Determining and Evaluating the Quality of Corpora in the LLM Era

Lucía Sevilla-Requena

Dept. of Software and Computing Systems, University of Alicante, Apdo. de Correos 99, E-03080, Alicante, Spain.

Abstract

The present study addresses the need for a systematic and scalable framework to determine the quality of linguistic corpora in the era of Large Language Models (LLMs). As the success of LLMs increasingly depends on the quality of their training data, traditional corpus evaluation methods have become obsolete. This research analyses and proposes a new methodology that defines comprehensive quality criteria, introduces a classification system, and designs interpretable and partially automatable metrics. The resulting framework will be validated through empirical experiments that assess how corpus quality impacts model performance, robustness, and fairness. Therefore, this work aims to bridge the methodological gap in Natural Language Processing (NLP) by providing an updated, reproducible, and practical corpus assessment and creation tool.

Keywords

Corpus Linguistics, Natural Language Processing, Artificial Intelligence, Quality Corpora, Corpus Evaluation, Datasets

1. Justification of the Research

Evaluating the quality of linguistic corpora is a fundamental aspect of their design and usage, as it plays a key role in determining the reliability and overall usefulness of the data [1]. In an era marked by the spread of digital content, accessibility to information is no longer an obstacle. However, the vast abundance of sources presents a new challenge: distinguishing quality data from overwhelming information available. According to Austermühl [2], accessing online data is relatively easy, but finding accurate information can be complex and often frustrating. This highlights the need to establish solid criteria that allow the development and assessment of the quality of digital resources.

In recent years, large language models (LLMs) have marked a profound transformation in the field of Artificial Intelligence (AI) [3]. Although much of this progress is attributed to innovations in model design and training techniques, another crucial factor has gained prominence: revising the criteria used to determine the validity and usefulness of the data employed to train these systems.

Early LLMs highlighted the importance of having coherent and high-quality textual data for training models [4]. To achieve this, they began using structured document-level corpora drawn from specific domains, such as Wikipedia and BookCorpus[1] [5], thus moving away from earlier approaches based on minimal linguistic units, like individual sentences [6]. This shift responded to the need for longer and more contextually cohesive data. As these models increased in scale and complexity, large-scale web scraping became a widely adopted strategy for collecting massive volumes of textual data [7].

The limitations of web-scraped data without human supervision soon became apparent. Studies such as Radford et al. [8] emphasised the importance of data curation and cleaning, showing that carefully selected datasets consistently outperformed raw Web content. This insight led to the creation of so-called "high-quality" corpora such as The Pile [9], which integrates web data with books, scientific articles, and conversations from social media.

In addition, data cleaning has become a critical step in corpus preparation [7], offering additional benefits such as reduced dataset size [10] and more efficient training cycles. Along these lines, a

Doctoral Symposium on Natural Language Processing, 25 September 2025, Zaragoza, Spain.

luciasevillarequena@gmail.com (L. Sevilla-Requena)

0009-0003-8144-8543 (L. Sevilla-Requena)

[1] https://github.com/soskek/bookcorpus?tab=readme-ov-file

recent study by Eldan and Li on TinyStories [11], a synthetically generated dataset for training neural networks in English, demonstrates the potential impact of high-quality data. Their results suggest that well-crafted, high-quality data can alter scaling laws, allowing smaller models to achieve performance levels traditionally associated with much larger systems.

Beyond this, developing domain-specific and multimodal models further reinforces the importance of using specialised data. Corpora focused on specific topics have been successfully used to build biomedical models [12] and conversational models [13].

In this context, the present thesis addresses a critical challenge in the field of Natural Language Processing, which is the need for a systematic and scalable framework for assessing corpus quality in the era of LLMs. As LLMs increasingly advance, it has become clear that the quality of training data is a decisive factor in model performance, fairness, and reliability. Nevertheless, current corpus evaluation practices are often based on outdated or fragmented criteria that are insufficient for the complexity and demands of modern models. This PhD thesis proposes a structured set of quality criteria, a classification methodology, and interpretable metrics that will guide the critical assessment of existing corpora, particularly regarding their suitability for fine-tuning LLMs.

2. Background and Related Work

The present section provides an overview of the existing literature and frameworks relevant to corpus quality evaluation, starting with a conceptual clarification of the term "quality" and progressing through the main criteria developed in the pre-LLM era. Finally, it examines recent advances in data curation and quality assessment in the context of LLMs training, highlighting how modern developments have reshaped traditional evaluation paradigms.

