

Is educational content enough? The linguistic and thematic effects of educational filtering

Urho Kunnas, Veronika Laippala, Sampo Pyysalo, Tomasz Galica

TurkuNLP, University of Turku

{urho.r.kunnas, mavela, spyysalo, tomasz.galica}@utu.fi

Abstract

Corpus filtering strategies have evolved steadily, leading to increasingly effective data selection techniques that improve downstream LLM performance. However, while most filters focus on knowledge-rich and educational content, their effect on the linguistic and thematic composition of the documents in pre-training corpora remains underexplored. In this paper, we investigate how educational filtering affects both linguistic and thematic content by combining Propella-1 educational value annotations and the FinePDFs-Edu filter with the Multi-CORE register (genre) framework. Focusing on the Finnish subset of the HPLT 4.0 corpus, we examine how educational filtering changes the distribution of registers and how educational value relates to register categories. By studying machine translated documents with high educational value we show that educational filters focus on content over textual quality. We further analyze changes in the thematic composition using Propella-1 business sector labels, examining whether educational filtering shifts the representation of different content domains. Finally, to assess the impact of filtering, we trained three small language models and evaluated their downstream performance on benchmark tasks. We find that educational filters disagree in their assessment of document educational value, show similar preferences for linguistic registers, but differ in the thematic content they select, resulting in different downstream performance across task types.

1 Introduction

Web-corpora are the main source of pre-training data for large language models (LLMs) such as Apertus (Hernández-Cano et al., 2026), Olmo3 (Olmo et al., 2026) or DCLM-7B (Li et al., 2025). As the amount of web data continues to grow, data curation objectives have shifted from solely maximizing data volume towards improving average

data quality. Higher-quality pre-training data enables LLMs to achieve the same or better downstream performance using fewer training tokens. This in turn decreases LLM training cost and time (Li et al., 2025).

A popular strategy for improving the quality of pre-training corpora is the use of educational filters offering either binary or ordinal classification. These filters estimate the educational value of documents, defined as their usefulness for learning and knowledge transfer. Documents containing clear explanations, structured information, instructional or factual content are graded with higher educational value than e.g. entertainment-oriented text.

Studies on educational filters demonstrate that higher-quality data can substantially improve training efficiency. Models trained on FineWeb-Edu (Lozhkov et al., 2024) and Ultra-FineWeb (Wang et al., 2025) matched or exceeded the English downstream benchmark performance of a model trained on the original FineWeb dataset (Penedo et al., 2024) while using less training data. However, these gains come at the cost of aggressive filtering: FineWeb-Edu removes 92% and Ultra-FineWeb 93% of the data originally available in FineWeb.

Such substantial reductions may change the distribution of pre-training corpora by reducing topic coverage and removing cultural references or other information that is not directly captured by standard benchmarks. Moreover, relying on downstream performance is challenging, as such benchmarks capture only a subset of language model capabilities and do not directly assess how changes in training data composition affect the model’s coverage of different knowledge domains, language varieties, or text types. Furthermore, it remains unclear how content-based filters such as FineWeb-Edu respond to stylistic and linguistic features of texts. This is particularly important given the many dimensions of linguistic variation that may be overlooked when filtering primarily for educational value, including

dialects, informal language, and situational variation.

Our linguistic analysis focuses on situational variation using the Multilingual CORE (MultiCORE) register (genre) framework (Henriksson et al., 2024) targeting the entire multilingual web. The framework categorizes web documents according to *registers* (Biber and Conrad, 2019), situationally defined text varieties such as news articles, discussion forums and how-to pages. Registers cover systematic differences in the situational context of the text, including communicative purpose, writing style, and linguistic characteristics, and have been shown to be one of the strongest predictors of linguistic variation (Biber, 2012).

We analyze how educational filtering changes the distribution of registers and thematic content in the Finnish subset of HPLT 4.0. In our study, we employ the FinePDFs-Edu classifier-based quality filter (CQF) (Lozhkov et al., 2024) and Propella-1-4b annotations (Idahl et al., 2026) for educational value and business sector labels.

FinePDFs-Edu is a well established CQF, trained on 1.3M annotations generated by Qwen3-235B-A22B-Instruct-2507 (Team, 2025) for FinePDFs (Kydlíček et al., 2025) dataset samples. FinePDFs-Edu assigns continuous scores, where higher scores indicate a better educational value of the document.

Propella-1 is a family of multilingual models that annotate documents across six semantic dimensions. In this work, we focus on two of the dimensions: educational value and business sector. The educational value annotation enables a direct comparison with the FinePDFs-Edu filter, while business sector labels are used to analyze how educational filtering affects the thematic composition of the corpus. Business sector categories include, for example, academic research, human resources, media and entertainment, and the gaming industry.

To investigate how educational filters and associated changes of register and thematic content distribution affect model behavior, we trained three small language models. We use model trained on unfiltered Finnish subset of HPLT 4.0 as a reference model, and compare it with models trained on data filtered by FinePDFs-Edu and Propella-1.

2 Related work

Data filtering strategies have been discussed in the past in multiple contexts beyond purely improving the quality of pre-training corpora. Albalak et al.

(2024) show two main approaches to data filtering: distribution matching and distribution diversification, with classifier based quality filtering (CQF) typically belonging to the former category. Albalak et al. (2024) shows that CQF can introduce bias in processed datasets: against certain groups, demographics, dialects, or data types. Additionally, one-dimensional quality classifiers such as educational filters can unintentionally favor stylistic features over semantic content. For example, Klimaszewski and Andruszkiewicz (2026) demonstrate how simply rewriting a document in Wikipedia-like style can bypass the FineWeb-Edu filter (Lozhkov et al., 2024), allowing non-educational or otherwise undesirable content to pollute the dataset. This study indicates that more focus on understanding what document features are prioritized by CQFs are needed.

While classifier-based filtering might unintentionally distort the distribution of data, several approaches to intentionally alter the composition of pre-training corpora exist. Source-domain or knowledge-domain re-weighting is commonly used to improve LLM downstream-task performance. Gao et al. (2020) uses heuristically determined data sampling weights to compile multiple source-domains (e.g. GitHub¹ or PubMed²) into the PILE dataset. Xie et al. (2023) demonstrates the Domain Reweighting with Minimax Optimization (DoReMi) approach. DoReMi automatically optimizes sampling proportions of source domains with small proxy models, producing a data mixture that improves LLM downstream performance.

Longpre et al. (2024) notes that domain heterogeneity, the diversity of syntactic, topic and reasoning structures, is beneficial for downstream performance, hence source-domain diversity should be prioritized. Additionally, Longpre et al. (2024) shows that the effectiveness of one-dimensional (single quality) filtering depends on the characteristics of the underlying pre-training corpus, indicating that no universally optimal strategy exists.

While data filtering has been extensively studied in the terms of downstream performance, comparatively little attention has been paid to how filtering alters the linguistic variations of pre-training corpora. In a recent study, Wettig et al. (2025) present the WebOrganizer tool to study domain and format topics, indicating that topic re-weighting can improve downstream evaluation tasks as well.

