

# Is a Document Educational or Just Wikipedia-Style? – Pitfalls of Classifier-Based Quality Filtering

Mateusz Klimaszewski<sup>1</sup>, Piotr Andruszkiewicz<sup>1,2</sup>

<sup>1</sup>Warsaw University of Technology, <sup>2</sup>IDEAS Research Institute  
firstname.lastname@pw.edu.pl

## Abstract

Classifier-based Quality Filtering has recently emerged as a fundamental technique in constructing pre-training corpora. The ability to deploy a single model that can replace or supplement a set of heuristics has proven effective across numerous Large Language Models. In this work, we expose a critical vulnerability in this approach by demonstrating how a straightforward Wikipedia-style reformatting operation can substantially alter a model’s quality assessment and enable low-quality content to surpass filtering thresholds. Our analysis reveals that the FineWeb-Edu CQF model would reverse its filtering decision for approximately 7% of evaluated documents, thereby admitting content into the pre-training corpus that would otherwise have been excluded.<sup>1</sup>

## 1 Introduction

Large Language Models (LLMs) are pre-trained on massive data corpora, and the quality of these corpora is one of the main factors in achieving state-of-the-art performance. The standard elements on the majority of pipelines include a series of heuristic filters (e.g., language identification, near-deduplication, characters-to-other-signs ratio (Wenzek et al., 2020; Burchell et al., 2025)) that aim to improve downstream performance. An alternative to or supplement of heuristic filters is Classifier-based Quality Filtering (CQF) (Penedo et al., 2024). This solution offloads the task of filtering to pre-trained models that provide a quality score, which is used not only to filter the pre-training corpora but also to design a curriculum based on the scores. Nowadays, CQF is a cornerstone of many prominent datasets (Fineweb-Edu (Penedo et al., 2024), DCLM (Li et al., 2024), Nemotron-CC (Su et al., 2025)) and LLMs (e.g. LLama 3 (Grattafiori et al., 2024), EuroLLM (Martins et al., 2025), SmolLM3

<sup>1</sup>Repository:  
<https://github.com/mklimasz/cqf-pitfalls>

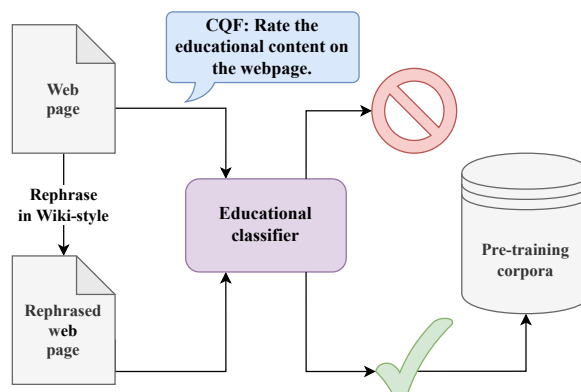


Figure 1: An educational classifier can be manipulated to assign a higher score to Wiki-style rephrased text, thereby passing the sample to the pre-training data.

(Bakouch et al., 2025)), showing its usefulness not only in monolingual (mostly English-based filtering) but also multilingual scenarios (Martins et al., 2025; Waldendorf et al., 2025).

CQF operates on the assumption that educational data enhances the performance of the LLM (Gunasekar et al., 2023; Abdin et al., 2024) and aims to filter such data from raw corpora. Filtering operates through a proxy; i.e., an "educational" quality classifier is developed from synthetic annotations generated by an LLM, e.g., LLama-3-70B (Penedo et al., 2024). The proxy, typically a BERT-sized model, is necessary to operate on a pre-training scale and make filtering feasible due to computational overhead. The "educational" score (typically a value between 0 and 5) is assigned based on specific criteria by the LLM (see Appendix A), and the proxy model is trained to replicate its behaviour.

In this work, we challenge the "educational" aspect of the obtained datasets and demonstrate (Sections 3.1 and 3.2) that CQFs can be Wikipedia-like or style-oriented and biased tools for filtering data, where the "educational" aspect may not be the sole outcome, as presented in Figure 1. Our findings align with the related work of Saada et al. (2025),

which highlights the limitations of CQF in between high-quality and low-quality examples.