2.1. The concept of quality corpus

Before reviewing the existing criteria for evaluating high-quality corpora, it is essential to clarify what "quality" means and what constitutes a quality corpus. The notion of quality, much like the concept of information, is employed in various contexts, often without a clear or consistent definition [14]. According to the Oxford English Dictionary, quality is "the standard of something when it is compared to other things like it; how good or bad something is" [15]. It is a synonym for the term "excellence", but defining quality solely as that offers a limited practical guide to establish and apply specific criteria in corpus evaluation.

The British Standards Institution (BSI) provides a more functional definition which describes quality as "the totality of characteristics of an entity that bear on its ability to satisfy stated and implied needs" [16]. In other words, the quality reflects how much a product or service fulfils its purpose [14]. Because those needs vary from context to context, there is no absolute quality standard. Instead, the focus is on specific attributes, such as reliability, efficiency, or robustness, directly influencing a resource's ability to satisfy its intended audience. This highlights the importance of establishing clear and appropriate evaluation parameters while building a corpus and ensuring its suitability for the intended purpose.

When turning this towards linguistic corpora, the notion of "quality" again depends on the project's aims. In the words of Wieczorkowska [17], "to obtain a representative and balanced corpus, the purpose of creating a given corpus should be defined first, as should the target user".

In empirical theoretical linguistics, for example, researchers prioritise rigorously controlled sampling and minimally invasive data cleaning to preserve authenticity. In contrast, computational linguistics and language technology applications often require more aggressive preprocessing to optimise algorithmic performance [1]. These different priorities illustrate that the "quality corpus" is a set of context-driven criteria.

2.2. Pre-LLM Evaluation Criteria

Several scholars have proposed criteria frameworks aimed at standardising corpora evaluation (see Appendix). To start, Cooke [14] proposed a comprehensive set of ten parameters: (1) purpose, (2) coverage, (3) authority and reputation, (4) accuracy, (5) currency and maintenance, (6) accessibility, (7) information presentation and layout, (8) ease of use, (9) comparison with other sources, and (10) overall quality. This extensive list reflected an integrative approach to the evaluation of digital content, encompassing both technical and content-related considerations.

In addition, other authors advocated for simpler models, such as Alexander and Tate [18], who suggested five core parameters: authority, coverage, objectivity, accuracy, and currency. These dimensions balance reliability and relevance while ensuring that the content reflects real and current perspectives. Codina [19], by contrast, proposed a slightly different configuration, adding usability-oriented parameters such as ergonomics, brightness, and visibility to more traditional factors like authority and content.

Jiménez Piano and Ortiz-Repiso Jiménez [20] offered another variation by identifying five core aspects in evaluating digital resources: search and retrieval functionalities, authority, content, resource management, and design. These categories highlighted the interplay between user interaction and content quality. Added to this, the comparative analysis conducted by Gordon-Murnane [21] on twelve web evaluation services also underscored this variability. The findings revealed that content was the only universally recognised evaluation parameter, suggesting a significant divergence in how evaluation priorities are conceived and implemented.

Across the literature, certain parameters, such as authority, accuracy, presentation, and currency, emerged as consistently valued, as shown below in Table 1. These are often cited either as stand-alone criteria [18] or as components of broader constructs like "content quality" [22]. Additionally, more recent contributions highlighted the growing importance of design in digital environments. Other scholars, such as Gaffney [23], Adreon et al. [24], and the Pearl K. Wise Library [25], point to visual and structural design as critical factors influencing users' perceptions of quality and usability [26].

However, there is still a lack of consensus on which criteria are the most essential or how they should be prioritised, according to Buendía Castro and Ureña Gómez-Moreno [26]. The variability among proposed frameworks reflects different priorities and evaluative traditions, often shaped by the technological and academic contexts in which they have emerged over the years.

