

Register Always Matters: Analysis of LLM Pretraining Data Through the Lens of Language Variation

Amanda Myntti, Erik Henriksson, Veronika Laippala & Sampo Pyysalo

University of Turku

{amanda.a.myntti,erik.henriksson,mavela,spyysalo}@utu.fi

Abstract

Pretraining data curation is a cornerstone in Large Language Model (LLM) development, leading to growing research on quality filtering of large web corpora. From statistical quality flags to LLM-based labelling systems, datasets are divided into categories, frequently reducing to a binary: those passing the filters are deemed as valuable examples, others are discarded as useless or detrimental. However, a more detailed understanding of the contribution of different kinds of texts to model performance is still largely lacking. In this article, we present the first study utilising *registers* or *genres*—a widely used standard in corpus linguistics to model linguistic variation—to curate pretraining datasets and investigate the effect of register on the performance of LLMs. We train small generative models with register classified data and evaluate them using standard benchmarks, and show that the register of pretraining data substantially affects model performance. We uncover surprising relationships between the pretraining material and the resulting models: using the *News* register results in subpar performance, and on the contrary, including the *Opinion* class, covering texts such as reviews and opinion blogs, is highly beneficial. While a model trained on the entire unfiltered dataset outperforms those trained on datasets limited to a single register, combining well-performing registers such as *How-to-Instructions*, *Informational Description*, and *Opinion* leads to major improvements. Furthermore, analysis of individual benchmark results reveals key differences in the strengths and drawbacks of specific register classes as pretraining data: *How-to-Instructions* excels at physical reasoning and sentence completion while barely crossing random baselines on world-knowledge benchmarks, while *Narrative* boosts performance on social interaction tasks but struggles with scientific questions. These findings show that register is an important explainer of model variation and can facilitate more deliberate and detailed future data selection practices.

1 Introduction

Recent work on scaling laws and the limits of human-generated training data available for Large Language Model (LLM) training have caused an uptick in research on pretraining data quality as opposed to quantity (Kaplan et al., 2020; Hoffmann et al., 2022; Muennighoff et al., 2023; Villalobos et al., 2024). While state-of-the-art models are still trained predominantly on large web corpora, there is increasing focus on filtering such data to remove potentially detrimental material. The selection of pretraining data often relies on simple heuristics or quality signals (Longpre et al., 2024; Ostendorff et al., 2024), sometimes together with LLM-created quality flags (Weber et al., 2024; Henriksson et al., 2025). However, language use varies along a number of dimensions other than the presumed dichotomy of positive or negative value for LLM training, ranging from persuasive, narrative, to informative texts. The effects of this variation on LLM capabilities remain unknown. In this paper, we approach pretraining data curation from a novel perspective: how do different registers (or genres) of training data affect LLM performance on commonly used benchmarks.

Register studies is a widely applied paradigm in corpus linguistics that examines language variation across different situations, from casual spoken conversations to persuasive and informational settings (Biber, 1988; 1995; Biber & Conrad, 2019). Registers are defined as situationally characterised text varieties, with typical classes encompassing news, reviews and song lyrics. Register has been shown to be one of the most important predictors of linguistic variation (Biber, 2012)—and, as put by the founder of register studies, Douglas Biber, “register always matters” (Biber, 2013).

Web register classification has long aimed to characterise different kinds of web texts in order to better utilise and understand the web as a corpus (Kilgariff & Grefenstette, 2003; Baroni & Kilgariff, 2006). However, early studies focused on pre-selected sets of registers presumed to be typical of the web, and thus lacked generalizability (Santini, 2007; Sharoff et al., 2010) and subsequently failed to capture the full range of linguistic variation found on the unrestricted web. Egbert et al. (2015) addressed this gap by creating a corpus representing the full spectrum of English web registers, with Laippala et al. (2023) presenting the first deep learning-based web register classifier trained on the data and targeting the entire unrestricted web. Other studies expanded this approach to multilingual settings (Repo et al., 2021; Rönqvist et al., 2021; Kuzman et al., 2023a; Henriksson et al., 2024), showing that web registers are identifiable in web-scale data. These advances now enable the integration of corpus linguistics frameworks with LLM pretraining data curation—our study being to the best of our knowledge the first to bridge these areas, despite the well-documented role of registers in linguistic variation.