The potential effects of CQF pitfalls should not be overlooked. Unlike instruction data, which is heavily filtered and, at times, manually verified, the pre-training scale makes the data, to some degree, unverifiable. Malicious actors could leverage these limitations to bypass filtering and promote their agenda by manipulating web corpora (Carlini et al., 2024). Moreover, recent studies have shown that a precisely injected backdoor can lead to permanent damage that persists even after the LLM’s alignment phase (Zhang et al., 2025).

## 2 Related Work

Model-based filtering has played a crucial role in constructing large-scale pre-training corpora (Wenzek et al., 2020). Traditional n-gram-based approaches, including FastText (Barrios et al., 2009) for language identification (NLLB Team et al., 2022; Burchell et al., 2023) and KenLM (Heafield, 2011) for perplexity-based filtering, have played foundational roles in ensuring corpus quality and relevance. As LLMs have scaled in size and capability, BERT-based models (Devlin et al., 2019; Conneau et al., 2020) have emerged as comparatively “lightweight” alternatives, gradually replacing simpler statistical and shallow neural network methods.

This transition is especially noticeable in the Machine Translation field, where Quality Estimation (QE) models such as COMET (Rei et al., 2022, 2023) and MetricX (Juraska et al., 2024) have become established tools for filtering parallel corpora. However, despite the advances these QE models have enabled in translation quality (Specia et al., 2021; Zerva et al., 2022; Blain et al., 2023; Zerva et al., 2024), subsequent research has revealed their limitations and identified unintended biases that filtering can introduce into downstream models (Zarnis et al., 2025).

CQF has expanded the scope of model-based filtering beyond parallel corpora to orders of magnitude larger pre-training datasets. While CQF methods have demonstrated effectiveness in producing English-centric LLMs and have shown recent promise in multilingual applications (Martins et al., 2025; Waldendorf et al., 2025), questions remain regarding their boundaries and limitations (Saada et al., 2025). These unresolved concerns challenge the presumed universality of CQF as a

filtering paradigm and require further investigation. Finally, while Wikipedia-style paraphrasing has been utilised to enhance low-quality data, prior works (Maini et al., 2024; Su et al., 2025) have employed it as a post-processing step applied solely to data that has already undergone filtering, and without explicit guidelines defining the extent of permissible modifications.

## 3 CQF Pitfalls

### 3.1 Wikipedia-style rephrasing

Our Wikipedia-style rephrasing experiment, illustrated in Figure 1, comprises two distinct stages. In the first stage, we transform the original webpage content into a Wikipedia-style format, with a primary focus on restructuring the presentation to align with the conventional characteristics of Wikipedia articles. In the second stage, we apply CQF models to evaluate both the original and rephrased versions of the text, thereby generating filtering decisions for each variant. This two-stage methodology enables us to isolate the impact of formatting conventions on quality assessment.

In the study, we investigate three CQF models: the original FineWeb-Edu<sup>2</sup> (Penedo et al., 2024) model and two models derived from Nemotron-CC, specifically NemoCurator Mixtral<sup>3</sup> and NemoCurator Nemotron<sup>4</sup> (Su et al., 2025). All three models utilise the Snowflake-Arctic-Embed-M embedding model as their foundation, a BERT-based model that contains 110 million parameters and supports a context window of 512 tokens (Merriam-Webster, 2024). Our analysis is based on a randomly sampled subset of 100,000 examples drawn from the FineWeb corpora (Penedo et al., 2024). To generate Wikipedia-style paragraphs, we employ the Qwen 2.5 72B Instruct<sup>5</sup> model (Yang et al., 2024), which receives instructions to restructure the text webpage content while preserving the original information and modifying only the presentation format. The complete prompt used for this rephrasing task is available in Appendix B and the examples of raw and Wikipedia-style rephrased documents can be found in Appendix C.