Evaluation criteria	Cooke (1999)	Alexander & Tate (1999)	Codina (2000)	Jiménez Piano & Ortiz-Repiso (2007)	Other authors*
Purpose	x				
Coverage	x	x			
Authority and reputation	x	x	x	x	x
Accuracy	x	x			x
Currency and maintenance	x	x			x
Accessibility	x			x	
Presentation / Layout Design	x		x (visibility, brightness)	x	x
Ease of use / Usability	x		x (ergonomics)	x (search/retrieval)	
Objectivity		x			x
Content			x	x	x
Resource management				x	

Table 1
Comparison of evaluation criteria across different authors. Other authors* include Gaffney (1998), Anderson et al. (1999), Adreon et al. (2002) and the Pearl K. Wise Library (2006).

Lastly, it is crucial to recognise that these evaluation frameworks were developed in a pre-Large Language Model (LLM) era. As valuable as these criteria may be, they emerged before the widespread adoption of Artificial Intelligence tools capable of generating, summarising, interpreting, and even creating corpora and digital content with unprecedented fluency and scale. The emergence of LLMs represents a paradigm shift in how linguistic corpora are produced, accessed, and used.

2.3. Quality corpus criteria in the LLM era

Training LLMs requires huge amounts of textual data, but quantity alone is not enough: Data quality plays a decisive role in achieving good model performance. Although LLMs are often trained on massive aggregated corpora [27], these datasets must carefully find a balance between data volume and quality. Raw sources such as CommonCrawl [28], while abundant, are often noisy and unstructured, making them inefficient and less effective for immediate use in pre-training.

To address this issue, the NLP community has developed refined corpora that apply rigorous filtering and cleaning processes to transform raw data into high-quality, structured training resources. Examples include C4 [29], RedPajama [30], SlimPajama [31] and DCLM-baseline [32], which use techniques such as scoring models, deduplication with MinHash and heuristic rule-based filtering. These data sets represent a major step forward in addressing issues such as redundancy, low linguistic quality, and the presence of irrelevant or toxic content.

More recently, datasets such as RefinedWeb [4], FineWeb [33] and FineWeb-2 [34] have set new benchmarks for data quality, significantly improving the efficiency and effectiveness of LLM pre-training. FineWeb, for instance, was created by Hugging Face using a multi-stage pipeline that includes URL filtering, language classification and custom quality filters targeting line punctuation ratios, short line prevalence, average words per line, n-gram repetition, and document length [33].

Additionally, a seminal contribution by Zhou et al.[35] is the development of LIMA (Less Is More for Alignment), an LLM designed to investigate the art of instruction tuning: rather than the volume of data, it is the quality that dictates the model's performance. Remarkably, LIMA demonstrates that even a limited amount of carefully curated, high-quality data can significantly enhance a model's ability to follow instructions. While this underscores the critical role of data quality, the question of automatically identifying high-quality data from a vast ocean of available datasets remains under investigation.

These advances reflect a broader consensus: while data volume remains important, high-quality, diverse and well-selected sources, often drawn from specific domains such as scientific literature, books or encyclopaedic content, are critical for improving the capabilities of models [36],[37]. Content scraped from the web alone is no longer considered sufficient; rather, careful data selection has become an essential component of the training pipeline.

Lastly, these developments highlight a shift in the way the NLP field approaches data: no longer as a mere volume-driven input, but as a carefully designed and evaluated component that is central to the success of modern LLMs.

3. Main Hypothesis and Objectives

The present thesis addresses a critical methodological gap in NLP: the lack of a systematic, interpretable, and scalable framework for evaluating the quality of linguistic corpora used in the fine-tuning of Large Language Models (LLMs). While the field has significantly progressed in model architecture, scale, and capabilities, corpus evaluation practices have failed to keep pace. Current methods are often based on paradigms that precede LLMs and do not capture the nuanced aspects of data quality essential for the effective adaptation of these models to specific domains and tasks.

This study is grounded in the hypothesis that corpus quality is a key determinant of model performance. A well-constructed, high-quality corpus can significantly enhance a language model's accuracy, robustness, and fairness, whereas low-quality data can result in biased, unreliable, or ineffective systems. Consequently, developing rigorous and clearly defined evaluation methodologies is essential to ensure that existing corpora meet the standards required for fine-tuning LLMs responsibly and effectively.

Given the fundamental role of the corpora's quality in model results, there is an urgent need for an evaluation system that goes beyond technical metrics. Such a system must also integrate linguistic knowledge to identify potential weaknesses or representational gaps in the data, factors that can profoundly influence model performance. The challenge is to create an evaluation framework that is robust, reproducible, scalable and adaptable to the increasing complexity of LLMs.