¹<https://github.com/>

²<https://pubmed.ncbi.nlm.nih.gov/>

Recent work on the Olmo3 model (Olmo et al., 2026) utilized WebOrganizer, and featured studies on the distribution of topics in pre-training corpora. However, the topic ratios used for constructing the pre-training mixture were optimized purely for downstream performance.

Similarly Myntti et al. (2025) studied how registers and different register compositions influence downstream performance for English language. They extracted datasets consisting of single registers, and used them to train series of small language models, which were subsequently evaluated against benchmarks. Their studies indicate that individual registers as well as combination of registers can lead to changes in model downstream performance. However, the study did not compare the register-based filtering with other CQFs such as FineWeb or FinePDFs-Edu.

In this paper we address the challenges by investigating how educational filters reshape the distribution of web registers and thematic content in pre-training corpus. We compare CQF FinePDFs-Edu with LLM-based Propella-1 annotations, and with the MultiCORE framework.

3 Methods

3.1 Data

Our study focused on the Finnish subset of the "clean" partition of the HPLT 4.0 dataset³. The "clean" partition contains documents that have been filtered to remove low-quality and noisy content such as documents consisting mostly of HTML navigation elements, or boilerplate text. HPLT 4.0 is an updated version of HPLT 3.0 described by Oepen et al. (2026), and contains documents in 195 languages, sourced from Internet Archive, the Common Crawl and ArchiveBot. The Finnish subset used in our study consists of 51 million documents. Data was annotated by dataset creators with the MultiCORE register classifier (Henriksson et al., 2024), Propella-1-4b (Idahl et al., 2026), FinePDFs-Edu (Kydľčec et al., 2025) and several other classifiers.

3.2 Data labeling

We used the web register probabilities included in HPLT 4.0 to classify documents with registers. The MultiCORE classifier follows the register scheme developed by Laippala et al. (2023) based on the original work by Biber and Egbert (2018). The

scheme is hierarchical with 9 main registers and 47 sub-register categories. Following the approach of Myntti et al. (2025), we use a threshold of 0.4 to assign register labels. A document may receive one or several register labels depending on its linguistic characteristics. Such multi-label documents are called "hybrid" registers. The main register labels and their descriptions are explained in table 1. The sub-registers are more detailed categories; for example, the NA register includes News articles and Personal blogs, the IN register includes Encyclopedia and Research articles, and the OP register includes Reviews and Advice.

Register label	Register description
HI	How-to-instructions
ID	Interactive discussion
IN	Informational description
IP	Informational persuasion
LY	Lyrical
MT	Machine translation
NA	Narrative
OP	Opinion
SP	Spoken

Table 1: Main register labels used in our study

In addition to registers, we study LLM-based Propella-1 annotations `educational_value` and `business_sector`, as well as FinePDFs-Edu scores.

The `educational_value` feature estimates the instructional value of a document, ranging from content with no educational value to structured material that explicitly supports learning through clear explanations, examples, and progressive knowledge building. Propella-1 `educational_value` assigns five categories to documents: None, Minimal, Basic, Moderate, High.

The Propella-1 `business_sector` is a multi-label feature with 37 possible classes such as "academic research" and "technology software". In the HPLT 4.0 cleaned Finnish subset, the average number of labels per document is 1.9. To simplify our analysis we logged each class in a document as a separate entry, resulting in 98 million business sector entries.

FinePDFs-Edu is a CQF and it assigns a continuous numerical score ranging from 0 to 5, reflecting its suitability as educational content for primary to grade school levels. Documents containing detailed explanation, coherent reasoning, and clear writing receive higher scores, while documents characterized by non-academic writing, poor coherence or superficial treatment of the topic re-

³<https://hplt-project.org/datasets/v4.0>

ceive lower scores.

Propella-1 educational_value and FinePDFs-Edu scores were used as present in HPLT 4.0 and were not post processed by us.

3.3 Training and evaluation of small language models

We trained three small language models with the Llama architecture (Touvron et al., 2023), each with 0.6 billion trainable parameters and a training budget of 100 billion tokens. Training was conducted using the NVIDIA Megatron-LM framework (Shoeybi et al., 2019). The models had 22 hidden layers, a hidden size of 1024, and 16 attention heads. We used a sequence length of 4096, a global batch size of 1024, and the Adam optimizer. Hyperparameters were selected based on our initial experiments. All computing was carried out on the CSC-LUMI supercomputer. Additional details are given in Appendix A table 5

To measure the effect of filtering, we trained a model on the unfiltered Finnish subset of HPLT 4.0, which we use as the baseline for downstream performance comparisons. We used Propella-1 to construct a training corpus consisting only of the high, moderate, and basic educational classes, resulting in a collection of 33 billion tokens (BT). We then selected the 33 BT with the highest FinePDFs-Edu scores to construct a corpus of the same size. The two filtered datasets were subsequently used to train separate small language models. All models were trained on 100 BT budget with the same number of steps. However because filtered datasets were smaller than original Finnish subset of HPLT 4.0, the baseline model underwent 1 epoch, and filtered models 3 epochs of training.

To evaluate the three trained models, we used the LM-evaluation-harness (Gao et al., 2024) framework. We selected the following tasks:

- ARC:Challenge (Clark et al., 2018): multiple-choice exam questions requiring multi-hop reasoning. We used the LumiOpen/arc_challenge_mt implementation ⁴.
- Belebele (Bandarkar et al., 2024): Multiple-choice reading comprehension task with 900 machine translated and human annotated questions. ⁵

⁴https://huggingface.co/datasets/LumiOpen/arc_challenge_mt

⁵<https://huggingface.co/datasets/facebook/>

- MultiBLiMP (Jumelet et al., 2025): Linguistic minimal pairs test, evaluating subject-verb agreement for number, person, and gender. ⁶
- HellaSwag (Zellers et al., 2019): Common sense reasoning task for sentence completion. We used the LumiOpen/opengpt-x_hellaswagx implementation ⁷ (Thellmann et al., 2024)

This selection allowed us to observe different dimensions of model performance. All tasks were run with a zero-shot approach.

4 Results and discussion

We extracted the metadata from the Finnish subset of HPLT 4.0, and divided it into five separate groups respective to the Propella-1 educational_value classes. We compared the register and business_sector distributions between the groups as well as with the distribution of the entire Finnish HPLT 4.0. Table 2 shows the proportion of documents belonging to each Propella-1 class in the dataset.

The recommended use of the FinePDFs-Edu filter is to utilize only the top 10% of the highest scored documents. However, to enable a direct comparison between FinePDFs-Edu and Propella-1, we calculated the FinePDFs-Edu score threshold values which retain the same amount of documents as Propella-1 educational_value classes, and repeated the metadata extraction. The thresholds and corresponding Propella-1 classes are: <0.3771 (none), 0.3771-0.8929 (minimal), 0.8929-1.1998 (basic), 1.1998-2.6491 (moderate), >2.6491 (high). Matching the amount of retained data is important from LLM pre-training perspective, where models trained with a fixed token budget can otherwise be exposed to different numbers of epochs depending on the size of the filtered dataset. Controlling the amount of training data therefore allows for direct comparison.