Figure 2 presents the distributions of CQF scores

<sup>2</sup><https://hf.co/HuggingFaceFW/fineweb-edu-classifier>

<sup>3</sup><https://hf.co/nvidia/nemocurator-fineweb-mixtral-edu-classifier>

<sup>4</sup><https://hf.co/nvidia/nemocurator-fineweb-nemotron-4-edu-classifier>

<sup>5</sup><https://hf.co/Qwen/Qwen2.5-72B-Instruct>

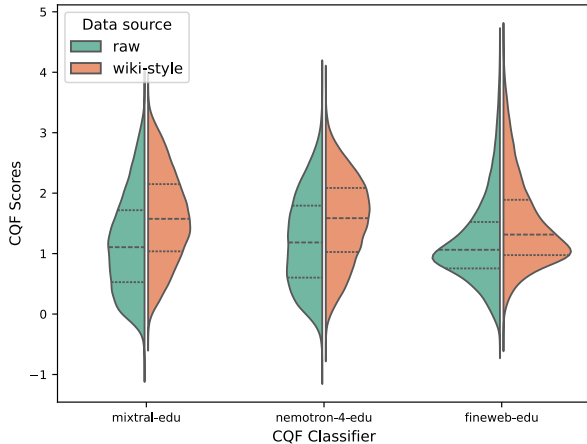


Figure 2: Wikipedia-style rephrasing impact on the CQF models’ scores. The original documents (named “raw” - green, left part of the violin plots) score significantly lower for all evaluated models compared to their Wikipedia-style modified counterparts (orange, right).

assigned to both the original and rephrased versions of the text. All three models exhibit consistent behaviour, assigning higher scores to Wikipedia-style rephrasing. As shown in Table 1, FineWeb-Edu demonstrates the highest robustness among the evaluated models, evidenced by the smallest disparity between scores for raw and Wikipedia-style content. Nevertheless, all models allow a substantial proportion of reformatted data to pass through their filters. Specifically, at a filtering threshold of 3, NemoCurator Mixtral admits over 7% of the rephrased data that scores two or worse in the original form. The NemoCurator Nemotron and FineWeb-Edu admit 5% and 6%, respectively. Furthermore, even with a more rigorous threshold of 4, FineWeb-Edu fails to filter approximately 1% of the Wikipedia-style data, which consistently receives elevated scores. Hence, while FineWeb-Edu performs best on average, it proves least effective in practice at eliminating false positives among the most highly-scored examples.

### 3.2 Domain sensitivity

CQF models demonstrate bias not only toward Wikipedia-style articles but also toward particular content domains. For this phase of our analysis, we focus exclusively on the FineWeb-Edu model. We annotated the FineWeb corpus using Nvidia’s domain classifier<sup>6</sup> (He et al., 2023), which assigns text to 26 distinct categories, and sampled 20,000 examples from each domain.

<sup>6</sup><https://hf.co/nvidia/domain-classifier>

CQF Model	Mean score	
	raw	wiki-style
FineWeb-Edu	1.19	1.49
NemoCurator Mixtral	1.17	1.60
NemoCurator Nemotron	1.18	1.59

Table 1: CQF models’ scores on the 100,000 examples without and with the Wikipedia-style rephrasing. While, on average, the FineWeb-Edu model is the most robust, as the difference between the original (raw) document and its rephrased version is the smallest, in practice, it is “the worst” choice (see Section 3.1 for details).

Figure 3 presents the score distributions for both original and Wikipedia-style documents across these domains. The results reveal both anticipated behaviours and problematic patterns in CQF scoring. As expected, domains such as “Science”—which intuitively should receive high-quality ratings—do indeed score among the highest. However, the rephrasing intervention consistently produces substantial score increases across all domains. In none of the 26 domains do the rephrased versions receive scores comparable to their original counterparts on average. The effect might reach substantial magnitude in certain domains; for instance, in “Books and Literature”, the first quartile of Wikipedia-style rephrased documents nearly reaches the median score of the original documents.

Finally, comparing the distributions across domains, we can clearly see that “educational” filtering has a preference side effect. There is a definite over-representation of specific domains given a standard 3 or 4 threshold. While the result is not directly a pitfall of the CQF, it may end up being one if our downstream domain falls on the lower end of the model’s preference. This finding aligns with the recent Nemotron-CLIMB (Diao et al., 2025) work, which highlights the limitations of CQF in specific domains and constructs pre-training corpora by searching for optimal pre-training data mixtures within predefined domains.