To this end, the following specific objectives are proposed:

- **O1:** To define and systematise a comprehensive set of quality criteria for linguistic corpora used in fine-tuning LLMs.
- **O2:** To propose a classification methodology that categorises corpora into quality levels (raw, bronze, silver, and gold) according to the defined criteria.
- **O3:** To design an interpretable and, as far as possible, automatable corpus quality metric that operationalises these criteria in measurable terms.
- **O4:** To apply the proposed framework to evaluate existing corpora, thereby demonstrating its applicability.
- **O5:** To validate the framework's utility by analysing existing corpora and conducting fine-tuning experiments to explore the relationship between corpus quality and model performance.

Tho summarise, the present thesis aims to address a notable gap in the literature through these interconnected objectives. It seeks to provide researchers with a scalable, linguistically grounded tool for ensuring the quality for fine-tuning datasets. This applies regardless of the language in which the datasets are created or the domain for which they are intended, ensuring compliance with the stringent requirements of contemporary LLM architectures.

4. Methodology

Following Creswell and Plano Clark [38], this thesis will adopt a methodology based on a mixed approach, combining the theoretical definition of quality criteria with their practical application and empirical validation through experiments with language models fine-tuned on corpora of different quality levels. The research unfolds through the following successive stages:

1. **Literature Review.** This study begins with a comprehensive review of the state of the art of corpus evaluation and data quality in the field of NLP, covering both pre-LLMs frameworks and subsequent emerging proposals. Through this review, it is intended to identify methodological gaps, define the most recurrent quality criteria, and discover possible indicators specifically relevant for corpora used in fine-tuning LLMs.
2. **Examining Corpus Quality Criteria.** Based on the literature review, the most commonly used quality criteria in corpus evaluation (such as authority, accuracy, and coverage) will be identified and systematised. In addition, new indicators will be proposed, specifically designed to address the challenges and particularities posed using corpora in contexts involving large-scale language models (LLMs). Each criterion, both traditional and newly introduced, will be precisely defined and accompanied by a reasoned justification outlining its relevance and impact on the overall quality of the corpus.
3. **Design of a Quality Metric.** Based on the defined criteria, a quantitative evaluation methodology will be developed to assign each corpus a quality level: raw, bronze, silver or gold. This metric will combine automated measures with supervised linguistic evaluations to ensure interpretability and reproducibility.
4. **Evaluation of Existing Corpora.** The proposed metrics will be applied to a selection of reference datasets ranking them according to established quality levels. This will test the metric's behaviour on various types of resources and calibrate the threshold values separating each quality level.
5. **Empirical Validation through Fine-Tuning.** Finally, lightweight, parameter-efficient fine-tuning experiments (e.g., using LoRA or adapters) may be conducted on corpora of different quality levels. The goal is to explore the relationship between corpus quality and LLM behaviour by assessing how differences in data quality influence model performance, robustness, and fairness.

Through this methodology, the present thesis will provide a solid conceptual model for corpus evaluation in the LLM era and the NLP community with effective and empirically validated tools to improve the reliability and fairness of modern language technologies.

5. Research Issues to Discuss

To conclude, this section poses key questions to guide the research, identify gaps and develop a comprehensive framework for assessing corpus quality. Addressing these questions will help refine the focus of the study and contribute to the broader field of NLP by exploring more effective approaches to corpus evaluation for fine-tuning LLMs.

- **RQ1. How can the quality of a corpus be evaluated?** Corpus quality evaluation is central to this research. What are the most effective ways to assess the corpus quality, particularly in the context of fine-tuning large language models? Can existing evaluation methods be adapted to reflect modern datasets' complexities and model requirements more accurately? It is crucial to explore whether traditional evaluation metrics can capture the nuanced aspects of data quality that affect the effectiveness of LLMs, especially as the scale and diversity of corpora continue to grow.

- **RQ2. What methodologies exist for corpus evaluation?** Numerous methodologies have been proposed for evaluating corpus quality, but many are outdated or insufficient for the needs of large language models. What are the strengths and limitations of current evaluation methodologies, and are there other, potentially more suitable methods for assessing corpus quality? This question emphasises the need for an updated and comprehensive approach incorporating technical and linguistic considerations to determine a corpus's suitability for LLM fine-tuning.