4.1 Comparison of Propella-1 educational values and FinePDFs-Edu

The relationship between FinePDFs-Edu and Propella-1 educational_value can be seen in Figure 1. The visible mismatch between both filters

belebele

⁶<https://huggingface.co/datasets/jumelet/multiblmp>

⁷https://huggingface.co/datasets/LumiOpen/opengpt-x_hellaswagx

Educational value	Percent of dataset
None	38.19%
Minimal	34.13%
Basic	17.54%
Moderate	9.27%
High	0.87%

Table 2: Proportions of Propella-1 educational values in the dataset

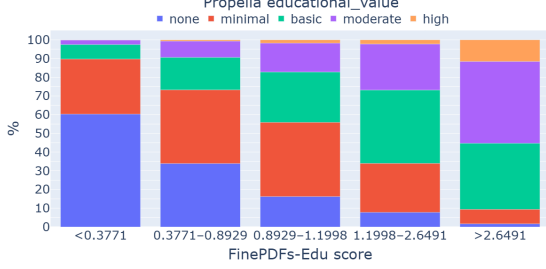


Figure 1: Distribution of propella-1 educational value within FinePDFs-Edu score buckets of matching size

can be seen. At the same time, a trend is observed: with increasing FinePDFs-Edu score, the amount of documents labeled with Propella-1 none and minimal educational classes decreases, while the amount of moderate and high documents increases. It is worth to note that both classifiers mismatch on all educational classes: Propella-1 educational labels none and high can be found throughout all the FinePDFs-Edu score ranges. We further calculated the Spearman rank correlation between both filters by assigning scores (0-4) to Propella-1 educational categories (with 0 = none and 4 for high). The filters correlate with a score of 0.49, which indicates moderate correlation. All these values support finding of weak-to-moderate correlation between filter scores. In addition we calculated Kendalls $\tau_B = 0.37$ and quadratic Cohen’s $\kappa = 0.46$ by assigning FinePDFs-Edu to 5 categories as shown in table 3b. Our finding indicates that use of only a single filter might be insufficient to identify the educational content in the corpus. Moreover, the disagreement between filters suggests that both CQF and LLM-based educational filters may misclassify the educational value of documents, either by overlooking educational content or by retaining documents with limited educational value.

4.2 Changes in register distributions

To quantify the effect of educational filtering, we calculated the χ^2 distance between the distributions of registers in full corpus and each educational

category. Let p_j denote the overall proportion of observations relating to register j :

$$p_j = \frac{\sum_c n_{jc}}{\sum_j \sum_c n_{jc}} \quad (1)$$

where n_{jc} is the number of observations belonging to register j and educational category c . The proportion of observations belonging to register j within educational category c is defined as:

$$p_{j|c} = \frac{n_{jc}}{\sum_j n_{jc}}. \quad (2)$$

And finally, the χ^2 distance for educational category c is then calculated as:

$$D_{\chi^2}(c) = \sum_j \frac{(p_{j|c} - p_j)^2}{p_j}, \quad (3)$$

The distance is not normalized to a fixed range. A value zero indicates that the sector distribution within the category is identical to the overall distribution, while larger values indicate greater divergence. As shown in Table 3, the two filters produce broadly similar patterns: register distributions remain close to the full corpus for the lower educational categories, while divergence increases for the highest categories. This can be associated with lower educational classes having more documents, and only the highest classes/scores being very selective. For Propella-1, both moderate and high show strong divergence, whereas for FinePDFs-Edu, it is observed only in the highest score bucket. Interestingly, for both filters, the second category exhibits lower divergence than the first, despite the lowest category containing the most documents.

We evaluated this finding further by inspecting register distributions within the buckets as seen in Figure 2. For simplicity, we show only the 9 main register categories, and due to a large number of hybrid register classes, all hybrids are grouped together. A numerical breakdown of the registers is given in Appendix B tables 6 and 7.

Both filters produce similar register distributions. Figure 2 shows that, for both filters, the proportion of Informational Description (IN) and How-to (HI) documents increases with educational value, while Narrative (NA) and Informational Persuasion (IP) documents show the opposite trend. Both effects are stronger for the Propella-1 filter. In addition, the Interactive Discussion (ID) register is less present in the two highest educational categories. Lyrical (LY) constitutes a marginal part

(a) Educational value	χ^2 distance
None	0.1097
Minimal	0.0467
Basic	0.1226
Moderate	1.0636
High	2.5612

(b) FinePDFs-Edu score	χ^2 distance
<0.3771	0.117457
0.3771–0.8929	0.044733
0.8929–1.1998	0.078526
1.1998–2.6491	0.225581
>2.6491	1.125385

Table 3: χ^2 distances of main registers distribution for a) Propella-1 educational value categories, b) FinePDFs-Edu score buckets

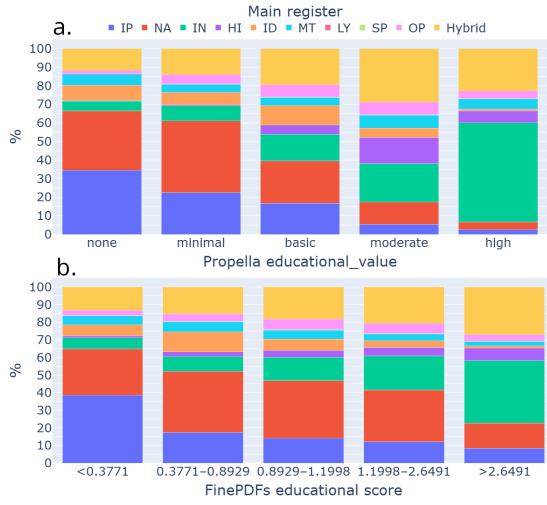


Figure 2: Register distributions for a) Propella-1 educational categories and, b) FinePDFs-Edu score categories of matching size

of the set (<1%), however, LY is mostly present in the lowest educational categories. The changes of distributions of Opinion (OP), Spoken (SP) and Machine Translation (MT) registers are smaller, and their proportion in the educational categories stays relatively constant.

These findings do not completely agree with the results described by Myntti et al. (2025). While numerous studies, such as (Lozhkov et al., 2024), show that educational filters improve the downstream performance of LLMs, Myntti et al. (2025) used register-based filtering. Their results show that datasets consisting only of HI, IN, or OP documents achieve higher downstream performance than datasets consisting of IP or MT documents. However, the presence of MT in the high educational classes and OP in the low educational classes stands in contrast to these findings. Moreover, our finding of an inverse relationship between ed-

ucational class and the NA register, together with the prevalence of OP in lower educational classes, stands in contrast to Myntti et al. (2025)’s finding that a combination of HI, IN, OP, and NA registers improves the performance of the baseline model. This could indicate that the main source of improved downstream performance is related to the HI and IN registers.