## 4 Manual “educational” score annotation

To determine whether the observed limitations of CQF models originate from the models themselves or from their training data, we revisited the LLM labelling process. Given the “educational” prompt of FineWeb-Edu (see Appendix A), we instructed human annotators to evaluate webpages according

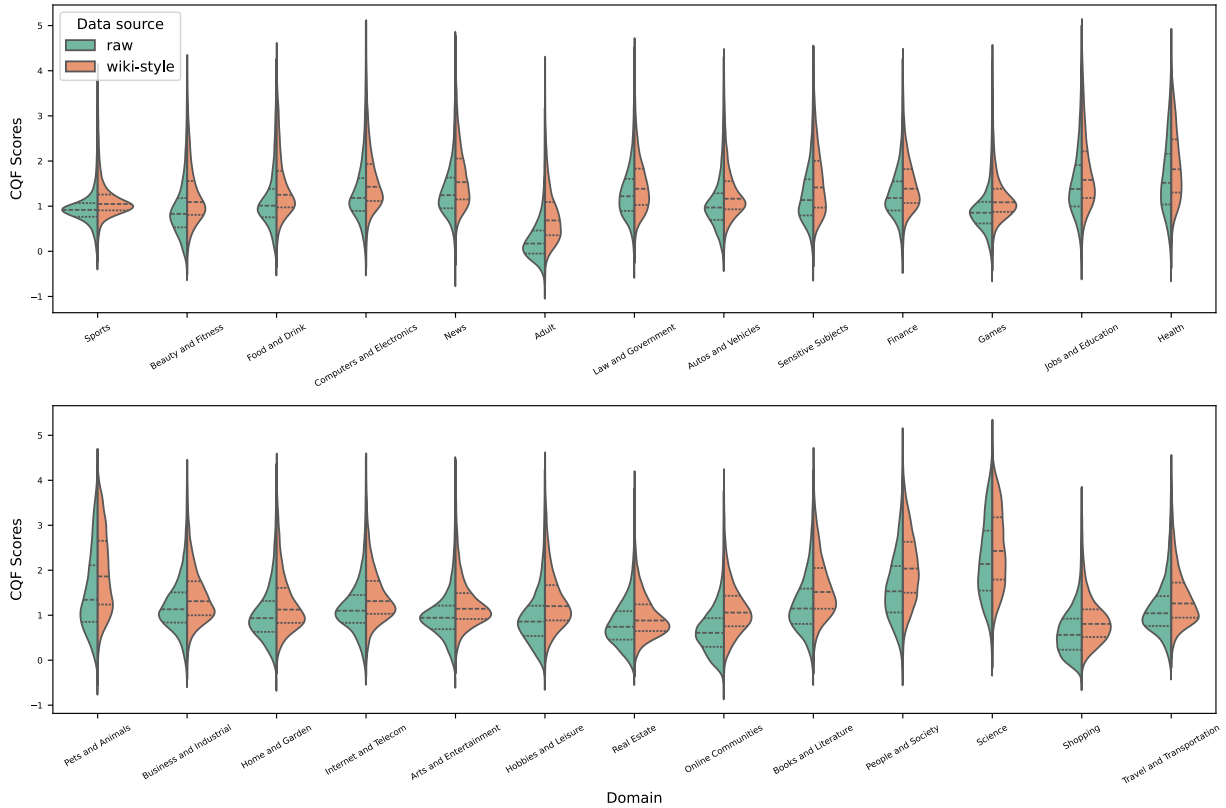


Figure 3: Wikipedia-style rephrasing impact on the FineWeb-Edu CQF model across domains. In each of 26 domains, the original documents score lower than their Wikipedia-style modified counterparts. Moreover, the distributions indicate that the "educational" approach implies a preference towards specific domains.

to the same additive criteria employed in the original LLM-based annotation task. The annotation corpora include 100 documents, each of which was rated by three independent annotators<sup>7</sup> according to the prompt. To curate the annotation corpora, we chose samples with the highest difference of the classifier score of the original and rephrased versions of the text. Then we sampled across all the possible scores to form a uniform distribution across the LLM-based scores.

