Fine PT-PT Web: A High-Quality 41 Billion Tokens Data Collection of the European Portuguese Web

Gonçalo Vinagre[1,2], Rui Pedro Guerra[1,3], Pedro Gomes[3], Miguel Moura Ramos[4,5], Duarte Miguel Alves[4,5], Afonso Simplício[1,2], Diogo Tavares[1,2], David Semedo[1,2], Daniel Gomes[3], João Magalhães[1,2]

[1]NOVA School of Science and Technology, [2]NOVA LINCS, [3]Fundação para a Ciência e Tecnologia, [4]Instituto Superior Técnico, Universidade de Lisboa, [5]Instituto de Telecomunicações,

Correspondence: gv.martins@fct.unl.pt

arXiv:2609.07699v1 [cs.CL] 7 Sep 2026

Abstract

Curating Web corpora for regional language variants like European Portuguese (PT-PT) is heavily bottlenecked by dialectal overlap (mainly with PT-BR) and data processing scale. This paper presents an efficient pipeline to curate a production-ready PT-PT corpus from the Portuguese Web, spanning 411 TB of raw data from Arquivo.pt. We introduce a novel post-scraping block that removes boilerplate and line duplicates prior to filtering. This early-stage intervention increases final document yield by 19.04% by rescuing valid text that standard heuristic filters prematurely discard. Integrated with rigorous language identification, weighted fuzzy deduplication, and neural quality classification, our pipeline offers a scalable framework and a clean, representative corpus optimized for LLM pre-training.[1]

1 Introduction

The success of modern Large Language Models (LLMs) relies heavily on the availability of massive, diverse, and high-quality text corpora. While general web scrapes like Common Crawl provide an abundant source of raw text, turning these uncurated dumps into high-quality training data requires sophisticated, multi-stage processing pipelines. For dominant, high-resource languages like English, extensive investments have yielded highly optimized, open-source recipes. However, developing comparable resources for mid-to-lower-resource languages, and specifically localized regional variants, remains a major hurdle for the natural language processing community. This challenge is particularly visible in the context of European Portuguese (PT-PT): while Portuguese is a global pluricentric language spoken globally by hundreds of millions of people, it is divided primarily into European Portuguese and Brazilian Portuguese, alongside distinct African and Asian varieties. However, when it comes to data representativeness, it is overwhelmingly dominated by the Brazilian Portuguese (PT-BR) variety.

Standard Web-scraping and language identification filters often fail to adequately distinguish between PT-PT and PT-BR variants, and aggressively discard valid PT-PT content due to rigid heuristic boundaries. In this work, we present a highly efficient, end-to-end data processing pipeline built on top of the open-source Datatrove (Penedo et al., 2024b) library to build a high-quality, production-ready PT-PT Web corpus. Our pipeline processed several web collections from Arquivo.pt, the Portuguese Web Archive,[2] representing approximately 411 terabytes (TB) of raw WARC (Web ARChive) data spanning from 1997 to 2024.

In our pipeline, depicted in Figure 1, we design and implement a custom post-scraping mechanism that cleans and structures raw scraped text before passing it to downstream heuristic quality filters. This component removes short, fragmented lines and local duplicate boilerplate text within documents. Our development experiments demonstrate that post-scraping substantially improves the effective quality of Web documents, rescuing valuable, linguistically rich data that standard filters would normally reject. This intervention alone yields a 19.04% increase in the final number of retained high-quality documents without compromising the rigorous standards required for LLM training. The resulting pipeline combines geographical domain filtering, neural language identification, heuristic text filters adapted from FineWeb2 (Penedo et al., 2025), custom weighted MinHash fuzzy deduplication, neural quality classification via EuroFilter (Martins et al., 2025), and cross-collection deduplication. The result is a robust, clean, and repre-

[1]For the pipeline code, refer to https://github.com/AMALIA-LLM/arquivo_processing and https://github.com/AMALIA-LLM/datatrove-amalia

[2]https://arquivo.pt/collections/

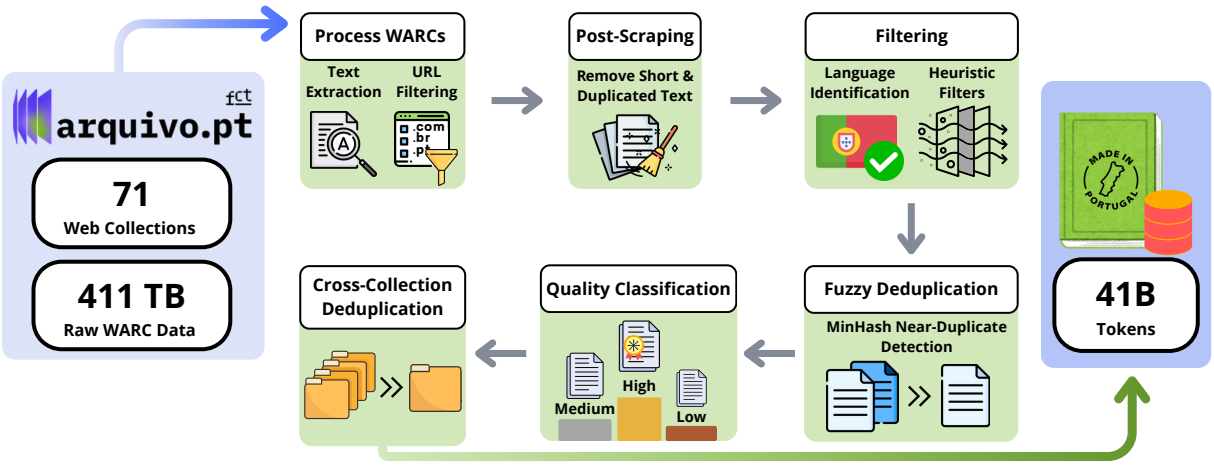

Figure 1: Overview of the developed data processing pipeline. Starting from 71 collections totaling 411 TB of raw WARC data, the pipeline applies six sequential curation stages, yielding a final corpus of 41 billion tokens of high quality European Portuguese text.

sentative PT-PT corpus specifically optimized to power the mid-training and pre-training phases of next-generation language models.

2 Related Work

Web Data Processing Pipeline. The curation of large-scale Web corpora for LLM training has become a well-established research area, with pipelines typically built around CommonCrawl,[3] combining filtering and deduplication steps (Penedo et al., 2023). FineWeb (Penedo et al., 2024a) represents one of the major contributions to the community, releasing a 18.5-trillion-token dataset alongside Datatrove (Penedo et al., 2024b), a modular data processing library, while also applying and adapting quality and repetition filters from MassiveText (Rae et al., 2022) and C4 (Raffel et al., 2020).

Deduplication is equally important, as training on deduplicated data consistently improves model quality and reduces memorization (Lee et al., 2022). Deduplication can be done at the URL level, through exact or semantic matching (Abbas et al., 2023), or through MinHash-based near-duplicate detection (Broder, 1997), which has become the standard approach at scale. Some approaches apply heuristics to lower-quality documents only to avoid discarding high-quality content (Su et al., 2025), while others complementary apply model-based filters (Henriksson et al., 2025).

[3] https://commoncrawl.org

Multilingual Web Corpora. Several efforts have been made to build high-quality multilingual Web datasets (Xue et al., 2021; Abadji et al., 2022; Laurençon et al., 2022; Nguyen et al., 2024; Ali et al., 2025), with recent work showing that language-specific pipelines are crucial to guarantee the best quality (DatologyAI et al., 2026; Burns et al., 2026; Norlund et al., 2025; Penedo et al., 2025). Despite these advancements, multilingual datasets often fail to represent each language thoroughly, blurring regional variants and language-specific characteristics, a limitation that motivates language- and variant-specific corpus efforts.