The failure for educational filters to remove MT documents from high educational categories can be explained by filters often prioritizing structured material, explanations, and examples (i.e. the content of the document) rather than style or language quality. The MT register is assigned to a document based on its linguistic characteristics (i.e., translation artifacts). This suggests that educational filters may be prone to overlooking linguistic correctness. To further support this claim, we investigated the distribution of MT hybrid registers. Figure 3 presents the distribution of MT hybrid registers for a) Propella-1 and b) FinePDFs-Edu filters. The label ‘multi’ refers to hybrid registers combining machine translation with two or more other registers (e.g. MT-HI-IN). The numeric breakdown of prevalence of machine-translated documents can be found in Appendix B, Tables 8 and 9.



Figure 3: Hybrid MT document distribution within a) Propella-1 educational categories and, b) FinePDFs-Edu score categories of matching size

The distribution of MT hybrid registers reveals the same trend as observed within main registers for IN and HI: the amount of MT-IN and MT-HI register documents increases with increasing educational_value.

To further strengthen our finding, we inspected a small random sample of texts with MT-IN or MT-HI register labels, Propella-1 educational value

(a) MT-HI register

Original text: Käytä kahta kaarevaa viivaa, jotka on yhdistetty typeriin pisteeseen, jotta dorsal fin on kalan takaosassa.

Translation: "Use two curving lines that are connected to stupid point so the dorsal fin is at the back end of the fish"

(b) MT-IN register

Original text: Näiden naisten, monien aatelisperheen, varallisuus auttoi häntä perustamaan löytöeläimen ja muita sairaaloita.

Translation: "The wealth of these women, many noble family's, helped him found a rescue animal and other hospitals."

Figure 4: Examples of documents labeled as (a) MT-HI register with educational_value: high, FinePDFs-Edu score: 2.7322; and (b) MT-IN register with educational_value: moderate, FinePDFs-Edu score 3.0498.

of high or moderate and a FinePDFs-Edu score of >2.6491 . Many of the examined texts exhibited clear characteristics of machine translation errors. In some of the examples, translation artifacts made the text difficult to interpret by native speakers. Common features included grammatical errors, contextually mismatched words, and left-over non-Finnish words. In some texts individual sentences are fairly coherent, but do not form one cohesive text. The examples of MT register documents are shown in Figure 4.

The example in Figure 4a contains detailed instructions for drawing fish meant for children. This example sentence contains several features that highlight the way educational filters focus on content, not linguistic features. The most notable is the use of the English "dorsal fin" rather than the Finnish "selkäevä". The noun phrase "typeriin pisteeseen" contains two additional errors. The adjective "typerä" ("stupid") is clearly a mistranslation. There is also a grammatical error, where the adjective is in its plural form rather than in the singular to match the singular of the noun "pisteeseen" ("to a point").

The MT-IN example shown in Figure 4b is describing the life of a historical saint. The sentence contains a number-agreement error ("monien aatelisperheen" instead of "monien aatelisperheiden") and a likely lexical translation error, where

"löytöeläimen" ("rescue animal") is semantically incompatible with "perustaa" ("to establish").

The above findings suggest that educational filters can be bypassed by MT documents even if only the highest educational categories are used. This in turn can affect linguistic correctness of text generated by LLMs.

4.3 Changes in business sector distributions

Similarly to the analysis of registers presented above, we calculated χ^2 distances for business_sector with equations 1, 2, 3 where j now denotes business_sector category. The scores for Propella-1 and FinePDFs-Edu are presented in Table 4. We observe a similar trend to that found in the register analysis: the Propella-1 high educational category deviates the most from the distribution of the general Finnish subset. Furthermore, as with the register distributions, the Propella-1 filter results in a stronger deviation from the general distribution than FinePDFs-Edu.

(a) Educational value	χ^2 distance
None	0.1931
Minimal	0.0426
Basic	0.0965
Moderate	0.3785
High	1.9368

(b) FinePDFs-Edu score	χ^2 distance
<0.3771	0.2545
0.3771–0.8929	0.0447
0.8929–1.1998	0.1771
1.1998–2.6491	0.3544
>2.6491	1.3690

Table 4: χ^2 distances of business sector distributions for (a) Propella-1 educational value categories and (b) FinePDFs-Edu score buckets.

Unlike educational_value, and similar to register labels, each document can have a single or multiple business_sector categories. Figure 5a shows the business_sector distribution for the Finnish HPLT 4 dataset. We treat each label as separate entry i.e., we split each multilabel-document pair to label-document pairs. The most populated categories within the Finnish subset of HPLT 4.0 are: general interest, media and entertainment, and government public. The distribution changes significantly for both filters. Figure 5a shows the distribution for the Propella-1 educational category high and 5c the distribution for the respective FinePDFs-Edu score range (the same proportion of the text as educational category high). Both Propella-1 and FinePDFs-Edu increase the proportion of academic

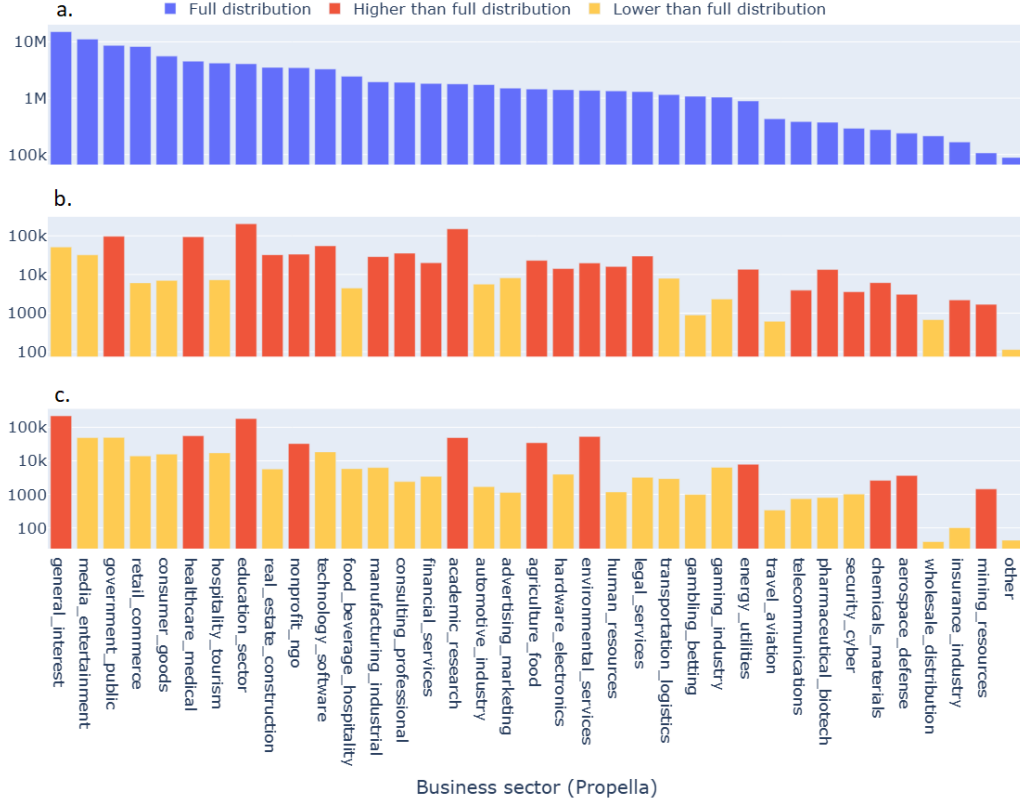


Figure 5: Distributions of business sectors in a) Finnish subset of HPLT 4.0, b) Subset with educational_value: high; c) Subset with FinePDFs score >2.6491. The bar colors on plots b) and c) indicate if the proportion of given sector is higher or lower than for the general distribution.

research documents (8.0x, 6.4x respectively), educational sector (4.8x, 2.9x respectively), as well as 8 others. However, the proportion of "general interest" documents is higher for FinePDFs-Edu than for unfiltered Finnish subset of HPLT 4.0, while for Propella-1 it is lower. In addition, Propella-1 has an increased proportion of 12 business sectors which are less present in the FinePDFs-Edu filtered data.