To understand the characteristics of web registers in the context of LLM pretraining, we train small generative models up to 100 billion tokens with web data with datasets curated to only contain texts from one register class. We evaluate these models using well-known benchmarks and analyse the results through average accuracy and performance on individual benchmarks, revealing how each register impacts model capabilities. Our results show that register has a substantial impact on model performance. Additionally, we find that training on a combination of register-specific datasets that displayed strengths in different benchmarks can lead to major improvements: registers like *Opinion* boosts model performance, while incorporating examples from register *Interactive Discussion* or *Spoken* at the expense of *How-to-Instructions*, *Informational Description* and *Opinion* leads to performance drops. With these findings, our study aims to open up discussion about linguistic selection of training data, and to show that register classification holds potential as a tool for analysing pretraining data.

2 Related work

Ablation studies, where small generative models are trained with different parts of a large dataset and evaluated on downstream tasks, are a well-established method for evaluating dataset quality and optimising data curation methods. Gao et al. (2020) validated the quality of their Pile datasets by evaluating models trained on the Pile, CommonCrawl, and CC-100. Longpre et al. (2024) pretrain small models on data filtered by different qualities (data age, data domain, data toxicity) and analyse the effects of this curation via model evaluation. Similarly, Sachdeva et al. (2024) use an ablation framework to find that diversity and coverage of different topics in the pretraining data are crucial to models that succeed in multiple benchmarks. Likewise, Burchell et al. (2025) released a multilingual dataset and performed ablation studies to measure the quality of different versions of their datasets. Henriksson et al. (2025) use an LLM to line-by-line classify training documents to optimise dataset composition and evaluate with a downstream task.

The methodology of our approach most closely resembles that of the FineWeb ablation studies (Penedo et al., 2024a). The FineWeb dataset was created by conducting analyses of tens of heuristic filters to select the data curation methods that result in the best pretraining data with respect to the downstream performance of the trained model. Currently, datasets produced by the FineWeb pipeline are available for over a thousand languages (Penedo et al., 2024b).

Current leading work on dataset composition optimisation includes Su et al. (2024) and Li et al. (2024). Lee et al. (2022) established deduplication as a standard procedure in pretraining data selection pipelines. Other common approaches for selecting pretraining data from web sources include URL-based filtering (e.g. Penedo et al. (2023)), text statistics such as word count, repetition, or presence of blacklisted words (e.g. Raffel et al. (2020)), filtering based on perplexity (e.g. Muennighoff et al. (2023), Weber et al. (2024)), and selection based on similarity against a selected good-quality corpus (e.g. Brown et al. (2020), Tirumala et al. (2023)). According to a survey by Albalak et al. (2024), it remains an open question whether quality filtering data improves model performance, and in which cases: Dodge et al. (2021) show that url-based pretraining data filtering leads to bias in the resulting data, while reference corpora used for similarity selection can be biased in language, topics, or demographics (Rae et al., 2022; Gururangan et al., 2022).

While register studies is a widely applied standard to analyse language variation in corpus linguistics, the term *genre* is more often used in the NLP field (Sharoff et al., 2010; Petrenz & Webber, 2011; Sharoff, 2020; Kuzman et al., 2023b). However, both approaches target text categories such as opinionated, informational, and lyrical.

Register studies has a long tradition in mapping the relationship between typical linguistic features associated with texts and different situational contexts where language is used, in both online and offline settings (e.g., Biber (1995); Biber & Egbert (2018); Ansarifar et al. (2025)). Biber (2012) even shows that register is one of the most important predictors of linguistic variation. Similarly, web register (or genre) classification is a widely applied approach to classify web content (Kilgariff & Grefenstette, 2003; Santini, 2007; Petrenz & Webber, 2011; Myntti et al., 2024).