The results reveal that a significant portion of the CQF models' shortcomings can be traced back to the LLM annotation process itself. Human annotators assigned "educational" scores that averaged 0.77 points lower than those generated by the LLM annotator (Llama 3.1 70B Instruct<sup>8</sup> (Grattafiori et al., 2024), consistent with the FineWeb approach). This discrepancy indicates that the CQF model, as a student model of LLMs, faces an even more challenging task in learning the meaning of rating educational content, given that its training

data is inherently noisy, reflecting the biases and limitations of the teacher LLM.

## 5 Conclusion

In this paper, we identify a critical vulnerability inherent to Classifier-based Quality Filtering models. We demonstrate that simple Wikipedia-style reformatting of documents can systematically bypass filtering thresholds, enabling content that would otherwise be rejected to gain access into pre-training corpora. Furthermore, we establish that this phenomena manifest consistently across different domains, while confirming prior findings indicating that CQF models exhibit domain-dependent biases in their quality assessments. Our findings suggest that the "educational" quality scores assigned by the CQF teacher model may be inflated, providing a judgment that is overly optimistic regarding the actual value of the data.

Although CQF has undeniably contributed to the development of higher-performing LLMs, we hope that our findings will inform the evaluation and design of future CQF architectures. By presenting these limitations, we aim to mitigate the

<sup>7</sup>The group of human annotators consist of master's students from the NLP course.

<sup>8</sup><https://hf.co/meta-llama/Llama-3.1-70B-Instruct>

risk of introducing unintended biases and malicious data injections that could compromise downstream model behaviour and capabilities.

## Acknowledgements

This research was funded in whole by the National Science Centre, Poland 2023/49/N/ST6/02691. For the purpose of Open Access, the authors has applied a CC-BY public copyright licence to any Author Accepted Manuscript (AAM) version arising from this submission. We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Center: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2025/018209.

## Limitations

Our analysis is based on automatic reformatting existing web corpora that must yield inaccuracies and include noisy outputs. Although scale and simplicity of the requested rephrasing should limit the variance of presented results, the exact numbers are imprecise and are an approximation of the actual problem degree.

## References

- Marah Abidin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, and 110 others. 2024. [Phi-3 technical report: A highly capable language model locally on your phone](#). *Preprint*, arXiv:2404.14219.
- Elie Bakouch, Loubna Ben Allal, Anton Lozhkov, Noumane Tazi, Lewis Tunstall, Carlos Miguel Patino, Edward Beeching, Aymeric Roucher, Aksel Joonas Reedi, Quentin Gallouédec, Kashif Rasul, Nathan Habib, Clémentine Fourrier, Hynek Kydlicek, Guilherme Penedo, Hugo Larcher, Mathieu Morlon, Vaibhav Srivastav, Joshua Lochner, and 4 others. 2025. [SmolLM3: smol, multilingual, long-context reasoner](#). <https://huggingface.co/blog/smollm3>.
- María A. Barrios, Guadalupe Aguado de Cea, and José Ángel Ramos. 2009. [Enriching a lexicographic tool with domain definitions: Problems and solutions](#). In *Proceedings of the 1st Workshop on Definition Extraction*, pages 14–20, Borovets, Bulgaria. Association for Computational Linguistics.
- Frederic Blain, Chrysoula Zerva, Ricardo Rei, Nuno M. Guerreiro, Diptesh Kanojia, José G. C. de Souza, Beatriz Silva, Tânia Vaz, Yan Jingxuan, Fatemeh Azadi, Constantin Orasan, and André Martins. 2023. [Findings of the WMT 2023 shared task on quality estimation](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 629–653, Singapore. Association for Computational Linguistics.
- Laurie Burchell, Alexandra Birch, Nikolay Bogoychev, and Kenneth Heafield. 2023. [An open dataset and model for language identification](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 865–879, Toronto, Canada. Association for Computational Linguistics.
- Laurie Burchell, Ona de Gibert, Nikolay Arefyev, Mikko Aulamo, Marta Bañón, Pinzhen Chen, Mariia Fedorova, Liane Guillou, Barry Haddow, Jan Hajič, Jindřich Helcl, Erik Henriksson, Mateusz Klimaszewski, Ville Komulainen, Andrey Kutuzov, Joonas Kytöniemi, Veronika Laippala, Petter Mæhlum, Bhavitvya Malik, and 16 others. 2025. [An expanded massive multilingual dataset for high-performance language technologies \(HPLT\)](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17452–17485, Vienna, Austria. Association for Computational Linguistics.
- Nicholas Carlini, Matthew Jagielski, Christopher A. Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. 2024. [Poisoning Web-Scale Training Datasets is Practical](#). In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 407–425, Los Alamitos, CA, USA. IEEE Computer Society.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Shizhe Diao, Yu Yang, Yonggan Fu, Xin Dong, Dan SU, Markus Kliegl, ZIJIA CHEN, Peter Belcak, Yoshi Suhara, Hongxu Yin, Mostofa Patwary, Yingyan Celine Lin, Jan Kautz, and Pavlo Molchanov. 2025. [Nemotron-CLIMB: Clustering-based iterative data mixture bootstrapping for language model pre-training](#). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