- **RQ3. How does corpus quality impact the performance of language models?** The relationship between corpus quality and model performance is crucial. To what extent do variations in data quality affect a model's effectiveness, robustness, and fairness during fine-tuning? Can improvements in corpus quality result in significant gains in the overall performance of a language model? Exploring this dynamic will help determine the role of corpus evaluation in responsible and effective LLM adaptation.

The discussion generated by these questions and additional considerations that may arise throughout the research process will play an essential role in enriching the direction of the PhD thesis.

Declaration on Generative AI

The author has not employed any Generative AI tools.

References

[1] C. Biemann, F. Bildhauer, S. Evert, D. Goldhahn, U. Quasthoff, R. Schäfer, T. Zesch, Scalable construction of high-quality web corpora, Journal for Language Technology and Computational Linguistics 28 (2013) 23–59. doi:10.21248/jlcl.28.2013.175.

[2] F. Austermühl, Electronic Tools for Translators, St. Jerome, Manchester, 2001.

[3] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, 2023. URL: https://arxiv.org/abs/2302.13971. arXiv:2302.13971.

[4] G. Penedo, Q. Malartic, D. Hesslow, R. Cojocaru, A. Cappelli, H. Alobeidli, B. Pannier, E. Almazrouei, J. Launay, The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only, 2023. URL: https://arxiv.org/abs/2306.01116. arXiv:2306.01116.

[5] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, S. Fidler, Aligning books and movies: Towards story-like visual explanations by watching movies and reading books, 2015. URL: https://arxiv.org/abs/1506.06724. arXiv:1506.06724.

[6] C. Chelba, T. Mikolov, M. Schuster, Q. Ge, T. Brants, P. Koehn, T. Robinson, One billion word benchmark for measuring progress in statistical language modeling, 2014. URL: https://arxiv.org/abs/1312.3005. arXiv:1312.3005.

[7] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, 2023. URL: https://arxiv.org/abs/1910.10683. arXiv:1910.10683.

[8] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.

[9] L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, S. Presser, C. Leahy, The pile: An 800gb dataset of diverse text for language modeling, 2020. URL: https://arxiv.org/abs/2101.00027. arXiv:2101.00027.

[10] S. Longpre, G. Yauney, E. Reif, K. Lee, A. Roberts, B. Zoph, D. Zhou, J. Wei, K. Robinson, D. Mimno, D. Ippolito, A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity, 2023. URL: https://arxiv.org/abs/2305.13169. arXiv:2305.13169.

[11] R. Eldan, Y. Li, Tinystories: How small can language models be and still speak coherent english?, 2023. URL: https://arxiv.org/abs/2305.07759. arXiv:2305.07759.

[12] I. Beltagy, K. Lo, A. Cohan, Scibert: A pretrained language model for scientific text, 2019. URL: https://arxiv.org/abs/1903.10676. arXiv:1903.10676.

[13] R. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, Y. Li, H. Lee, H. S. Zheng, A. Ghafouri, M. Menegali, Y. Huang, M. Krikun, D. Lepikhin, J. Qin, D. Chen, Y. Xu, Z. Chen, A. Roberts, M. Bosma, V. Zhao, Y. Zhou, C.-C. Chang, I. Krivokon, W. Rusch, M. Pickett, P. Srinivasan, L. Man, K. Meier-Hellstern, M. R. Morris, T. Doshi, R. Delos Santos, T. Duke, J. Soraker, B. Zevenbergen, V. Prabhakaran, M. Diaz, B. Hutchinson, K. Olson, A. Molina, E. Hoffman-John, J. Lee, L. Aroyo, R. Rajakumar, A. Butryna, M. Lamm, V. Kuzmina, J. Fenton, A. Cohen, R. Bernstein, R. Kurzweil, B. Aguera-Arcas, C. Cui, M. Croak, E. Chi, Q. Le, Lamda: Language models for dialog applications, 2022. URL: https://arxiv.org/abs/2201.08239. arXiv:2201.08239.

[14] A. Cooke, A Guide to Finding Quality Information on the Internet: Selection and Evaluation Strategies, 2nd ed., Library Association Publishing, 1999.

[15] Oxford English Dictionary, Quality, https://www.oed.com/dictionary/quality_n, n.d. Retrieved April 15, 2025.