The distribution of business sectors within educational categories is shown in Appendix C tables 10 and 11. Appendix D Figure 7 shows a version of Figure 5 with the top 10 % of texts sorted by FinePDFs educational score.

These findings contrast with the register analysis, where both filters showed largely consistent shifts towards the same types of registers in highly educational documents. This suggests that, while the two filters identify similar document characteristics associated with educational value such as step-by-step instructions and explanations of concepts, they differ more substantially in how they assess the thematic content of documents. Such differences in thematic composition may in turn affect

the knowledge represented in the resulting models and consequently, model performance across different types of evaluation tasks.

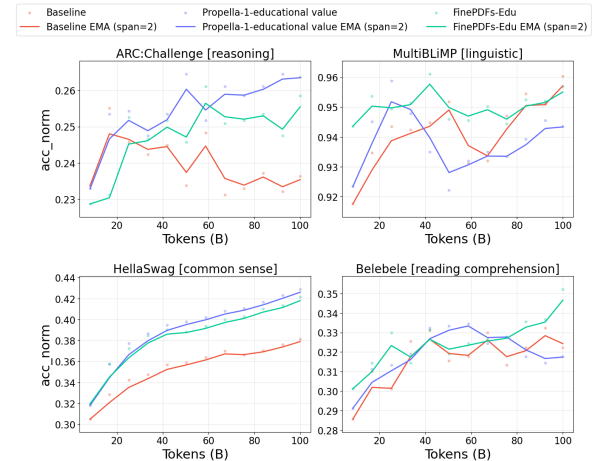


Figure 6: Downstream performance of three trained language models. The dots represent checkpoint performance, while lines show the exponential moving average. The standard error for ARC:Challenge, MultiBLiMP, HellaSwag and Belebele tasks are 0.012, 0.009, 0.004 and 0.015 respectively.

4.4 Downstream performance evaluation

We evaluated three small language models: a baseline model which was trained on the unfiltered Finnish subset of HPLT 4.0, and two models trained on subsets filtered by Propella-1 and FinePDFs-Edu. Figure 6 shows the performances of the evaluation tasks.

The effect of data filtering varies across evaluation tasks. For multi-hop and common sense reasoning tasks (ARC:Challenge and HellaSwag respectively), the model trained with data filtered with Propella-1 outperforms the two others. However, Propella-1 shows drop in performance for the MultiBLiMP linguistic minimal pairs task, which does not recover till end of training. Additionally, for the reading comprehension task (Belebele), all three models perform similarly during training, except for last three checkpoints, where FinePDFs-Edu model outperforms two other. The improved performance of filtered models on multi-hop and common sense reasoning tasks can be associated with increased amount of instructional and informative documents in training corpora.

5 Conclusions

We analyzed how educational filters affect the distribution of linguistic variation and thematic content (business sectors) in a pre-training corpus using Propella-1 annotations, FinePDFs-Edu, the MultiCORE register (genre) framework, and the Finnish subset of HPLT 4.0. Our results show that educational filters shift the distribution towards instructional and informative registers, and away from narrative and persuasive documents. Similarly, filtering increases the representation of academic, technical and medical business sectors.

Our analysis of machine-translated documents further shows that high educational scores do not always correspond to linguistically well-formatted text. Documents containing grammatical and lexical errors are present even among the highest educational categories of both filters. We determined that educational filters focus on content characteristics and ignore linguistic quality that introduces risk of including machine-translated documents in the filtered corpus.

Finally, by studying data distribution changes, training small language models, and evaluating them we show that educational filters have both shared and distinct effects on corpus and model performance. These findings highlight the impor-

tance of considering linguistic and thematic composition alongside downstream performance when evaluating pre-training data filtering strategies.

Limitations

In our work we analyze the effect of educational filters on the change of distributions of linguistic variance as well as business categories of documents within the HPLT 4.0 corpus. However, our study focuses only on the Finnish subset of the dataset. Additional studies should be carried out to fully describe the effect of filters on broader range of languages. Moreover we used only two educational filters: FinePDFs-Edu and Propella-1. Additional studies using other filters could clarify our claims and findings. For qualitative evaluation of machine translated text we only consider a small sample of 15 randomly selected texts, reviewed by one native speaker. Lastly, we train three small language models, but do not carry out model size ablation studies - observed effects might differ for larger model sizes. We carried out tests with multiple choice formula for ARC:Challenge and HellaSwag tasks, meanwhile a cloze formula test could result in higher values.

Acknowledgments

This project has received funding from OpenEuroLLM and LLMs4EU projects, under Digital Europe Programme grant agreements No 101195233 and 101198470, and was supported by the Research Council of Finland through several projects: FINCLARIAH research infrastructure (project 358720, which has also received funding from the European Union– NextGenerationEU instrument), “Mechanisms of Register Variation in Massively Multilingual Web-Scale Corpora” (project 362459), and the “Centre of Excellence Human Diversity through Contacts” (project 374223). We gratefully acknowledge CSC– IT Center for Science Ltd. for providing computational resources. The contents of this publication are the sole responsibility of its authors and do not necessarily reflect the opinion of the European Union.

References

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Hae-won Jeong, Colin Raffel, Shiyu Chang, Tatsunori