European Portuguese Language Resources While most multilingual corpora include Portuguese data, they treat it as a single variant despite significant lexical, grammatical, and orthographic differences between European and Brazilian Portuguese (Preda et al., 2024), with the latter being systematically over-represented due to its much larger Web presence (Almeida et al., 2025). GlórIA (Lopes et al., 2024) introduced a European Portuguese decoder-based LLM pre-trained on a large PT-PT corpus. This corpus includes an Arquivo.pt subset derived from approximately 7 TB of raw WARC data, filtered using basic heuristic cleaning and exact-duplicate removal, resulting in 1.5 million documents (0.8 billion tokens) of '.pt' news websites only. More recently, AMALIA (Simplício et al., 2026), the first fully open LLM for European Portuguese, was trained on a preliminary version of the corpus introduced in this work, which was made available to the AMALIA team

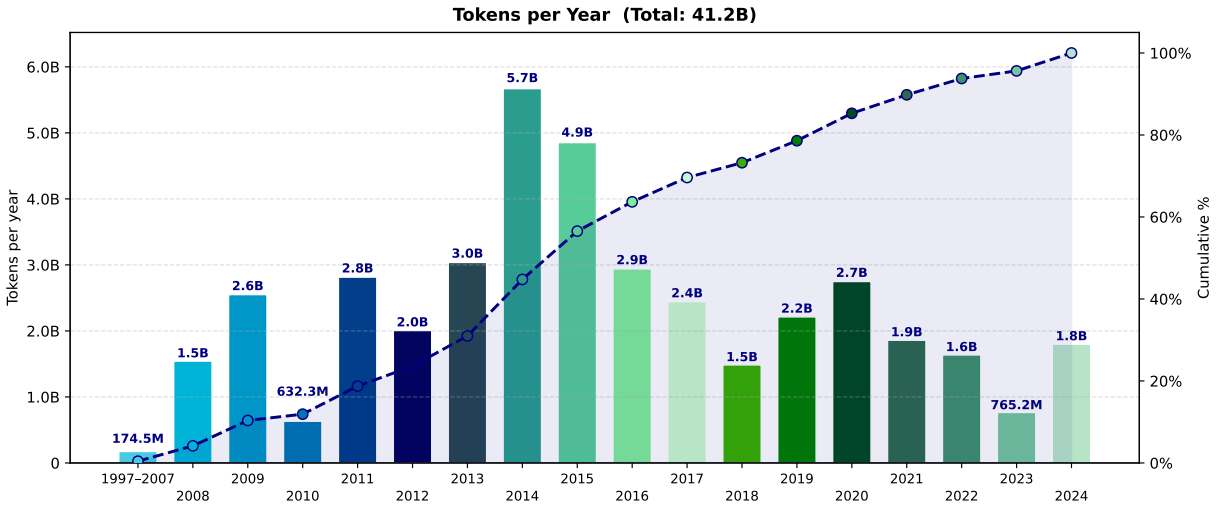

Figure 2: Token distribution across crawl years. The corpus spans 1997–2024, with peak coverage in 2014–2015. The dashed line shows the cumulative proportion of total tokens.

prior to publication.

Recent benchmark developments (Tavares et al., 2026; Ferreira et al., 2026; Vieira et al., 2026; Baucells et al., 2025) further reflect the growing PT-PT research ecosystem. Despite this progress, no large-scale, openly documented pipeline exists for processing Web archive data targeting European Portuguese specifically, a gap that this work addresses.

3 Corpus Creation Methodology

3.1 Data Collection

Our corpus is composed of publicly available data from Arquivo.pt, the Portuguese Web Archive, which has been continuously crawling and preserving Portuguese Web content since 1996, while respecting the Robots Exclusion Protocol.[4] We deployed a dedicated server within Arquivo.pt's infrastructure, enabling direct transfer and local processing of raw WARC archives, achieving a roughly 20× reduction in processing time relative to querying their public API,[5] due to bandwidth limitations and throttling.

Combining the data collected across the phases of this work, we processed 71 collections representing roughly 411 TB of raw WARC archives available in Arquivo.pt's database. These collections span a wide variety of content types, from news to blogs and the general Web, and were crawled between 1997 and 2024, ensuring that we respect an embargo of one year,[6] as seen in Figure 2 (collection selection details are in Appendix A).

The collections we processed were part of three main categories: frequently crawled content (FAWP), general Web content (AWP), and other smaller collections with high-quality content. Depending on the type of collection, the crawling frequency ranges from daily to quarterly, with some collections dedicated to special events and/or selected websites. Among the available collections, we have chosen a total of 71 collections, according to their time span, spacing in time, and crawling frequency, towards minimizing duplicates while still guaranteeing content comprehensiveness over the full period. Table 1 summarizes the proportions and content of each collection type processed.

3.2 Data Processing

We process the raw Web archive data from Arquivo.pt, using the WARC archives collected, into a high-quality European Portuguese corpus through a multi-stage pipeline built on top of Datatrove, an open-source library for large-scale text processing. Our pipeline takes inspiration from FineWeb2 (Penedo et al., 2025), adopting its threshold configuration for Portuguese,[7] but extending its approach with custom processing components and implementation improvements. Our overall pipeline consists of six stages that each WARC collection is submitted to: (1) WARC Processing,

[4] https://sobre.arquivo.pt/en/help/crawling-and-archiving-web-content/#qe-faq-2407

[5] https://arquivo.pt/api

[6] Arquivo.pt follows a one year embargo policy, to avoid drawing traffic away from original content publishers.

[7] https://github.com/huggingface/fineweb-2/blob/main/configs/por_Latn.yml

Name	Type	Update Rate	#Tokens	#URL	Description
AWP	General Web	Quarterly	36 827M	20 667k	Complete crawls of the Portuguese Web
MAWP	Media	Monthly	85M	84k	Monthly crawls of selected websites
FAWP	Media	Daily	3 702M	3 438k	Websites which publish new daily content
EAWP	Special Crawls	Varied	43M	31k	Selected pages about a given theme or event
RAQ	High-Quality	Varied	175M	140k	Selected websites carefully curated
Other	Donations	Varied	340M	199k	Collections archived/donated by users or institutions

Table 1: Summary of Arquivo.pt collection types processed in this work. AWP dominates the corpus by volume (36.8B tokens, 89% of the total), while smaller curated collections such as RAQ and EAWP contribute higher-quality but sparser content.

(2) Text Post-Scraping, (3) Data Filtering, (4) MinHash Deduplication, (5) Quality Classification, and (6) Cross-Collection Deduplication. An overview is shown in Figure 1.

3.2.1 WARC Processing

The first stage processes the raw WARC archives, using Datatrove's built-in *WarcReader*, which are submitted to two URL-level filters before any text extraction. First, we remove every document with URLs belonging to '.br' domains or containing similar variants (such as '/pt-br'), in order to reduce the presence of Brazilian Portuguese text and focus on European Portuguese. Second, we remove URLs matching a blacklist of known NSFW domains and banned keywords. We deliberately do not restrict to '.pt' domains alone, as this would discard a large proportion of valid European Portuguese content hosted under other top level domains.

We proceed with the text extraction from the HTML content in the WARC files using a custom Trafilatura-based (Barbaresi, 2021) block that extends Datatrove's built-in implementation to additionally preserve the page title as metadata alongside the extracted text. We find that this extra metadata provides richer document context for downstream use without decreasing extraction efficiency.

3.2.2 Text Post-Scraping

One of the main contributions of this work is a custom *post-scraping* block, which improves the effective quality of the documents. This improvement prevents some documents from being removed by the downstream heuristic filters, without worsening the final quality of the dataset. This represents an increase of 19.04% in the final number of documents obtained from our test collection (more details in Section 4.1). We remove short lines, i.e., with fewer than 10 alphabetic characters, and duplicate lines within a document, retaining only a single instance of each repeated line.