Since the release of the first web register corpus covering the full unrestricted web (Egbert et al., 2015), automatic register classification using machine learning has been successfully applied to the unrestricted web, achieving nearly human-level performance (Laippala et al., 2023; Kuzman et al., 2023b). Henriksson et al. (2024) also show that register characteristics transcend language boundaries, allowing classification models to identify registers even in languages that were not present in the training data (zero-shot languages). These advances enable the current study to select pretraining data by register. Our approach also aligns with calls for greater transparency in foundation models (Bommasani et al., 2023), as it provides structured metadata about the linguistic composition of training data, helping researchers and users to understand what types of texts a model has been exposed to during training.

3 Methods

3.1 Data

We use HPLT version 2.0¹ (Burchell et al., 2025) as our source data. This collection includes datasets for 193 languages and additional parallel datasets for 50 languages paired with English. HPLT v2 datasets have been processed from a combination of Internet Archive and Common Crawl using Trafilatura (Barbaresi, 2021) with language identification performed using OpenLID (Burchell et al., 2023). Two versions of each dataset are provided, *deduplicated*, with MinHash deduplication (Broder et al., 1998), and *cleaned*, with additional cleaning heuristics applied. An evaluation by Burchell et al. (2025) showed the cleaned English dataset to be of high quality, with models trained on the data achieving performance comparable to ones trained on the FineWeb dataset (Penedo et al., 2024a). In this study, we nevertheless chose to use the deduplicated English dataset, which yields slightly worse performance in LLM pretraining. This choice allows us to examine the effect of register with minimal interference from cleaning procedures that might disproportionately affect certain registers and thus bias our analysis. This decision is also motivated by the uneven distribution of register labels in web data; with fewer cleaning steps, we retain more data, which is especially important for the less frequent register classes.

¹Available at <https://hplt-project.org/datasets/v2.0>

Individual registers	Description	Available
<i>How-to-Instructions</i> (HI)	Recipes, instructions	100 B
<i>Interactive Discussion</i> (ID)	Discussion forums	314 B
<i>Informational Description</i> (IN)	Wikis, information sites	695 B
<i>Informational Persuasion</i> (IP)	Advertisements, commercial sites	421 B
<i>Lyrical</i> (LY)	Song lyrics, poems	20 B
<i>Machine Translation</i> (MT)	Machine translation	306 B
<i>Narrative</i> (NA)	News, blogs	545 B
<i>Opinion</i> (OP)	Opinionated texts, religious sermons	416 B
<i>Spoken</i> (SP)	Interviews, speeches	32 B
<i>Instructive-Informational</i> (HI-IN)	Hybrid; Documents predicted as both HI and IN	70B
<i>News</i> (ne)	Subregister of NA; news	404 B
<i>Description</i> (dtp)	Subregister of IN; decription of a thing or a person	781 B
Combinations		
<i>HI-IN-HI-dtp</i>	1/3rd sampled from HI-IN, HI, dtp	–
<i>HI-IN-HI-dtp-OP</i>	1/4th sampled from HI-IN, HI, dtp, OP	–
<i>HI-IN-HI-dtp-OP-NA</i>	1/5th sampled from HI-IN, HI, dtp, OP, NA	–
<i>HI-IN-HI-dtp-OP-NA-ID</i>	1/6th sampled from HI-IN, HI, dtp, OP, NA, ID	–
<i>HI-IN-HI-dtp-OP-NA-ID-SP</i>	1/7th sampled from HI-IN, HI, dtp, OP, NA, ID, SP	–
Baselines		
HPLT v2 deduplicated	Burchell et al. (2025)	–
FineWeb	Penedo et al. (2024a)	–

Table 1: Datasets used in our experiments. Each dataset corresponds to one trained model. The *Individual registers* section describes the main register labels and hybrid & subregisters. The *Combinations* section contains combined register classes, and *Baselines* lists the baseline datasets we use in our experiments. The *Available* column contains the approximate token counts available in the HPLT v2 deduplicated corpus.

