational Linguistics.

Anton Lozhkov, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. 2024. Fineweb-edu: the finest collection of educational content.

Mihai Masala, Denis Ilie-Ablachim, Alexandru Dima, Dragos Georgian Corlatescu, Miruna-Andreea Zavelca, Ovio Olaru, Simina-Maria Terian, Andrei

Terian, Marius Leordeanu, Horia Velicu, Marius Popescu, Mihai Dascalu, and Traian Rebedea. 2024. "vorbești românește?" a recipe to train powerful Romanian LLMs with English instructions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11632–11647, Miami, Florida, USA. Association for Computational Linguistics.

Mihai Masala, Stefan Ruseti, and Mihai Dascalu. 2020. Robert–a romanian bert model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6626–6637.

Luke Merrick, Danmei Xu, Gaurav Nuti, and Daniel Campos. 2024. Arctic-embed: Scalable, efficient, and accurate text embedding models. *Preprint*, arXiv:2405.05374.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. *Preprint*, arXiv:1809.02789.

Mistral AI. 2025. Mistral Small 24B Instruct 2501. `https://huggingface.co/mistralai/Mistral-Small-24B-Instruct-2501`. Accessed: [Current Date, e.g., 2025-06-04].

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*.

Thao Nguyen, Yang Li, Olga Golovneva, Luke Zettlemoyer, Sewoong Oh, Ludwig Schmidt, and Xian Li. 2025. Recycling the web: A method to enhance pre-training data quality and quantity for language models. *Preprint*, arXiv:2506.04689.

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.

Stephan Oepen, Nikolay Arefev, Mikko Aulamo, Marta Bañón, Maja Buljan, Laurie Burchell, Lucas Charpentier, Pinzhen Chen, Mariya Fedorova, Ona de Gibert, Barry Haddow, Jan Hajič, Jindřich Helcl, Andrey Kutuzov, Veronika Laippala, Zihao Li, Risto Luukkonen, Bhavitvya Malik, Vladislav Mikhailov, and 13 others. 2025. Hplt 3.0: Very large-scale multilingual resources for llm and mt. mono- and bi-lingual data, multilingual evaluation, and pre-trained models. *Preprint*, arXiv:2511.01066.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.

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. *Preprint*, arXiv:2406.17557.

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, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Gemma Team, Aishwarya Kamath, and Johan Ferret et al. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Gemma Team, Morgane Riviere, Shreya Pathak, and Pier Giuseppe Sessa et al. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Hugo Touvron, Louis Martin, and Kevin Stone et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Alexander Wettig, Kyle Lo, Sewon Min, Hannaneh Hajishirzi, Danqi Chen, and Luca Soldaini. 2025. Organize the web: Constructing domains enhances pre-training data curation. *Preprint*, arXiv:2502.10341.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *Preprint*, arXiv:1905.07830.

Xinlin Zhuang, Jiahui Peng, Ren Ma, Yinfan Wang, Tianyi Bai, Xingjian Wei, Jiantao Qiu, Chi Zhang, Ying Qian, and Conghui He. 2025. Meta-rater: A multi-dimensional data selection method for pre-training language models. *Preprint*, arXiv:2504.14194.

A Taxonomy Definition

The comprehensive list of topics and formats is adopted from WebOrganizer (Wettig et al., 2025) (we are interested in cross-lingual distribution comparisons, so we use their classifiers for English texts).

Topics (24): *Adult Content*, *Art & Design*, *Software Development*, *Crime & Law*, *Education & Jobs*, *Electronics & Hardware*, *Entertainment*, *Social Life*, *Fashion & Beauty*, *Finance & Business*, *Food & Dining*, *Games*, *Health*, *History & Geography*, *Home & Hobbies*, *Industrial*, *Literature*,

Politics, *Religion*, *Science, Math & Technology*, *Software*, *Sports & Fitness*, *Transportation*, and *Travel & Tourism*.

Formats (24): *Academic Writing*, *Content Listing*, *Creative Writing*, *Customer Support Page*, *Discussion Forum / Comment Section*, *FAQs*, *Incomplete Content*, *Knowledge Article*, *Legal Notices*, *Listicle*, *News Article*, *Nonfiction Writing*, *Organizational About Page*, *Organizational Announcement*, *Personal About Page*, *Personal Blog*, *Product Page*, *Q&A Forum*, *Spam / Ads*, *Structured Data*, *Technical Writing*, *Transcript / Interview*, *Tutorial / How-To Guide*, and *User Reviews*.

We propose the following classification for education level (6): *Preschool*, *Primary School*, *Middle School*, *High School*, *Bachelor's Degree*, and *Postgraduate*.

B Choosing the LLM

Table 5 provides the main metrics that influenced the decision of using Gemma3-12B (Team et al., 2025) for the creation of the training dataset that was further distilled.

C Lightweight Classifier Setup

Table 6 summarizes the lightweight model architecture.

D Impact of Filtering Thresholds

The results in Table 7 demonstrate that applying different filters drastically changes the size of the dataset. As expected, there are far fewer high-quality texts available than low-quality ones.

E Educational Level Distribution

In Figure 6, we present the distribution of educational levels as predicted by our trained classifier on the FineWeb2 Romanian split. A substantial majority of the data (over 80%) corresponds to primary and middle school levels. In contrast, data corresponding to education levels beyond high school account for only about 1%.

Distribution of educational level after filtering (with the same 3.5 threshold on educational value) is presented in Figure 7. We note that, documents related to lower educational levels (preschool and primary school) are largely filtered out, resulting in a notable decrease in their overall proportion, whereas those associated with higher educational levels are more prevalent in the dataset.