3.2.3 Filtering

The extracted and processed text is then submitted to quality filters to discard low-quality documents. First, each document is passed through a *Language Filter* using the GlotLID model (Kargaran et al., 2023). Documents where the estimated probability of European Portuguese falls below a threshold of 0.799 (consistent with the FineWeb2 configuration for Portuguese) are discarded.

For the heuristic quality filters, we start with the *Gopher Repetition Filter* (Rae et al., 2022) to remove documents with excessive repetition, removing documents with a ratio of duplicated n-grams following two different criteria: for short n-grams (2–4), we apply a top n-gram criterion that flags documents where the most frequent repeated n-gram accounts for a disproportionate share of characters; for longer n-grams (5–10), we apply a duplicate n-gram criterion over all repeated n-gram occurrences. Next, we apply the *FineWeb Quality Filter*, removing documents with excessive lines ending without appropriate punctuation and with an excessive ratio of newlines to words. Finally, the *Gopher Quality Filter* (Rae et al., 2022) is applied, where documents are removed if they fall outside acceptable word count bounds, have an average word length outside the defined range, exceed limits on the proportion of hash symbols, ellipses, bullet-prefixed lines, or ellipsis-terminated lines, fall below a minimum alphabetic character ratio, or contain fewer than two Portuguese stop words from a fixed list. An example of a sample removed by each heuristic filter is shown in Appendix D. Following the heuristic filters, we apply a formatting step to correct possible character encoding issues, redact personal data (public IP addresses, email addresses, and phone numbers), and remove residual HTML table artifacts introduced by the scraper.

3.2.4 Deduplication and Rehydration

For each collection, we independently perform fuzzy deduplication using MinHash (Broder, 1997), implemented via Datatrove's four-stage deduplication pipeline: (1) computing a MinHash signature for each document, (2) identifying pairs of duplicate signatures and outputting them as buckets, (3) creating clusters of duplicates from the buckets, and (4) filtering the original documents by retaining only one representative document per cluster.

After deduplication, a *Rehydrater* block assigns an upsampling weight to each retained document based on its cluster size. Documents from larger clusters receive proportionally higher weights, reflecting the implicit density of the content they represent, whereas documents from very large clusters (indicative of spam or boilerplate) have their weights capped. These weights are stored in the document metadata to support optional data rehydration during training. Besides these weights, our custom Datatrove implementation also lets users keep all duplicate documents and only annotate their clusters' ids and sizes, to allow for more cluster filtering techniques besides random sampling, as Datatrove defaults to.

3.2.5 Quality Classification

Following per-collection deduplication, all collections are jointly processed by the EuroFilter quality classification model (Martins et al., 2025). With the scores given by this classifier, we discarded the low-quality entries and proceeded to the following pipeline stage only with the remaining ones. Unlike the other stages, this step is run outside of Datatrove.

3.2.6 Cross-Collection Deduplication

The per-collection deduplication in Section 3.2.4 does not remove duplicates that appear across collections. Since different Arquivo.pt collections often crawl the same Web domains at different points in time, they may overlap substantially in content. Thus, a second fuzzy deduplication pass, using an identical procedure, is performed across all collections jointly, applied to the output of the quality classification stage. This step yields the final corpus used for language model training.

3.3 Hardware and Processing Configuration

As mentioned in Section 3.1, we deployed a dedicated server in Arquivo.pt's infrastructure, achieving a 20× reduction in total processing time compared to querying their public API. This dedicated server has two AMD CPUs with 32 cores each, one JBOD storage array with 1.2 PB, and a 20 Gbps network connection. WARC processing was performed on this dedicated server and took three weeks, while filtering and deduplication were performed on the MareNostrum 5 supercomputer hosted at the Barcelona Supercomputing Center, using seven compute nodes, and together took one and a half weeks to complete. All three stages used Datatrove, submitting jobs as SLURM job arrays with its predefined parameters: WARC processing ran with 10,000 tasks across 32 workers, using 2 CPUs per task and 4GB of memory per CPU; data filtering ran with 280 tasks across 128 workers, using 6 CPUs per task and 2GB of memory per CPU; deduplication ran with 100 tasks across 100 workers, using 8 CPUs per task and 2GB of memory per CPU.

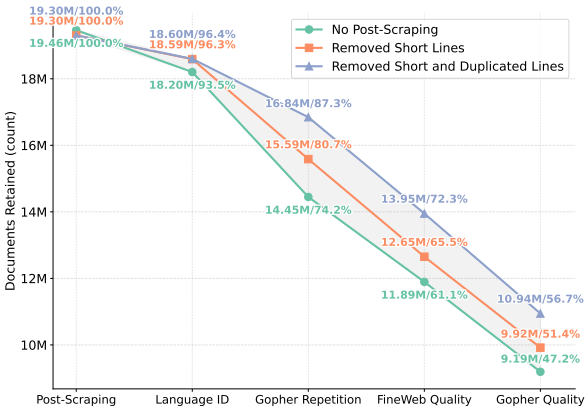

Figure 3: Comparison of document retention through the filtering stages between not including any of the post-scraping operations, including only the short line removal operation, and including the short line and duplicated lines removal operations.

4 Data Curation Results

We validate the design decisions in our pipeline through experiments on a single collection (FAWP21), composed primarily of Portuguese news media, frequently updated websites, and government and public body pages, crawled between April and June 2015.

4.1 Post-Scraping Steps

We remove lines with fewer than 10 characters and duplicate lines within each document before heuristic filtering. To understand the contribution of each added step, we compare three configurations: no post-scraping (baseline), short line removal only,

and both steps combined.

Performing both steps consistently increases the number of documents remaining after the full filtering pipeline. Starting from a baseline of 9.19 million retained documents, adding short line removal yields 9.92 million documents (+7.94%), while further adding duplicate line removal yields 10.94 million documents (+19.04% over baseline). This yield improvement does not come at the expense of quality: post-scraping improves the quality of documents which would otherwise be rejected. Every document, regardless of post-scraping, still passes through the same downstream filtering pipeline. Each step rescues more content which would have been discarded by the heuristic filters in its raw scraped form, as shown in Figure 3. In our test collection, only 47.2% of documents survive to the final Gopher Quality stage without post-scraping, compared to 56.7% with both steps applied. Examples of rescued documents before and after being subject to post-scraping can be seen in Appendix C. We also experimented removing residual HTML tags from the extracted text, but found that many were part of legitimate technical content such as HTML tutorials, where their presence is meaningful. Consequently, we utilize the text scraper's native tag cleanup feature

4.2 Model-Based Quality Pre-Filtering

Motivated by Nemotron-CC (Su et al., 2025), we evaluated whether applying quality classification before heuristic filtering, exempting high-quality documents, could improve our pipeline. We thus applied the EuroFilter quality classifier (Martins et al., 2025) to our test collection prior to filtering. The results showed negligible benefit, as this pipeline retained 0.4% more documents than heuristic filtering alone. Given that this step significantly increases the processing time for a single collection, the marginal yield improvement does not justify the added computational cost. As a result we used the simpler heuristic-only filtering approach.