- 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, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. [Textbooks are all you need](#). *Preprint*, arXiv:2306.11644.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. [DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing](#). In *The Eleventh International Conference on Learning Representations*.
- Kenneth Heafield. 2011. [KenLM: Faster and smaller language model queries](#). In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.
- Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. [MetricX-24: The Google submission to the WMT 2024 metrics shared task](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.
- Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Yitzhak Gadre, Hritik Bansal, Etash Kumar Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee F Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, and 40 others. 2024. [Datacomp-LM: In search of the next generation of training sets for language models](#). In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Pratyush Maini, Skyler Seto, Richard Bai, David Grangier, Yizhe Zhang, and Navdeep Jaitly. 2024. [Rephrasing the web: A recipe for compute and data-efficient language modeling](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14044–14072, Bangkok, Thailand. Association for Computational Linguistics.
- Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M. Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M. Alves, José Pombal, Nicolas Boizard, Manuel Faysse, Pierre Colombo, François Yvon, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2025. [Eurollm-9b: Technical report](#). *Preprint*, arXiv:2506.04079.
- Luke Merrick. 2024. [Embedding and clustering your data can improve contrastive pretraining](#). *Preprint*, arXiv:2407.18887.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation.
- 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](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 30811–30849. Curran Associates, Inc.
- Ricardo Rei, Nuno M. Guerreiro, Josã© Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. 2023. [Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore. Association for Computational Linguistics.
- Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. [CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Thiziri Nait Saada, Louis Bethune, Michal Klein, David Grangier, Marco Cuturi, and Pierre Ablin. 2025. [The data-quality illusion: Rethinking classifier-based quality filtering for llm pretraining](#). *Preprint*, arXiv:2510.00866.
- Lucia Specia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André F. T. Martins. 2021. [Findings of the WMT 2021 shared task on quality estimation](#). In *Proceedings of the Sixth Conference on Machine Translation*, pages 684–725, Online. Association for Computational Linguistics.
- Dan Su, Kezhi Kong, Ying Lin, Joseph Jennings, Brandon Norrick, Markus Kliegl, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2025. [Nemotron-CC: Transforming Common Crawl into a refined long-horizon pretraining dataset](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2459–2475, Vienna, Austria. Association for Computational Linguistics.

Jonas Waldendorf, Barry Haddow, Alexandra Birch, and Mateusz Klimaszewski. 2025. [Multilingual data filtering using synthetic data from large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 9317–9334, Suzhou, China. Association for Computational Linguistics.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. [CCNet: Extracting high quality monolingual datasets from web crawl data](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

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, and 22 others. 2024. [Qwen2.5 technical report](#). *CoRR*, abs/2412.15115.

Emmanouil Zaranis, Giuseppe Attanasio, Sweta Agrawal, and Andre Martins. 2025. [Watching the watchers: Exposing gender disparities in machine translation quality estimation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25261–25284, Vienna, Austria. Association for Computational Linguistics.