- Hashimoto, and William Yang Wang. 2024. [A survey on data selection for language models](#). *Preprint*, arXiv:2402.16827.
- Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabza. 2024. [The belebele benchmark: a parallel reading comprehension dataset in 122 language variants](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.
- Douglas Biber. 2012. [Register as a predictor of linguistic variation](#). *Corpus Linguistics and Linguistic Theory*, 8(1):9–37.
- Douglas Biber and Susan Conrad. 2019. *Register, Genre, and Style*. Cambridge University Press, Cambridge.
- Douglas Biber and Jesse Egbert. 2018. *Register Variation Online*. Cambridge University Press.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. [Think you have solved question answering? try arc, the ai2 reasoning challenge](#). *Preprint*, arXiv:1803.05457.
- Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. [The pile: An 800gb dataset of diverse text for language modeling](#). *Preprint*, arXiv:2101.00027.
- Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac’h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. [The language model evaluation harness](#).
- Erik Henriksson, Amanda Myntti, Saara Hellström, Anni Eskelinen, Selcen Erten-Johansson, and Veronika Laippala. 2024. [Automatic register identification for the open web using multilingual deep learning](#). *Preprint*, arXiv:2406.19892.
- Alejandro Hernández-Cano, Alexander Hägele, Allen Hao Huang, Angelika Romanou, Antoni-Joan Solergibert, Barna Pásztor, Bettina Messmer, Dhia Garbaya, Eduard Frank Ďurech, Ido Hakimi, Juan Garcia Giraldo, Mete Ismayilzada, Negar Foroutan, Skander Moalla, Tiancheng Chen, Vinko Sabolčec, Yixuan Xu, Michael Aerni, Badr AlKhamissi, and 83 others. 2026. [Apertus: Democratizing open and compliant LLMs for global language environments](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 46877–46955, San Diego, California, United States. Association for Computational Linguistics.
- Maximilian Idahl, Benedikt Droste, Björn Plüster, and Jan Philipp Harries. 2026. [propella-1: Multi-property document annotation for llm data curation at scale](#). *Preprint*, arXiv:2602.12414.
- Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. 2025. [Multiblimp 1.0: A massively multilingual benchmark of linguistic minimal pairs](#). *Preprint*, arXiv:2504.02768.
- Mateusz Klimaszewski and Piotr Andruszkiewicz. 2026. [Is a document educational or just wikipedia-style? – pitfalls of classifier-based quality filtering](#). *Preprint*, arXiv:2605.23721.
- Hynek Kydlíček, Guilherme Penedo, and Leandro von Werra. 2025. Finepdfs. https://huggingface.co/datasets/HuggingFaceFW/finepdfs_edu.
- Veronika Laippala, Samuel Rönqvist, Miika Oinonen, Aki Juhani Kyröläinen, Anna Salmela, Douglas Biber, Jesse Egbert, and Sampo Pyysalo. 2023. [Register identification from the unrestricted open web using the corpus of online registers of english](#). *Language Resources and Evaluation*, 57(3):1045–1079. Funding Information: Open Access funding provided by University of Turku (UTU) including Turku University Central Hospital. Funding was provided by Academy of Finland (Grant No. 331297), Emil Aaltosen säätiö, National Science Foundation (Grant No. 1147581). Publisher Copyright: © 2022, The Author(s).
- Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, and 40 others. 2025. [Datacomp-llm: In search of the next generation of training sets for language models](#). *Preprint*, arXiv:2406.11794.
- Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ippolito. 2024. [A pretrainer’s guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3245–3276, Mexico City, Mexico. Association for Computational Linguistics.
- Anton Lozhkov, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. 2024. [Fineweb-edu: the finest collection of educational content](#).
- Amanda Myntti, Erik Henriksson, Veronika Laippala, and Sampo Pyysalo. 2025. [Register always matters: Analysis of LLM pretraining data through the lens of language variation](#). In *Second Conference on Language Modeling*.

- Stephan Oepen, Nikolay Arefev, Mikko Aulamo, Marta Bañón, Maja Buljan, Laurie Burchell, Lucas Charpentier, Pinzhen Chen, Mariya Fedorova, Ona de Gibert, Barry Haddow, Jan Hajič, Jindřich Helcl, Andrey Kutuzov, Veronika Laippala, Zihao Li, Risto Luukkonen, Bhavitvya Malik, Vladislav Mikhailov, and 13 others. 2026. [Hplt 3.0: Very large-scale multilingual resources for llms and mt. mono- and bi-lingual data, multilingual evaluation, and pre-trained models](#). *Preprint*, arXiv:2511.01066.
- Team Olmo, :, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, and 50 others. 2026. [Olmo 3](#). *Preprint*, arXiv:2512.13961.
- Guilherme Penedo, Hynek Kydlíček, Loubna Ben al-lal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. [The fineweb datasets: Decanting the web for the finest text data at scale](#). *Preprint*, arXiv:2406.17557.
- Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*.
- Qwen Team. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Klaudia Thellmann, Bernhard Stadler, Michael Fromm, Jasper Schulze Buschhoff, Alex Jude, Fabio Barth, Johannes Leveling, Nicolas Flores-Herr, Joachim Köhler, René Jäkel, and Mehdi Ali. 2024. [Towards cross-lingual llm evaluation for european languages](#). *Preprint*, arXiv:2410.08928.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. [Llama: Open and efficient foundation language models](#). *Preprint*, arXiv:2302.13971.
- Yudong Wang, Zixuan Fu, Jie Cai, Peijun Tang, Hongya Lyu, Yewei Fang, Zhi Zheng, Jie Zhou, Guoyang Zeng, Chaojun Xiao, Xu Han, and Zhiyuan Liu. 2025. [Ultra-fineweb: Efficient data filtering and verification for high-quality llm training data](#). *Preprint*, arXiv:2505.05427.
- Alexander Wettig, Kyle Lo, Sewon Min, Hannaneh Hajishirzi, Danqi Chen, and Luca Soldaini. 2025. [Organize the web: Constructing domains enhances pre-training data curation](#). *Preprint*, arXiv:2502.10341.
- Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy S Liang, Quoc V Le, Tengyu Ma, and Adams Wei Yu. 2023. [Doremi: Optimizing data mixtures speeds up language model pretraining](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 69798–69818. Curran Associates, Inc.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

A Appendix

The small language models were trained with the following hyperparameters:

Parameter	Value
Number of layers	28
Hidden size	1024
Number of attention heads	16
Sequence length	4096
Learning rate	0.002
Learning rate decay type	WSD
Decay style	linear
Warmup iterations	2000
Cooldown fraction	0.2
Weight decay	0.1
Optimizer	Adam
Adam β_1	0.9
Adam β_2	0.95
Adam ϵ	1e-8
Global batch size	512
Micro batch size	2
Training token budget	100BT
Tokenizer	openeurollm/tokenizer- 256k

Table 5: Most important hyperparameter of trained models. The openerollm tokenizer can be found under <https://huggingface.co/openeurollm/tokenizer-256k>

B Appendix

Table 6 shows the register composition of each Propella-1 educational category. The same composition for FinePDFs-Edu score categories is shown in table 7.

	none	minimal	basic	moderate	high
IP	34.42 %	22.57 %	16.77 %	5.51 %	2.64 %
NA	32.04 %	38.58 %	22.92 %	11.95 %	4.09 %
IN	5.28 %	8.05 %	14.26 %	20.69 %	53.40 %
HI	0.23 %	0.59 %	4.91 %	14.08 %	6.40 %
ID	8.31 %	6.67 %	10.60 %	4.98 %	0.95 %
MT	5.99 %	4.22 %	4.30 %	7.05 %	5.54 %
LY	0.16 %	0.03 %	0.01 %	0.02 %	0.02 %
SP	0.09 %	0.20 %	0.29 %	0.28 %	0.18 %
OP	1.74 %	5.03 %	6.38 %	6.63 %	3.95 %
Hybrid	11.73 %	14.07 %	19.57 %	28.81 %	22.83 %