4.3 Final Corpus Statistics

This pipeline resulted in a high-quality European Portuguese Web corpus, enabling pre-training or mid-training of PT-PT-focused language models. Table 2 summarizes the filtering efficiency at each stage of the processing pipeline over all 71 collections. Starting from approximately 5.4 billion raw entries in the WARC archives, URL filtering reduces this to 3.4 billion entries by removing Brazilian Portuguese and NSFW domains. Text extraction further reduces the count to 1.7 billion entries, after which heuristic filtering removes the majority of low-quality documents, leaving 731 million entries. Single-collection MinHash deduplication then produces a dramatic reduction to 148 million entries, an ~80% reduction from the previous step and ~97% from the raw input, reflecting the high degree of content repetition inherent in a Web archive crawled over multiple years, particularly in news-heavy collections where pages are crawled daily.

Stage	Documents
Raw WARC data (§3.1)	5 392 417 365
After URL filtering (§3.2.1)	3 440 781 193
After text extraction (§3.2.1)	1 653 754 362
After heuristics (§3.2.3)	731 048 214
After deduplication (§3.2.4)	147 699 937

Table 2: Document counts at each stage of the processing pipeline across all 71 Arquivo.pt collections before cross-collection deduplication and quality classification

	GlórIA	Ours
#Documents	1.5M	24.6M
#Tokens	0.8B	41B
Scope	News only	All domains

Table 3: Comparison between the Arquivo.pt data used in GlórIA vs our approach.

After discarding low-quality content through quality classification, the remaining documents are subject to a cross-collection deduplication stage. This results in 24.6M documents and 41B tokens (calculated using the GPT-2 tokenizer (Radford et al., 2019), as are all token counts in this section). These processing steps reduce the corpus token count from 176B tokens (after per-collection deduplication) to 41B tokens, a 76.6% reduction, reflecting substantial content overlap across collections and highlighting the importance of this processing step. When comparing with the Arquivo.pt data included in GlórIA's training corpus, this represents a 16.4× increase in document count and a 51.3× increase in token count, large enough for Arquivo.pt alone to serve as a primary pretraining source rather than a minor supplementary one, as it was in GlórIA's corpus, as summarized in Table 3.

The final corpus covers a wide variety of content areas, with news and blogs dominating; Fig-

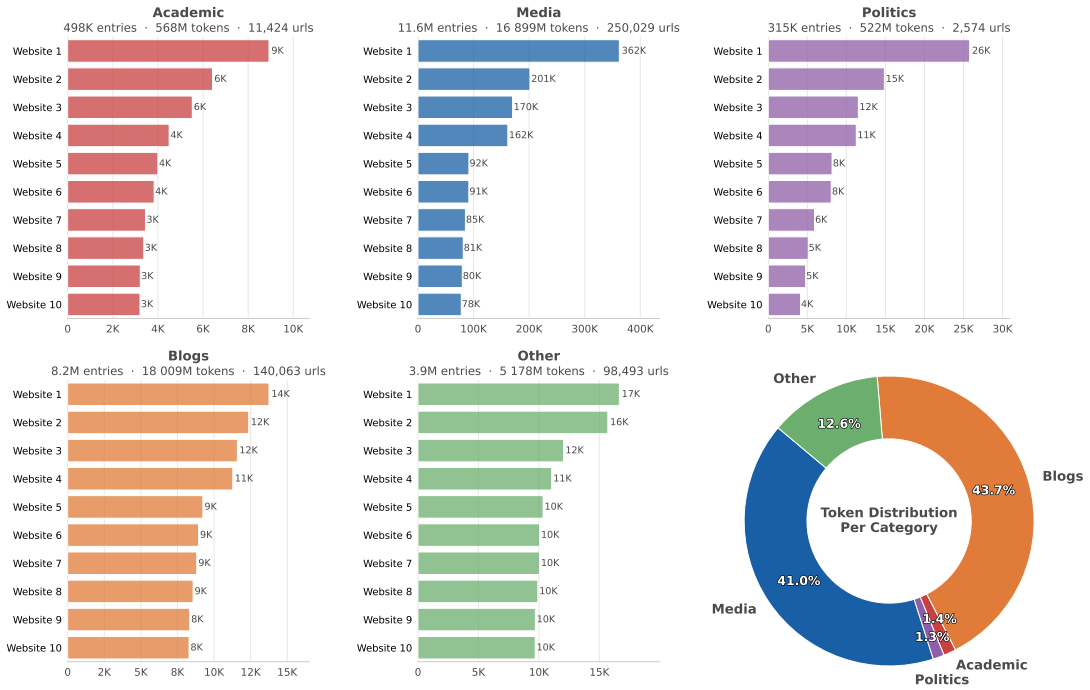

Figure 4: Top 10 domains by entry count across content categories, with overall token distribution by category (bottom right). The corpus is dominated by media and blogs (41% and 43.7% of tokens, respectively), reflecting the strong presence of professional Portuguese-language publishing in Arquivo.pt.

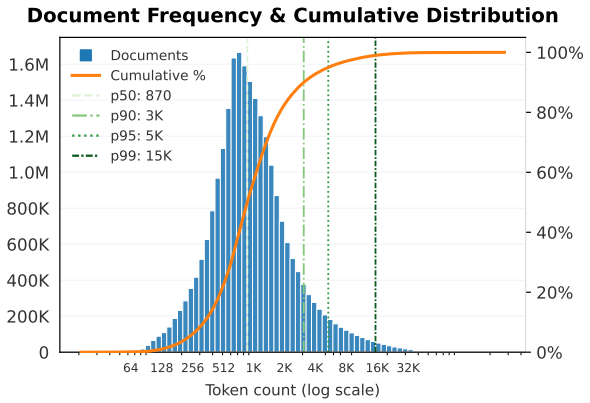

Figure 5: Document token-length distribution of the processed Arquivo.pt collections.

ure 4 shows the 10 most common domains per category and the category distribution of the corpus. Figure 5 shows the cumulative token length distribution of the final corpus, with over 90% of documents containing fewer than 4k tokens.

5 Experiments and Results

5.1 Filter Stage Quality Ablations

To empirically validate the effect of each filtering stage on document quality, we analyze the next-token prediction behavior of a set of language models on documents removed at each stage of the heuristic filtering pipeline. For each filter, we sample 10,000 documents from the removed set, selecting the worst-case examples according to the filter metric that caused their removal (e.g., the documents with the highest duplicate n-gram fraction for Gopher Repetition), while preserving the original proportion of filter reasons. We additionally sample 10,000 documents uniformly at random from the pipeline output after filtering, which serves as our quality baseline. These sets are deduplicated before sampling to ensure diversity. For each document, we compute bits-per-byte (bpb) and top-1 next-token prediction accuracy using four base language models of comparable size: AMALIA 9B (Simplício et al., 2026), a PT-PT-focused model mid-trained in our collection; EuroLLM 9B (Martins et al., 2025), the base model on top of which AMALIA was trained; Llama 3.1 8B (Dubey et al., 2024), an established multilingual model; and OLMo 2 7B (OLMo et al., 2024), an English-only model included as a control.

The results, shown in Figure 6, reveal a consistent pattern across all four models. Documents removed by the FineWeb Quality and Gopher Quality filters are systematically harder to predict than the final output, with higher bits-per-byte across all models, confirming that these filters remove genuinely low-quality text that deviates from the nat-

Model	Consecutive prefix words						BLEU-3						BLEU-4					
	mean	p75	p90	p95	p99	max	mean	p75	p90	p95	p99	max	mean	p75	p90	p95	p99	max
AMALIA-9B	0.696	1.0	2.0	3.0	6.0	26.0	0.051	0.068	0.104	0.138	0.219	0.391	0.031	0.038	0.067	0.098	0.169	0.326
EuroLLM-9B	0.704	1.0	2.0	3.0	7.0	26.0	0.048	0.062	0.098	0.134	0.203	0.474	0.028	0.034	0.062	0.092	0.15	0.462
Llama-3.1-8B	0.622	1.0	2.0	3.0	5.0	18.0	0.045	0.058	0.094	0.124	0.197	0.256	0.026	0.031	0.056	0.082	0.146	0.213
OLMo-2-7B	0.449	1.0	1.0	2.0	4.0	18.0	0.036	0.046	0.072	0.096	0.166	0.307	0.021	0.024	0.042	0.062	0.13	0.276

Table 4: Memorization-eval statistics per model: means, percentiles and maximum.