[16] British Standards Institution, Quality management and quality assurance: vocabulary, British Standards Institution, London, 1995.

[17] A. Wieczorkowska, Methodology for obtaining high-quality speech corpora, Applied Sciences 15 (2025) 1848. URL: https://doi.org/10.3390/app15041848. doi:10.3390/app15041848.

[18] M. Tate, M. Tate, J. Alexander, Web Wisdom: How To Evaluate and Create Information Quality on the Web (1st ed.), CRC Press, 1999. URL: https://doi.org/10.1201/9780429195556. doi:10.1201/9780429195556.

[19] L. Codina, Parámetros e indicadores de calidad para la evaluación de recursos digitales, in: Actas de las VII Jornadas Españolas de Documentación. La gestión del conocimiento: retos y soluciones de los profesionales de la información, Bilbao, España, 2000, pp. 135–144.

[20] M. Jiménez Piano, V. Ortiz-Repiso Jiménez, Evaluación y calidad de sedes web, Ediciones Trea, S.L., Gijón, 2007.

[21] L. Gordon-Murname, Evaluating net evaluators, Searcher 7 (1999) 57–66.

[22] P. F. Anderson, N. Allee, S. Grove, S. Hill, Development of a web evaluation tool in a clinical environment, http://www-personal.umich.edu/~pfa/pro/courses/WebEvalNew.pdf, 1999. Accessed: 2025-04-9.

[23] G. Gaffney, Website evaluation checklist v1.1, http://www.infodesign.com.au/ftp/WebCheck.pdf, 1998. Accessed: 2025-04-12.

[24] H. Adreon, A. Catey, K. Strysick, An educator's guide to credibility and web evaluation, http://www.ed.uiuc.edu/wp/credibility-2002/index.html, 2002. Accessed: 2025-4-16.

[25] Pearl K. Wise Library, Web evaluation form, http://www.cpsd.us/CRLS/Library/PDFs/

WebEvaluationForm.pdf, 2006. Accessed: 2025-4-16.
[26] M. Buendía Castro, J. M. Ureña Gómez-Moreno, ¿cómo diseñar un corpus de calidad? parámetros de evaluación, Sendebar: Revista de la Facultad de Traducción e Interpretación (2010) 165–180.
[27] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, in: Advances in Neural Information Processing Systems (NeurIPS), volume 33, 2020. URL: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967411061d95457d756-Abstract.html.
[28] Q. Team, Qwen2.5: A party of foundation models!, http://qwenlm.github.io/blog/qwen2.5/, 2024. Blog; accessed 2025.
[29] J. Dodge, M. Sap, A. Marasović, W. Agnew, G. Ilharco, D. Groeneveld, M. Mitchell, M. Gardner, Documenting large webtext corpora: A case study on the colossal clean crawled corpus, arXiv preprint arXiv:2104.08758 (2021).
[30] Together Computer, Redpajama: An open source recipe to reproduce llama training dataset, https://github.com/togethercomputer/RedPajama-Data, 2023. Accessed 2025.
[31] D. Soboleva, F. Al-Khateeb, R. Myers, J. R. Steeves, J. Hestness, N. Dey, SlimPajama: A 627B token cleaned and deduplicated version of RedPajama, https://cerebras.ai/blog/slimpajama-a-627b-token-cleaned-and-deduplicated-version-of-redpajama, 2023. URL: https://huggingface.co/datasets/cerebras/SlimPajama-627B.
[32] J. Li, A. Fang, G. Smyrnis, M. Ivgi, M. Jordan, S. Gadre, H. Bansal, E. Guha, S. Keh, K. Arora, S. Garg, R. Xin, N. Muennighoff, R. Heckel, J. Mercat, M. Chen, S. Gururangan, M. Wortsman, A. Albalak, Y. Bitton, M. Nezhurina, A. Abbas, C.-Y. Hsieh, D. Ghosh, J. Gardner, M. Kilian, H. Zhang, R. Shao, S. Pratt, S. Sanyal, G. Ilharco, G. Daras, K. Marathe, A. Gokaslan, J. Zhang, K. Chandu, T. Nguyen, I. Vasiljevic, S. Kakade, S. Song, S. Sanghavi, F. Faghri, S. Oh, L. Zettlemoyer, K. Lo, A. El-Nouby, H. Pouransari, A. Toshev, S. Wang, D. Groeneveld, L. Soldaini, P. W. Koh, J. Jitsev, T. Kollar, A. G. Dimakis, Y. Carmon, A. Dave, L. Schmidt, V. Shankar, Datacomp-lm: In search of the next generation of training sets for language models, 2024.
[33] G. Penedo, H. Kydlicek, L. Ben Allal, A. Lozhkov, M. Mitchell, C. Raffel, L. Von Werra, T. Wolf, The fineweb datasets: Decanting the web for the finest text data at scale, in: The Thirty-eighth Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2024. URL: https://openreview.net/forum?id=n6SCkn2QaG.
[34] G. Penedo, H. Kydlıcek, V. Sabolcec, B. Messmer, N. Foroutan, M. Jaggi, L. von Werra, T. Wolf, Fineweb2: A sparkling update with 1000s of languages, december 2024b, URL https://huggingface.co/datasets/HuggingFaceFW/fineweb-2 (????).
[35] C. Zhou, P. Liu, P. Xu, S. Iyer, J. Sun, Y. Mao, X. Ma, A. Efrat, P. Yu, L. Yu, S. Zhang, G. Ghosh, M. Lewis, L. Zettlemoyer, O. Levy, Lima: Less is more for alignment, 2023. URL: https://arxiv.org/abs/2305.11206. arXiv:2305.11206.
[36] T. Le Scao, T. Wang, D. Hesslow, L. Saulnier, S. Bekman, M. S. Bari, S. Bideman, H. Elsahar, N. Muennighoff, J. Phang, et al., What language model to train if you have one million gpu hours?, arXiv preprint arXiv:2210.15424 (2022). URL: https://arxiv.org/abs/2210.15424.
[37] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019). URL: https://arxiv.org/abs/1907.11692.
[38] J. W. Creswell, V. L. P. Clark, Designing and Conducting Mixed Methods Research, 3rd ed., Sage Publications, Thousand Oaks, CA, 2018.