Table 6: Proportion of registers within each Propella-1 educational category

	<0.3771	0.3771–0.8929	0.8929–1.1998	1.1998–2.6491	>2.6491
IP	38.45 %	17.44 %	14.17 %	12.20 %	8.43 %
NA	26.49 %	34.77 %	32.72 %	29.39 %	14.24 %
IN	6.21 %	8.33 %	13.29 %	19.34 %	35.55 %
HI	1.23 %	2.59 %	3.73 %	4.76 %	7.13 %
ID	6.23 %	11.38 %	6.52 %	3.85 %	1.23 %
MT	5.13 %	5.77 %	5.07 %	3.74 %	2.57 %
LY	0.04 %	0.10 %	0.11 %	0.07 %	0.01 %
SP	0.06 %	0.20 %	0.33 %	0.29 %	0.20 %
OP	2.91 %	4.25 %	5.85 %	5.68 %	3.57 %
Hybrid	13.25 %	15.16 %	18.21 %	20.69 %	27.06 %

Table 7: Proportion of registers within each FinePDFs-Edu score category

Tables 8 and 9 show what percentage of texts in each educational score category belong to the MT register or a hybrid register containing MT. The label 'multi' denotes hybrid registers composed of three or more main registers.

	none	minimal	basic	moderate	high
MT	5.9933 %	4.2243 %	4.2998 %	7.0457 %	5.5436 %
IP-MT	2.2730 %	0.5112 %	0.2784 %	0.1285 %	0.1038 %
IN-MT	0.1015 %	0.2167 %	0.3308 %	0.8144 %	1.9455 %
HI-MT	0.0122 %	0.0839 %	0.2221 %	1.3349 %	0.7746 %
MT-OP	0.2636 %	0.1226 %	0.1020 %	0.1345 %	0.0217 %
MT-NA	0.1185 %	0.1035 %	0.0340 %	0.0188 %	0.0072 %
multi	0.0782 %	0.0138 %	0.0152 %	0.0451 %	0.0408 %
ID-MT	0.0325 %	0.0329 %	0.0365 %	0.0197 %	0.0256 %
LY-MT	0.0082 %	0.0016 %	0.0003 %	0.0000 %	0.0000 %
MT-SP	0.0003 %	0.0010 %	0.0010 %	0.0016 %	0.0009 %

Table 8: Amount of machine translated registers as percentage of texts in Propella educational categories

	<0.3771	0.3771–0.8929	0.8929–1.1998	1.1998–2.6491	>2.6491
MT	5.1300 %	5.7697 %	5.0669 %	3.7396 %	2.5721 %
IP-MT	2.1767 %	0.6489 %	0.2376 %	0.1019 %	0.0248 %
IN-MT	0.0519 %	0.2747 %	0.5814 %	0.4845 %	0.3064 %
HI-MT	0.1037 %	0.2697 %	0.2435 %	0.2878 %	0.1924 %
MT-OP	0.3588 %	0.0811 %	0.0378 %	0.0176 %	0.0074 %
MT-NA	0.0921 %	0.1179 %	0.0613 %	0.0233 %	0.0027 %
multi	0.0824 %	0.0200 %	0.0132 %	0.0113 %	0.0134 %
ID-MT	0.0243 %	0.0452 %	0.0324 %	0.0179 %	0.0029 %
LY-MT	0.0011 %	0.0064 %	0.0060 %	0.0009 %	0.0002 %
MT-SP	0.0008 %	0.0010 %	0.0008 %	0.0002 %	0.0000 %

Table 9: Amount of machine translated registers as percentage of texts in FinePDFs educational score categories

C Appendix

Figure 7 shows how the proportion of business sector labels is affected by data selection. Figure 7a shows the baseline distribution of the whole dataset, while 7b shows distribution in documents labelled high by the Propella-1 educational score classifier. 7c shows the distribution in the 10 % of texts with the highest FinePDFs-Edu score. Bar colors in plots b and c indicate whether the proportion of the given sector is higher or lower than in the whole dataset.

D Appendix

Table 10 contains a breakdown of Propella-1 business sectors within the 5 educational categories given by Propella-1. Table 11 shows the same for educational score categories from FinePDFs-Edu.

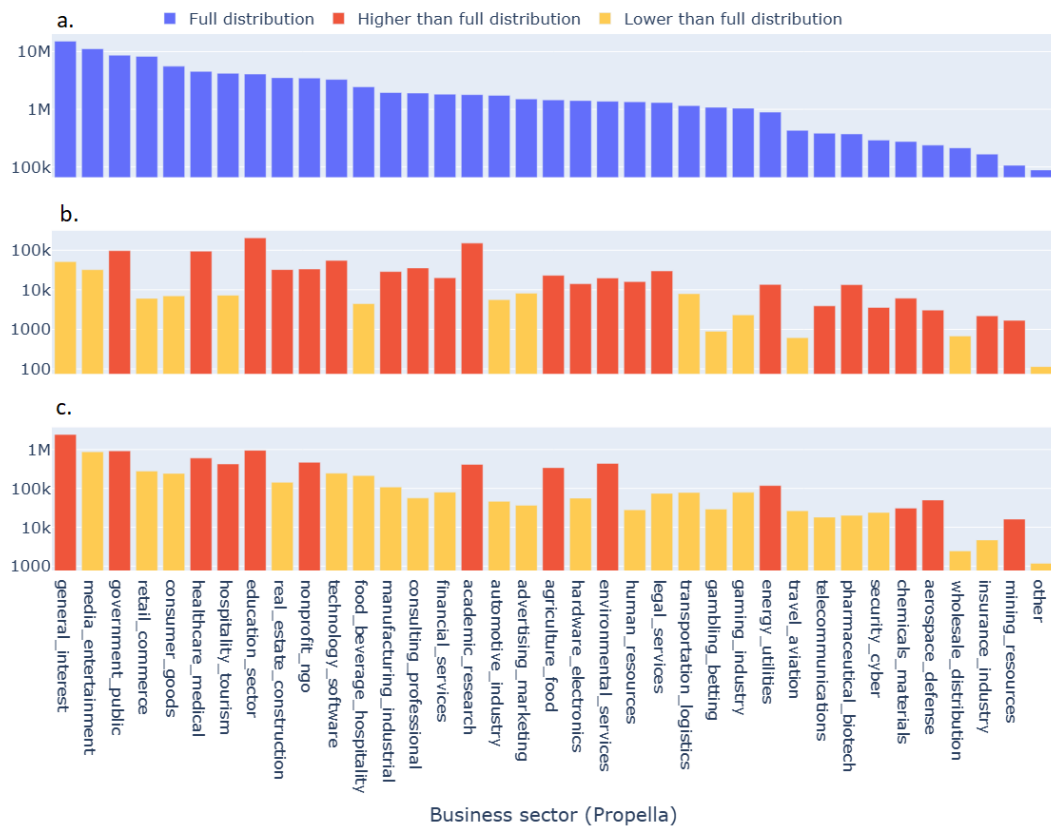


Figure 7: Distributions of business sectors in a) Finnish subset of HPLT 4.0, b) Subset with Propella-1 educational value high, c) Top 10 % of documents with the highest FinePDFs-Edu score