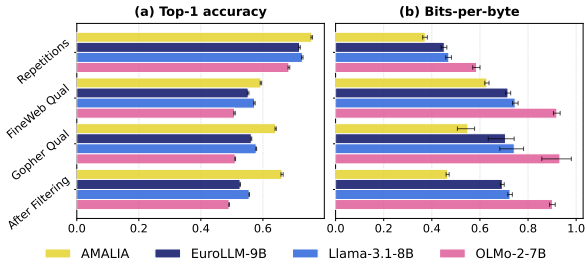

Figure 6: Model comparison on (a) top-1 next-token prediction accuracy and (b) bits-per-byte (lower=better) on the removed samples by each filter stage and the final output.

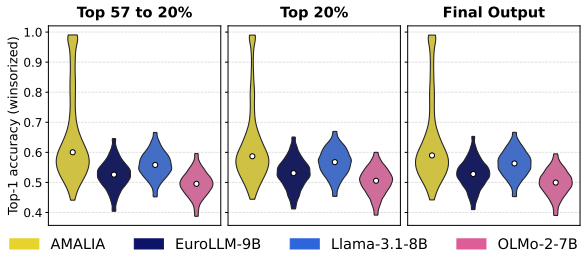

Figure 7: Quality thresholds comparison in the next-token prediction task.

ural language patterns learned during pre-training. However, the documents removed at the Gopher Repetition stage have the lowest bits-per-byte and highest top-1 accuracy. This is expected, since repetitive documents are inherently predictable and a language model that has learned any recurring pattern will assign high probability to its continuations, scoring better than the other stages' samples.

AMALIA shows markedly lower bits-per-byte and higher top-1 accuracy on the output set, reflecting the effect of its mid-training on Arquivo.pt data from this exact collection. However, EuroLLM's bits-per-byte on the output is comparable to or lower than its performance on the Gopher Quality removed set, suggesting that EuroLLM's general multilingual training does not specifically favour high-quality European Portuguese web text over low-quality text from the same domain.

OLMo-2, an English-only model, shows the highest bits-per-byte and lowest top-1 accuracy across all stages, reflecting its lack of exposure to Portuguese during pre-training. Despite this, it mirrors the ordering observed in the multilingual models. This suggests that these patterns reflect intrinsic properties of the text rather than any model's familiarity with Portuguese.

5.1.1 Quality Thresholds Comparison

Applying our quality filter followed by cross-collection deduplication on the medium and high-quality documents, leaves us with 57% of the original documents. To assess the impact of our quality threshold, we experimented running our next-token prediction evaluation at different splits of our final corpus: top 57 to 20%, which we'll call medium-quality; top 20%, which we'll call high-quality; and the final output, consisting of both medium- and high-quality splits. Figure 7 shows that AMALIA achieves slightly higher top-1 accuracy on medium-quality documents than on high-quality ones, with averages of 0.68 and 0.63, respectively, likely reflecting the larger proportion of medium-quality text in its training data. This suggests that increasing the proportion of high-quality documents during training, achievable through the upsampling weights assigned by the *Rehydrater* block, could improve the quality of the final corpus, and, consequently, the trained model's capabilities.

5.2 Memorization Experiments

Language models are known to memorize and regurgitate verbatim sequences from their training data, raising privacy and copyright concerns (Huang et al., 2024) and inflating perplexity-based quality estimates (Lee et al., 2022). It is therefore crucial to understand if a model trained with this data is generalizing or memorizing the data. To evaluate this, we adopt a prefix-continuation protocol: 1) we sample documents from the corpus, 2) truncate each sample to at most 4,096 tokens, 3) split them in half, and 4) prompt the model with the first half. We then greedy-decode a continuation of the same length as the

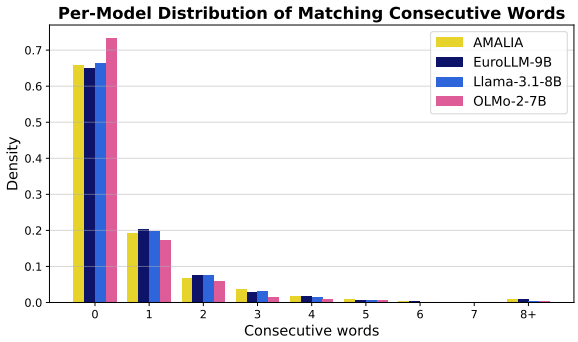

Figure 8: Per-model distribution of the count of leading consecutive words that exactly match the ground-truth continuation (n = 1,000 prompts each).

held-out second half and measure the overlap between the generated and the true continuation. Our primary metric is *consecutive words*: the length of the longest word sequence, starting from the first generated token, that the model reproduces exactly from the true continuation. This directly captures verbatim memorization by measuring how far the model can extend a document word-for-word before diverging. We complement this with *BLEU-3* and *BLEU-4*, which quantify n-gram precision at orders 3 and 4 between the generated and reference continuations. For each metric we report the mean, maximum, and upper percentiles across samples, summarized in Table 4; per-model distributions of *consecutive words* are shown in Figure 8. The results indicate that memorization is minimal: over 99% of samples yield six or fewer consecutive matching words, and BLEU-4 scores remain below 0.169, confirming that the models rarely reproduce verbatim spans from the training data. The few exceptions are still small as illustrated in Table 7. Notably, AMALIA's memorization metrics remain comparable to those of EuroLLM, the base model on top of which it was mid-trained, suggesting that the additional training on this corpus did not lead to meaningful memorization beyond what is already attributable to general language patterns and potential overlap with other pretraining sources.

5.3 Brazilian Portuguese Contamination

Due to the significant overlap between European and Brazilian Portuguese variants, it is difficult to filter out all PT-BR content without losing high-quality PT-PT text. Our URL-based filtering approach does not exclude generic top-level domains such as `.com` and `.net`, which may include content in either Portuguese variant. We therefore estimated PT-BR contamination in the final corpus using two independent, well-established language-variety classifiers. The `duarteocarmo/fasttext-euptvid` classifier (Carmo, 2026) labeled 86.4% of the corpus as PT-PT and 13.6% as PT-BR, while the `bastao/PeroVaz_PT-BR_Classifier`[8] assigned 82.9% to PT-PT and 17.1% to PT-BR. Averaging the estimates from the two classifiers results in 84.65% PT-PT and 15.35% PT-BR, indicating that PT-PT remains the predominant variant in the final corpus and providing evidence for the effectiveness of our filtering steps.

6 Conclusion

This paper presents an efficient pipeline for producing and curating a PT-PT corpus at scale, in order to address the low availability of text resources of the language variant when compared to PT-BR. Using the Web content preserved by Arquivo.pt, the Portuguese Web Archive, and taking advantage of a dedicated server infrastructure, we processed 411 TB of raw WARC archives, spanning from 1997 to 2024. Our pipeline introduced a novel post-scraping block, improving the overall quality of the text, leading to an increase of 19.04% in the document count of the final corpus. In combination with URL, language, and heuristic-based filtering, model-based quality classification, and MinHash fuzzy deduplication steps, we were able to create a **41 billion tokens of high-quality European Portuguese dataset**.