Appendix: Definitions of Evaluation Criteria

Purpose: Refers to its goals and communicative intent, including the intended audience, thematic scope, and any declared limitations in coverage; this helps assess whether the resource meets the user's information needs [14].

Coverage: Refers to the subject areas and types of materials included, as well as the breadth (variety of topics), depth (level of detail), and any stated limitations that define the resource's scope [14][18].

Authority and Reputation: Evaluates the identity and credibility of the authors or institutions responsible for the resource, considering their expertise, professional affiliation, prior publications, and recognised standing in the relevant field [14][18][19][20].

Accuracy: Refers to how factually correct and verifiable the information in a resource is, including the presence of reliable sources, data, and objective evidence to support claims [14][18].

Currency and Maintenance: Assesses whether the resource is regularly updated, includes recent information, removes outdated content, and offers clear mechanisms for indicating revisions or accessing previous versions [14][18].

Accessibility: Encompasses the ease with which a user can access and use the resource, considering technical factors such as URL stability, device compatibility, open or restricted access, and the absence of unnecessary usage barriers [14].

Presentation/Layout Design: Refers to the visual and structural design of the resource, including the use of colours, images, the layout of text, and navigation tools, and the extent to which these elements support understanding and improve the delivery of information [14][19][20].

Luminosity: Refers to the number of external links a website contains to other web pages. It measures how many outgoing references the site provides, contributing to its integration within the broader web ecosystem [19].

Visibility: Refers to the number of other websites that link to the site being analysed. Also known as "popularity", it is a factor used by some search engines to estimate the relevance of a website [19].

Ease of Use/Usability: Refers to the overall user experience while navigating the resource, including structural clarity, availability of help tools, intuitive browsing, and design choices that reduce cognitive effort-closely related to accessibility and design [14].

Ergonomics: Refers to the ease of reading and using a website, considering factors such as the appropriate contrast between text and background that facilitate the readability of the information [19].

Objectivity: Assesses whether the information is presented in a neutral and balanced way, especially on controversial topics, and whether the author's position is clearly stated without manipulating facts [18].

Content: Refers to the core information provided by the resource, evaluated based on its validity, accuracy, completeness, originality, intellectual organisation, timeliness, and relevance to the intended audience [20].

Resource Management: Includes the planning, task assignment, maintenance protocols, quality control, and organisational policies that support the long-term stability and continuous improvement of the web resource [20].