Chrysoula Zerva, Frederic Blain, José G. C. De Souza, Diptesh Kanojia, Sourabh Deoghare, Nuno M. Guerreiro, Giuseppe Attanasio, Ricardo Rei, Constantin Orasan, Matteo Negri, Marco Turchi, Rajen Chatterjee, Pushpak Bhattacharyya, Markus Freitag, and André Martins. 2024. [Findings of the quality estimation shared task at WMT 2024: Are LLMs closing the gap in QE?](#) In *Proceedings of the Ninth Conference on Machine Translation*, pages 82–109, Miami, Florida, USA. Association for Computational Linguistics.

Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. de Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, André F. T. Martins, and Lucia Specia. 2022. [Findings of the WMT 2022 shared task on quality estimation](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 69–99, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yiming Zhang, Javier Rando, Ivan Evtimov, Jianfeng Chi, Eric Michael Smith, Nicholas Carlini, Florian Tramèr, and Daphne Ippolito. 2025. [Persistent pre-training poisoning of LLMs](#). In *The Thirteenth International Conference on Learning Representations*.

## A Educational Prompt

Figure 4 presents the FineWeb prompt used for “educational” LLM labelling. We provide

the prompt for reference and the reader’s convenience; the original prompt is available at <https://huggingface.co/HuggingFaceFW/fineweb-edu-classifier/blob/main/utils/prompt.txt>.

## B Wikipedia Rephrasing Prompt

Figure 5 presents prompt used for the Wikipedia-style rephrasing.

## C Wikipedia Rephrasing Examples

Figures 6, 7, 8 present the Wikipedia-style rephrasing examples.

Below is an extract from a web page. Evaluate whether the page has a high educational value and could be useful in an educational setting for teaching from primary school to grade school levels using the additive 5-point scoring system described below. Points are accumulated based on the satisfaction of each criterion:

- Add 1 point if the extract provides some basic information relevant to educational topics, even if it includes some irrelevant or non-academic content like advertisements and promotional material.
- Add another point if the extract addresses certain elements pertinent to education but does not align closely with educational standards. It might mix educational content with non-educational material, offering a superficial overview of potentially useful topics, or presenting information in a disorganized manner and incoherent writing style.
- Award a third point if the extract is appropriate for educational use and introduces key concepts relevant to school curricula. It is coherent though it may not be comprehensive or could include some extraneous information. It may resemble an introductory section of a textbook or a basic tutorial that is suitable for learning but has notable limitations like treating concepts that are too complex for grade school students.
- Grant a fourth point if the extract is highly relevant and beneficial for educational purposes for a level not higher than grade school, exhibiting a clear and consistent writing style. It could be similar to a chapter from a textbook or a tutorial, offering substantial educational content, including exercises and solutions, with minimal irrelevant information, and the concepts aren't too advanced for grade school students. The content is coherent, focused, and valuable for structured learning.
- Bestow a fifth point if the extract is outstanding in its educational value, perfectly suited for teaching either at primary school or grade school. It follows detailed reasoning, the writing style is easy to follow and offers profound and thorough insights into the subject matter, devoid of any non-educational or complex content.

The extract:  
<EXAMPLE>.

After examining the extract:

- Briefly justify your total score, up to 100 words.
- Conclude with the score using the format: "Educational score: <total points>"

Figure 4: Educational prompt listing.

You are a Wikipedia-style rephraser. Your objective is to rephrase the following web page to imitate a Wikipedia article.