Acknowledgments

This work was supported by the AMALIA project under Measure RE-C05-i08 of the Portuguese national Programa de Recuperação e Resiliência. We also acknowledge the support of Fundação para a Ciência e Tecnologia (FCT) and the NOVA LINCS project (UID/04516/2025). We thank Arquivo.pt for providing the archived data and collaborating on setting up a dedicated server within their infrastructure, which made this work possible. Finally, we thank the Barcelona Supercomputing Center (BSC) for providing the needed computational resources.

[8] https://huggingface.co/bastao/PeroVaz_PT-BR_Classifier

Limitations

The corpus is also topically skewed towards news media and frequently updated websites, which dominate the Arquivo.pt collections we processed. Despite its naturally high journalistic quality and comprehensive coverage of Portugal events, the resulting corpus may still underrepresent certain content types such as academic, technical, or literary text, which are less prevalent in web archive data in general. We counteracted this skewness by including more general collections (AWP) and other smaller curated collections to improve topical diversity.

The heuristic filters applied in this pipeline operate at the document level, discarding entire documents when any quality threshold is exceeded. As a result, documents containing mostly high-quality text alongside small portions of low-quality content are removed in their entirety. This conservative approach was deliberately chosen to minimise the risk of noisy or malformed text influencing LLM training, but it inevitably discards some useful content. A sentence- or paragraph-level filtering approach could recover this content at the cost of additional complexity.

Ethical Considerations

The development of this pipeline and the collection of Arquivo.pt data were carried out in accordance with relevant ethical and legal guidelines. Following GDPR legislation, personal data was removed from the final corpus. Only publicly available information was collected, while following all copyright and licensing restrictions. The Robots Exclusion Protocol (robots.txt) directives of each crawled domain were also respected throughout the collection process.

References

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

Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S. Morcos. 2023. Semdedup: Data-efficient learning at web-scale through semantic deduplication. *Preprint*, arXiv:2303.09540.

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 Koehler, Patrick Schramowski, Michael Fromm, and Kristian Kersting. 2025. Judging quality across languages: A multilingual approach to pretraining data filtering with language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8859–8898, Suzhou, China. Association for Computational Linguistics.

Thales Sales Almeida, Rodrigo Nogueira, and Helio Pedrini. 2025. Building high-quality datasets for portuguese llms: From common crawl snapshots to industrial-grade corpora. *Preprint*, arXiv:2509.08824.

Adrien Barbaresi. 2021. Trafilatura: A web scraping library and command-line tool for text discovery and extraction. In *ACL 2021 - System Demonstrations*, pages 122–131. ACL.

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

A.Z. Broder. 1997. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997*, pages 21–29.

Thomas F Burns, Letitia Parcalabescu, Stephan Waeldchen, Michael Barlow, Gregor Ziegltrum, Volker Stampa, Bastian Harren, and Björn Deiseroth. 2026. Aleph-alpha-GermanWeb: Improving German-language LLM pre-training with model-based data curation and synthetic data generation. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1267–1283, Rabat, Morocco. Association for Computational Linguistics.

Duarte O. Carmo. 2026. fasttext-euptvid: Fast portuguese variety identification.

DatologyAI, :, Aldo Gael Carranza, Kaleigh Mentzer, Ricardo Pio Monti, Alex Fang, Alvin Deng, Amro Abbas, Anshuman Suri, Brett Larsen, Cody Blakeney, Darren Teh, David Schwab, Diego Kiner, Fan Pan, Haakon Mongstad, Haoli Yin, Jack Urbanek, Jason Lee, and 16 others. 2026. überweb: Insights from multilingual curation for a 20-trillion-token dataset. *Preprint*, arXiv:2602.15210.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Rafael Ferreira, Inês Vieira, Inês Calvo, James Furtado, Iago Paulo, Diogo Glória-Silva, Diogo Tavares, David Semedo, and Joao Magalhaes. 2026. P3B3: A multi-turn conversational benchmark for measuring European and Brazilian Portuguese variety bias in LLMs. In *Proceedings of the 1st Workshop on Multilinguality in the Era of Large Language Models (MeLLM 2026)*, pages 240–248, San Diego, United States. Association for Computational Linguistics.

Erik Henriksson, Otto Tarkka, and Filip Ginter. 2025. FinerWeb-10BT: Refining web data with LLM-based line-level filtering. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 258–268, Tallinn, Estonia. University of Tartu Library.

Jing Huang, Diyi Yang, and Christopher Potts. 2024. Demystifying verbatim memorization in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 10711–10732. Association for Computational Linguistics.

Amir Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. Glotlid: Language identification for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, page 6155–6218. Association for Computational Linguistics.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, Jörg Frohberg, Mario Šaško, Quentin Lhoest, Angelina McMillan-Major, Gerard Dupont, Stella Biderman, Anna Rogers, Loubna Ben allal, Francesco De Toni, and 35 others. 2022. The bigscience roots corpus: A 1.6tb composite multilingual dataset. In *Advances in Neural Information Processing Systems*, volume 35, pages 31809–31826. Curran Associates, Inc.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, Dublin, Ireland. Association for Computational Linguistics.

Ricardo Lopes, João Magalhães, and David Semedo. 2024. Glória: A generative and open large language model for portuguese. In *PROPOR 2024, Volume 1*, pages 441–453. ACL.

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.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2024. CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4226–4237, Torino, Italia. ELRA and ICCL.

Tobias Norlund, Tim Isbister, Amaru Cuba Gyllensten, Paul dos Santos, Danila Petrelli, Ariel Ekgren, and Magnus Sahlgren. 2025. Sweb: A large web dataset for the scandinavian languages. In *International Conference on Learning Representations*, volume 2025, pages 82775–82798.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, and 21 others. 2024. 2 olmo 2 furious. *ArXiv*, abs/2501.00656.

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

Guilherme Penedo, Hynek Kydlíček, Alessandro Cappelli, Mario Sasko, and Thomas Wolf. 2024b. Datatrove: large scale data processing.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. Fineweb2: One pipeline to scale them all – adapting pre-training data processing to every language. *Preprint*, arXiv:2506.20920.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only. *Preprint*, arXiv:2306.01116.

David Preda, Tomás Osório, and Henrique Lopes Cardoso. 2024. Across the atlantic: Distinguishing between European and Brazilian Portuguese dialects. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 353–363, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, and 61 others. 2022. Scaling language models: Methods, analysis & insights from training gopher. *Preprint*, arXiv:2112.11446.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Afonso Simplício, Gonçalo Vinagre, Miguel Moura Ramos, Diogo Tavares, Rafael Ferreira, Giuseppe Attanasio, Duarte M. Alves, Inês Calvo, Inês Vieira, Rui Guerra, James Furtado, Beatriz Canaverde, Iago Paulo, Vasco Ramos, Diogo Glória-Silva, Miguel Faria, Marcos Treviso, Daniel Gomes, Pedro Gomes, and 3 others. 2026. AMALIA: A fully open large language model for European Portuguese. In *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1*, pages 380–391, Salvador, Brazil. Association for Computational Linguistics.

Dan Su, Kezhi Kong, Ying Lin, Joseph Jennings, Brandon Norick, Markus Kliegl, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2025. Nemotron-cc: Transforming common crawl into a refined long-horizon pretraining dataset. *Preprint*, arXiv:2412.02595.

Diogo C. Tavares, Rafael Ferreira, Afonso Simplício, Gonçalo Vinagre, Ana Carolina Condez, Inês Calvo, Inês Vieira, David Semedo, and Joao Magalhaes. 2026. Pheb: An european portuguese high school-level llm benchmark. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 4673–4683, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Inês Vieira, Inês Calvo, Iago Paulo, James Furtado, Rafael Ferreira, Diogo Tavares, Diogo Glória-Silva, David Semedo, and João Magalhães. 2026. ALBA: A European Portuguese benchmark for evaluating language and linguistic dimensions in generative LLMs. In *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1*, pages 697–707, Salvador, Brazil. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

A Collection Selection Details

As explained in Section 3.1, Arquivo.pt comprises several collection types, with varying contents, focuses, and crawling frequency. This variety required us to examine each collection type and define a tailored processing procedure for each.