Model (prompt lang.)	Edu. RMSE	Topic Acc.	Err.
Llama3.3-70B (en)	1.00	0.72	**0**
Llama3.3-70B (ro)	1.48	0.72	**0**
Llama3-70B (en)	1.42	0.71	21
Llama3-70B (ro)	1.61	0.69	35
Llama3.1-70B (en)	1.25	**0.73**	**0**
Llama3.1-70B (ro)	1.37	0.70	2
Llama3.1-8B (en)	1.03	0.52	5
Llama3.1-8B (ro)	1.23	0.45	17
Llama3-8B (en)	1.21	0.49	8
Llama3-8B (ro)	1.33	0.56	34
Gemma2-27B (en)	0.97	0.61	27
Gemma2-27B (ro)	1.04	0.58	23
Gemma2-9B (en)	0.99	0.57	71
Gemma2-9B (ro)	1.06	0.60	16
Gemma3-27B (en)	1.08	**0.73**	**0**
Gemma3-27B (ro)	1.26	**0.73**	1
Gemma3-12B (en)	1.02	0.68	1
Gemma3-12B (ro)	**0.96**	0.69	**0**
CohereAya-35B (en)	1.26	0.60	10
CohereAya-35B (ro)	1.35	0.50	29
Qwen2.5-72B (en)	1.35	0.72	**0**
Qwen2.5-72B (ro)	1.38	0.70	6
Mistral-S-24B (en)	1.07	**0.73**	5
Mistral-S-24B (ro)	1.03	0.70	7

Table 5: Evaluation of multiple models with both Romanian and English prompts on a manually annotated dataset. Reported metrics include root mean squared error for educational value, accuracy for topic classification, and the number of errors (instances where the model disregarded the instructions).

Configuration Parameter	Value
Encoder	RoBERT-base
Additional layer size	256
Tasks	All
Learning Rate	$1 \cdot 10^{-4}$
Encoder Learning Rate	$3 \cdot 10^{-6}$
Training Set Size	1M
Epochs	3
α (loss parameter)	0.8

Table 6: Lightweight classifier hyperparameters.

F Training Details

All training runs were performed on an NVIDIA DGX H100 equipped with 8 GPUs. Continual pretraining on the unfiltered data needed 656 GPU hours, while training on FineWeb2-RoEdu (our filtered data) needed 608 hours with training on

Threshold	#Tokens	#Samples
2.0	31.60B	18.7M
2.5	22.55B	12.0M
3.0	15.23B	7.3M
3.5	9.15B	3.9M
4.0	2.66B	1.0M

Table 7: Token counts for multiple filtering thresholds, without any truncation.

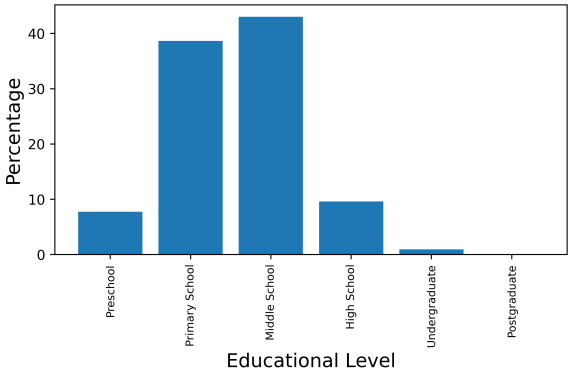

Figure 6: Educational level distribution of FineWeb2 Romanian split.

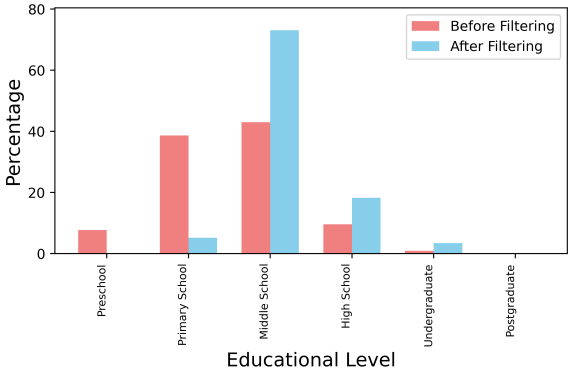

Figure 7: Educational level distribution of FineWeb2 Romanian split before and after filtering.

FineWeb2-JQL requiring 584 GPU Hours. All training runs were performed using the same hyperparameters: cosine learning rate schedule with $1e-4$ peak and $1e-5$ minimum learning rate and 5% warmup ratio; 4096 cutoff length, no packing; 64 batch size with 16 gradient accumulation steps, leading to an effective batch size of 1024. Due to resource constraints, we perform and report results on a single run.

G Annotation Details

The human annotations (100 documents) were conducted by a Romanian native MSc student in Artificial Intelligence who volunteered for the task. For evaluating education value, we translated the FineWeb-Edu English instructions[5], while for topic and format we translated the original WebOrganizer prompts (Wettig et al., 2025). The educational value prompt, with explanations for each category, is presented (translated) in Figure 8.

•**Preschool** – focused on social, emotional, cognitive, and physical development through play, creative activities, and social interactions.
•**Elementary school** – emphasizes literacy, basic mathematics, introduction to science, and development of fundamental skills.
•**Middle school** – deepens subjects from elementary school and introduces new disciplines (e.g., physics, chemistry, biology).
•**High school** – offers various tracks (science, humanities, technical) and prepares students for higher education or entering the workforce.
•**Bachelor's degree** – the first level of higher education. Focuses on an academic or professional discipline and provides foundational knowledge and practical skills.
•**Graduate degree** – includes advanced studies such as master's, doctorate, postdoctoral research, specialization, or residency. Involves rigorous training in a specific field.

Figure 8: Translated educational level prompt.

[5] https://huggingface.co/HuggingFaceFW/fineweb-edu-classifier/blob/main/utils/prompt.txt