Business sector	none	minimal	basic	moderate	high
general_interest	39.20 %	28.85 %	20.21 %	11.40 %	0.34 %
media_entertainment	47.25 %	38.72 %	10.44 %	3.30 %	0.29 %
government_public	18.72 %	51.38 %	20.63 %	8.15 %	1.12 %
retail_commerce	55.59 %	30.64 %	11.12 %	2.58 %	0.07 %
consumer_goods	50.14 %	29.85 %	14.71 %	5.17 %	0.13 %
healthcare_medical	15.34 %	35.67 %	27.93 %	18.98 %	2.08 %
hospitality_tourism	48.72 %	33.11 %	14.70 %	3.30 %	0.17 %
education_sector	11.71 %	47.26 %	24.46 %	11.55 %	5.03 %
real_estate_construction	43.10 %	32.69 %	15.38 %	7.92 %	0.92 %
nonprofit_ngo	26.09 %	46.91 %	19.29 %	6.75 %	0.96 %
technology_software	17.36 %	38.28 %	23.55 %	19.15 %	1.67 %
food_beverage_hospitality	27.56 %	25.84 %	25.53 %	20.89 %	0.18 %
manufacturing_industrial	32.14 %	40.74 %	17.12 %	8.52 %	1.48 %
consulting_professional	24.95 %	36.12 %	20.08 %	16.99 %	1.86 %
financial_services	19.43 %	39.53 %	25.02 %	14.92 %	1.10 %
academic_research	3.39 %	30.20 %	29.59 %	28.40 %	8.42 %
automotive_industry	40.82 %	31.92 %	17.46 %	9.48 %	0.32 %
advertising_marketing	44.34 %	28.58 %	15.98 %	10.56 %	0.54 %
agriculture_food	15.71 %	38.29 %	27.50 %	16.92 %	1.59 %
hardware_electronics	25.27 %	45.57 %	17.29 %	10.87 %	1.00 %
environmental_services	9.70 %	41.29 %	33.93 %	13.65 %	1.44 %
human_resources	37.55 %	30.93 %	17.82 %	12.51 %	1.19 %
legal_services	12.30 %	40.96 %	25.40 %	19.06 %	2.29 %
transportation_logistics	29.72 %	45.88 %	17.58 %	6.14 %	0.69 %
gambling_betting	50.26 %	31.05 %	13.54 %	5.06 %	0.08 %
gaming_industry	51.68 %	33.49 %	10.10 %	4.51 %	0.22 %
energy_utilities	17.30 %	44.71 %	25.02 %	11.45 %	1.52 %
travel_aviation	38.56 %	41.00 %	15.91 %	4.40 %	0.14 %
telecommunications	20.07 %	49.42 %	19.94 %	9.56 %	1.02 %
pharmaceutical_biotech	20.79 %	24.50 %	25.26 %	25.87 %	3.58 %
security_cyber	12.56 %	32.98 %	27.72 %	25.51 %	1.23 %
chemicals_materials	14.53 %	38.81 %	27.86 %	16.59 %	2.22 %
aerospace_defense	8.00 %	51.12 %	26.46 %	13.14 %	1.29 %
wholesale_distribution	60.66 %	28.20 %	7.47 %	3.35 %	0.32 %
insurance_industry	16.23 %	33.02 %	30.60 %	18.84 %	1.31 %
mining_resources	19.28 %	46.08 %	21.97 %	11.10 %	1.57 %
other	87.17 %	7.46 %	3.54 %	1.69 %	0.13 %

Table 10: Distribution of Propella-1 business sectors categories within Propella-1 educational value categories

Business sector	<0.3771	0.3771–0.8929	0.8929–1.1998	1.1998–2.6491	>2.6491
general_interest	7.89 %	18.22 %	20.02 %	24.13 %	25.62 %
media_entertainment	11.51 %	12.24 %	10.63 %	9.10 %	5.74 %
government_public	5.28 %	11.03 %	11.33 %	9.58 %	5.83 %
retail_commerce	15.27 %	5.28 %	3.37 %	2.89 %	1.65 %
consumer_goods	10.11 %	3.55 %	2.27 %	2.48 %	1.86 %
healthcare_medical	2.48 %	5.37 %	6.62 %	5.99 %	6.51 %
hospitality_tourism	4.62 %	3.86 %	4.24 %	4.46 %	2.03 %
education_sector	1.37 %	4.18 %	6.57 %	8.46 %	21.24 %
real_estate_construction	5.06 %	3.32 %	2.24 %	1.51 %	0.67 %
nonprofit_ngo	2.20 %	3.75 %	5.14 %	4.79 %	3.84 %
technology_software	4.11 %	3.24 %	2.45 %	2.49 %	2.16 %
food_beverage_hospitality	2.13 %	2.58 %	3.13 %	2.30 %	0.69 %
manufacturing_industrial	2.74 %	1.70 %	1.43 %	1.12 %	0.74 %
consulting_professional	2.25 %	2.29 %	1.40 %	0.60 %	0.28 %
financial_services	1.97 %	2.25 %	1.45 %	0.84 %	0.40 %
academic_research	0.35 %	1.78 %	3.54 %	3.99 %	5.76 %
automotive_industry	3.08 %	1.30 %	0.72 %	0.49 %	0.20 %
advertising_marketing	2.50 %	1.30 %	0.64 %	0.39 %	0.13 %
agriculture_food	0.81 %	1.24 %	2.15 %	3.35 %	4.06 %
hardware_electronics	2.62 %	0.90 %	0.52 %	0.57 %	0.47 %
environmental_services	0.39 %	1.00 %	2.47 %	4.23 %	6.20 %
human_resources	1.85 %	1.46 %	0.82 %	0.30 %	0.14 %
legal_services	1.22 %	1.71 %	1.15 %	0.78 %	0.38 %
transportation_logistics	1.17 %	1.30 %	1.14 %	0.83 %	0.34 %
gambling_betting	2.04 %	0.72 %	0.33 %	0.31 %	0.12 %
gaming_industry	1.60 %	0.81 %	0.55 %	0.79 %	0.75 %
energy_utilities	0.59 %	0.92 %	1.33 %	1.21 %	0.93 %
travel_aviation	0.44 %	0.50 %	0.42 %	0.29 %	0.04 %
telecommunications	0.59 %	0.33 %	0.21 %	0.19 %	0.09 %
pharmaceutical_biotech	0.25 %	0.51 %	0.49 %	0.21 %	0.09 %
security_cyber	0.27 %	0.36 %	0.25 %	0.25 %	0.12 %
chemicals_materials	0.28 %	0.25 %	0.32 %	0.31 %	0.31 %
aerospace_defense	0.09 %	0.27 %	0.34 %	0.51 %	0.43 %
wholesale_distribution	0.43 %	0.13 %	0.05 %	0.03 %	0.00 %
insurance_industry	0.21 %	0.20 %	0.10 %	0.05 %	0.01 %
mining_resources	0.07 %	0.10 %	0.16 %	0.16 %	0.17 %
other	0.17 %	0.06 %	0.03 %	0.01 %	0.01 %

Table 11: Distribution of Propella-1 business sectors categories within FinePDFs-Edu score categories