The FAWP collections contain websites with frequently updated content, such as news media, public and governmental pages, which were crawled every day and saved as quarterly collections. Given the highly variable content, this was the first collection type we processed, using almost every quarterly collection between the 2010-2016 and 2020-2024 time periods.

In order to vary the type of content, we also started processing AWP collections, which are complete crawls of the Portuguese Web, mainly from the .pt top level domain. They contain mainly blogs and other general content that was crawled and saved quarterly. Given the larger volume and higher chance of duplicate content, we selected roughly one collection of similar size per year in the 2008-2024 time period.

Lastly, we also selected to process an assortment of smaller collections, including high-quality, manually curated pages, donated pages, pages crawled during special events, and pages archived by Arquivo.pt users with their ArchivePageNow[9] service. These pages, collected during the 1997-2024 time period, were used to further improve the variety and quality of the extracted data.

B Dropped Samples per Stage

As explained in Section 4.1, the documents that are preserved in the early stages are still subject to the subsequent heuristic filters, and many are removed by them, demonstrating the effectiveness of these post-scraping steps. Figure 9 shows this evolution per stage.

C Rescued Samples by the Post-Scraping Stage

Table 5 shows examples of documents rescued by the post-scraping step described in Section 4.1. Each example shows a document in its raw scraped form and the resulting document after subjected to the post-scraping step. This table illustrates the range of cases the post-scraping step addresses, each of which would otherwise cause the document to be discarded by the downstream heuristic filters.

D Removed Samples

Table 6 shows an example of a discarded sample for each filter of the filter stages.

E Memorization Example

Table 7 show examples where AMALIA memorized more then usual, as these were exceptions of our experiments.

[9] https://arquivo.pt/archivepagenow

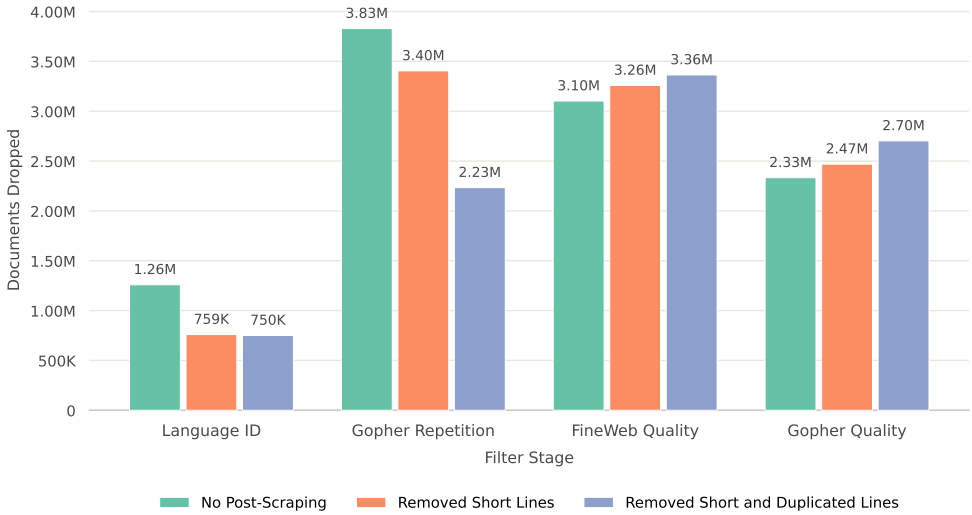

Figure 9: Comparison of the dropped documents with the different post-scraping approaches, per filtering stage.

Before Post-Scraping	**After Post-Scraping**
Duplicated paragraph	
Royal Blood ou a energia do rock\nNo último concerto desta digressão, a banda de rock britânica encheu quinta-feira o Coliseu dos Recreios, em Lisboa.\nRoyal Blood ou a energia do rock\nNo último concerto desta digressão, a banda de rock britânica encheu quinta-feira o Coliseu dos Recreios, em Lisboa.	Royal Blood ou a energia do rock\nNo último concerto desta digressão, a banda de rock britânica encheu quinta-feira o Coliseu dos Recreios, em Lisboa.
Repeated boilerplate label	
Vida e Carreira\nVida e Carreira\nVida e Carreira\nVida e Carreira\nJuntámos tudo o que gosta num só site! Mais simples, mais confortável e disponível também no seu telemóvel ou tablet. Esperemos que goste e que continue a acompanhar-nos diariamente.	Vida e Carreira\nJuntámos tudo o que gosta num só site! Mais simples, mais confortável e disponível também no seu telemóvel ou tablet. Esperemos que goste e que continue a acompanhar-nos diariamente.
Pagination header and footer	
Notícias\nLeiria: GNR afasta hipótese de crime na morte de casal\nEdição de:\nAs autoridades afastam a hipótese de ter ocorrido um crime na morte de um casal que foi encontrado dentro de uma viatura na passada sexta-feira à tarde num pinhal do Coimbrão, concelho de Leiria.\nPáginas\n- « primeira\n- ‹ anterior\n- 1\n- 2\n- 3\n- 4\n- 5\n- 6\n- 7	Leiria: GNR afasta hipótese de crime na morte de casal\nAs autoridades afastam a hipótese de ter ocorrido um crime na morte de um casal que foi encontrado dentro de uma viatura na passada sexta-feira à tarde num pinhal do Coimbrão, concelho de Leiria.
Low alphabetic ratio	
Ribeiro Cristóvão\n19-03-2015 07:24:37\nO crivo da Liga dos Campeões apenas deixou passar para os quartos-de-final da competição o Futebol Clube do Porto, Real Madrid, Barcelona, Atlético de Madrid, Bayern de Munique, Juventus, Mónaco e Paris Saint-Germain.	Ribeiro Cristóvão\nO crivo da Liga dos Campeões apenas deixou passar para os quartos-de-final da competição o Futebol Clube do Porto, Real Madrid, Barcelona, Atlético de Madrid, Bayern de Munique, Juventus, Mónaco e Paris Saint-Germain.
Structural/navigation markers	
|\n2015-06-12\nAssociações de Pais defendem que alunos deviam ter apenas um mês de férias, concorda?\n|\n|\n|\n|\n2015-06-09\nRui Vitória é a melhor opção para o Benfica?\n|\n|\n|\n|\n2015-06-04\nQuem ficou a ganhar com a mudança de Jesus para Alvalade?\n|\n|\n|\n|\n2015-05-31\nO Sporting é um justo vencedor da Taça de Portugal?\n|\n|\n|	Associações de Pais defendem que alunos deviam ter apenas um mês de férias, concorda?\nRui Vitória é a melhor opção para o Benfica?\nQuem ficou a ganhar com a mudança de Jesus para Alvalade?\nO Sporting é um justo vencedor da Taça de Portugal?

Table 5: Examples of documents rescued by the post-scraping step, comparing their raw form to the result after post-scraping. Each example is labeled with the issue that would otherwise have caused the document to be discarded by the heuristic filters.