Web page:  
```{web\_page}```

Follow the rules below during rephrasing:

- Focus on containing all the facts from the document, even if they are not essential.
- Do not include new facts, concepts and overall new content.
- Keep the exact dates, locations, names and other entities.
- Outcome should differ only in terms of style and formatting.
- The output document should have a similar number of tokens (with a maximum 10% margin).

Figure 5: Wikipedia-style rephrasing prompt listing.

Raw Document:

Teach your students how to encrypt and decrypt binary messages as well as understand binary code. This resource includes a hidden January themed message which students need to decrypt using the key (ASCII Table). Students are also given the option to write their own encrypted binary message.

This is a perfect CS unplugged activity and can be used as an introduction to a lesson, for homework, for early finishers or even when you have no access to computers in the computer lab. This is an ideal activity for the ISTE Standard of Computational Thinking/

Thanks to the following for all images:  
Teachers ClipArt: [www.edu-clips.com](http://www.edu-clips.com)  
Mike Rawls: <http://www.teacherspayteachers.com/Store/Monster-Wrangler-Mike>

Wikipedia-style rephrased document:

### Binary Message Encryption and Decryption Activity

This educational resource is designed to teach students how to encrypt and decrypt binary messages and to understand binary code. The activity includes a hidden message with a January theme, which students must decrypt using the ASCII table. Additionally, students have the option to create their own encrypted binary messages.

This activity serves as an excellent "CS Unplugged" exercise and can be utilized in various educational settings, including as an introduction to a lesson, for homework, for early finishers, or when computers are unavailable in the computer lab. It aligns well with the ISTE Standard of Computational Thinking.

#### Credits

- **Images**: Provided by Teachers ClipArt (<<https://www.edu-clips.com>>) and Mike Rawls (<<https://www.teacherspayteachers.com/Store/Monster-Wrangler-Mike>>)

This resource is ideal for educators looking to introduce fundamental concepts of binary code and encryption in a practical and engaging manner.

Figure 6: Wikipedia-style rephrasing example #1. The transformation reduced CQF score by 0.02 (4.17 vs 4.15).

Raw Document:

Ok, this is one of the most perfect cups of teas I've had. I made this yesterday, but ended up steeping it three times. Rather than have the chocolate filtered, I plucked the chips out of the filter and put them straight in my cup.  
The recommendations suggest a heaping teaspoon per 12 oz. I had about 10 oz of water, so left off the "heaping."  
I think because the chocolate chips were in the bottom on my cup, the sweetness was a little more concentrated towards the bottom of the cup. Most of the cup was subtly sweet and beautifully floral. Then, as I got to the bottom of the cup, the sweetness gets a bit stronger, but still beautiful. One of the best cups of comfort I've had in awhile.  
One the second and third steep, I increased the steeping time to 3 and 4 minutes, respectively.  
Flavors: Dark Chocolate, Rose, White Chocolate

Wikipedia-style rephrased document:

== Preparation and Experience of a Notable Cup of Tea ==

A user recently prepared a cup of tea, which they described as one of the most perfect they have ever had. The tea was prepared and steeped three times over the course of a single day. Instead of using a filter for the chocolate, the chocolate chips were removed from the filter and placed directly into the cup.

The recommended measurement is a heaping teaspoon of tea per 12 ounces of water. However, since the user had approximately 10 ounces of water, they used a level teaspoon instead of a heaping one.

The placement of the chocolate chips at the bottom of the cup resulted in a concentration of sweetness towards the bottom. The majority of the cup was subtly sweet and wonderfully floral, with the sweetness becoming more pronounced towards the end of the cup. The user described it as one of the most comforting cups of tea they have had in a long time.

For the second and third steepings, the steeping times were increased to 3 and 4 minutes, respectively. The notable flavors experienced were dark chocolate, rose, and white chocolate.

Figure 7: Wikipedia-style rephrasing example #2. The transformation increased CQF score by 2.24 (0.52 vs 2.76).

Raw Document:

I'm enjoying my Haiku series a bit too much lol. I would really love for you guys to try and reply my Haiku - use the comment box below.  
You can make it a bit more fun by starting yours with the last line of mine :-)  
Looking forward to reading!  
When the rain drops stop  
Can you tell me what happens?  
A rainbow or clouds...

Wikipedia-style rephrased document:

### Haiku Series and Interactive Engagement

The author expresses their enjoyment of writing haiku and encourages readers to participate by composing their own haiku in the comments section. To make the activity more engaging, the author suggests that readers start their haiku with the last line of the author's haiku. The author looks forward to reading the responses.

#### Sample Haiku

When the rain drops stop,  
Can you tell me what happens?  
A rainbow or clouds...

Figure 8: Wikipedia-style rephrasing example #3. The transformation increased CQF score by 1.83 (-0.04 vs 1.79).