Filter Stage	Filtering Reason	Discarded Text
Gopher Repetition	Top 2-gram	Rever episódios rtp.pt/beirais\nRever episódios rtp.pt/aguademar\nRever Episódios\nRever episódios rtp.pt/osfilhosdorock\nRever episódios
	Top 3-gram	Resultados e Classificações\nUEFA Champions League 2014/2015\nC. Amarelos | C. Vermelhos |\nMelhor Ataque |\nMelhor Defesa |\nESTATÍSTICAS DA UEFA Champions League\nClassificação
	Top 4-gram	A Microsoft quer estrear no Windows 10 uma loja única para comercializar vídeos, apps e músicas.\nLer mais: Windows 10 vai ter uma loja única para vídeos, músicas e apps
	Duplicated 5-grams	Os GNR voltam ao palco do Coliseu do Porto.\nOs GNR voltam ao palco do Coliseu dos Recreios.\nOs Apocalyptica dão um concerto no Coliseu do Porto para apresentar o novo álbum.\nOs Apocalyptica dão um concerto no Coliseu dos Recreios para apresentar o novo álbum.
	Duplicated 6-grams	Diana Chaves na capa da 'Women's Health'\nBom dia!!! Já viram quem está na capa da Women's Health Portugal deste mês??\n"Bom dia!!! Já viram quem está na capa da Women's Helath Portugal deste mês??", escreveu a atriz e apresentadora Diana Chaves na legenda da foto que publicou no Facebook.
	Duplicated 7-grams	Trailer - Mudar de Vida (versão digital restaurada)\nTrailer oficial de Mudar de Vida (versão digital restaurada) Com Geraldo Del Rey, Isabel Ruth e María Barroso. Realização de Paulo Rocha.
	Duplicated 8-grams	"Educação para uma Cultura de Segurança"\nRealizou-se o Colóquio intitulado "Educação para uma Cultura de Segurança", evento promovido pela Secretaria Regional de Educação e Recursos Humanos, em parceria com a Associação Insular de Geografia. Fotos: FM
	Duplicated 9-grams	Mariana Monteiro aventureira\nA tirar partido da Natureza , sinto-me como uma criança\nMariana Monteiro publicou uma fotografia no Facebook, onde aparece em modo desportivo a descontrair na natureza. "A tirar partido da Natureza , sinto-me como uma criança", escreveu na legenda da imagem.
	Duplicated 10-grams	Iva Lamarão e André Santos mais apaixonados que nunca\nO mundo é pequeno para o nosso amor. Contigo. Amo te\n"O mundo é pequeno para o nosso amor. Contigo. Amo-te", escreveu Iva Lamarão na legenda da imagem, onde aparece com o namorado, o futebolista André Santos.
FineWeb Quality	Line punctuation ratio	Descarregue a aplicação:\npublicidade
	List ratio	Que Música Tocou?\nOS AZEITONAS\nNOS DESENHOS ANIMADOS\nWHO'S GONNA RIDE YOUR WILD HORSES\nCHRISTINA PERRI\nJAR OF HEARTS\nSegunda-Feira\nTerça-Feira\nQuarta-Feira\nQuinta-Feira\nSexta-Feira
Gopher Quality	Below alpha threshold	A iniciativa acontece entre hoje e quarta-feira, de 11 a 13 de Maio, e alarga-se a todo o país. Os organizadores esperam atrair, no mínimo, 200 mil espectadores.\nLer mais: Festa do Cinema com bilhetes a 2,5 euros até quarta-feira
	Short document	Ronaldo pode assumir a liderança dos melhores marcadores\nA redireccionar...
	Long document	"O desnorte está a instalar-se nas hostes governamentais e para-governamentais.\n*etc.*." (This discarded document had 237k tokens)
	Too many end ellipsis	Adriano Lucas (1925-2011)\nAdriano Callé Lucas\nHá 25 anos a informar...\nA câmara do Bombarral aprovou ontem um Plano de Emergência Social para ajudar carenciados a pagar despesas básicas, como medicamentos, água e...\nNão existem notícias em destaque.
	Too many bullets	- As opiniões de Alexandre Mestre (Secretário de Estadodo Desporto),Fernando Gomes (Presidente da FPF) Vitor Baía, Rui Costa, Beto e Barroso.\n- O olhar feminino sobre os quatro "grandes".\n- Os gostos extrafutebol dos craques da bola.\n- Para que servem as equipas B dos maiores clubes?\n- Por que razão arrancam especificamente este ano?
	Not enough stop words	Diário de Notícias\nNecrologia\nPARTICIPAÇÃO\nMaria Ângela Jardim Felgueira\nSeus irmãos: Maria Inês Jardim Felgueira Soares, seu marido,...

Table 6: Examples of discarded documents for each filter. This shows the importance of a set of rigorous filtering stages, as these documents would undermine our final corpus.

Prompt (*italic*); continuation: verbatim match / divergence / ground truth.	
Example 1	
AMALIA *(26 w)*	*Se tiver com níveis de tensão arterial elevados, é importante que saiba que isso pode ter [. . .] nas artérias e na circulação. O nível de tensão arterial é considerado elevado quando os* valores de tensão arterial sistólica são iguais ou superiores a 140 mmHg e os valores de tensão arterial diastólica são iguais ou superiores a 90 mmHg. \nA tensão arterial elevada é uma doença silenciosa, uma vez que não apresenta sintomas. Por isso, é importante que faça exames[. . .] Porque é que a tensão arterial é tão importante? \n É importante que conheça os seus níveis de tensão arterial. A boa notícia é que é[. . .]
Example 2	
AMALIA *(18 w)*	*Pode consultar Legislação e Outros Documentos de três formas distintas: acedendo a [. . .] Quando se deparar com este símbolo, na secção de "Documentos Legais" ou "Publicações",* saiba que a legislação ou outro tipo de documento estão disponíveis para download, em formato PDF ou DOC. \nPara aceder à legislação ou outros documentos através da secção "Legislação e Regulamentação" deve pressionar em "Legislação e[. . .] Quer na área das "Eu Cidadão e. . . " quer das "Áreas de Interesse", pode aceder à legislação relativa ao tema que está a consultar[. . .]
Example 3	
AMALIA *(12 w)*	*Título: Bocage – a vida apaixonada de um genial libertino \n Autor: Luís Rosa \n Editora: [. . .] (passageiros e carga) têm origem e destino, predominantemente, na zona geográfica em que* se encontram localizados; que possuem uma bacia de atracção de tráfego aéreo mais extensa; com ligações aéreas regulares para um ou dois aeroportos de tipo superior, não mais – todos eles localizados, por[. . .] já bastante ampla e profunda, servida principalmente por outros meios de transporte; com ligações aéreas regulares a uma[. . .]
Example 4	
AMALIA *(11 w)*	*«A escola, que tão mal ensina a escrever, não ensina, de todo, a falar.» Quem o afirma é [. . .] e quando a falam.» \n Lembra, contudo, o que para ele é uma «obsessão»: «(...) A escola,* que tão mal ensina a escrever, não ensina, de todo, a falar.» \n «(...) A língua é um organismo vivo, e como tal, sujeito a mutações, a transformações, a adaptações, a reacções, a reacções[. . .] (...) A aprendizagem elementar da fala e o desenvolvimento da língua estão entregues às famílias, ao meio técnico e cultural em[. . .]
Example 5	
AMALIA *(10 w)*	*Segunda-feira, 18 de Junho de 2007 \n É como um mealheiro de sentimentos e de estados de [. . .] que amor são esses, quando uma pessoa segue para a esquerda e a outra para a direita? Que* amizade ou que amor são esses, quando uma pessoa se afasta e a outra se aproxima? Que amizade ou que amor são esses, quando uma pessoa se afasta e a outra se afasta? Que amizade ou[. . .] limita a seguir os passos da outra? A amizade e o amor só fazem sentido quando as pessoas caminham no sentido da convergência. Só[. . .]

Table 7: The five documents AMALIA reproduces most verbatim, ranked by the number of consecutive matching words. For each, green marks the words AMALIA reproduces exactly from the held-out continuation, red marks where it diverges, and blue is the ground-truth continuation. Even in the worst case the model reproduces only 26 consecutive words.