The HPLT v2 dataset includes predictions of register classes as probabilities for each document, generated using the multi-label register classifier by [Henriksson et al. \(2024\)](#). This classifier was fine-tuned from XLM-RoBERTa-Large ([Conneau et al., 2020](#)) using the multilingual CORE corpus introduced in the same study. The CORE register scheme is hierarchical, with 9 main registers divided into 25 subregisters that further describe the content of each document. Documents can have none, one, or multiple main registers (referred to as *hybrids*), along with any associated subregister labels. The full register scheme can be found in [Henriksson et al. \(2024\)](#), and example documents are given in Appendix A.

Although the register labels are available for multiple languages, we limit our analysis to English due to the availability of varied and well-established benchmarks. For our experiments, we assign register labels to documents from the given probabilities using a classification threshold of 0.4, optimised for English. From the register classes, we select all main registers, 2 subregisters, and one hybrid for our experiments. We included the subregisters and hybrids for specific reasons: the subregisters *News* and *Description* are among the largest subregister classes, each with over 100 billion tokens available—sufficient data to train models and draw meaningful comparisons with their parent registers *Narrative* and *Informational Description*. Moreover, there has been particular interest in using news articles (e.g. [de la Rosa et al. \(2025\)](#)) and web encyclopedias (i.a. Wikipedia) to train LLMs, which correspond to classes *News* and *Description* in the register scheme. The hybrid class *Instructive-Informational* was included to study the differences between individual registers and hybrids, based on preliminary results showing that register-specific models trained on these classes perform well overall but differ across individual benchmark tasks. The selected register classes and their abbreviations are described with examples in Table 1. Subregister classes not selected for our experiments are present in the data; however, they merged to their main-level class.

Using these labels, we sample 100 billion tokens per register, using all available data for registers with fewer than 100 billion tokens. We limit our analysis to documents over 200

characters and remove documents with extremely high token counts (corresponding to over 300,000 words²) due to hardware constraints. These excluded documents represent less than 0.3% of the total data and primarily belong to the *Informational Description* class.

We also experiment with training on combinations of certain registers, which are listed in Table 1 under “Combinations”. These combination datasets are created by sampling equally from individual register datasets: for example, in the combination dataset *HI-IN-HI-dtp*, 1/3rd of the training samples are from *Instructive-Informational* (HI-IN), 1/3rd from *How-to-Instructions* (HI), and 1/3rd from *Description* (dtp), with our dataset creation step ensuring no duplicates are present in the combined data. See Section 4 for further information and motivation for these datasets.

3.2 Model training

We replicate the training setup of [Penedo et al. \(2024a\)](#), using the same model architecture (Llama), model size (1.71 billion parameters), and other training parameters. In total, the models are trained up to 100 billion tokens. See Appendix B for further information on the training setup. As a baseline, we use a model trained on a random sample of the HPLT v2 deduplicated data. This allows us to compare the effects of using a single register as opposed to using data with the natural distribution of registers in the source dataset. Finally, as our setup mirrors that of [Penedo et al. \(2024a\)](#), we also include results for a model trained on a sample of the FineWeb dataset as a further point of comparison and as a reference for interpreting our findings.

3.3 Evaluation

To maintain comparability with previous work, we also follow [Penedo et al. \(2024a\)](#) in our evaluation. Specifically, we use the LightEval ([Fourrier et al., 2023](#)) evaluation harness with the following benchmarks in a zero-shot setting for all tasks:

- **HellaSwag** ([Zellers et al., 2019](#)) contains logical sentence completion tasks in the form of presenting a paragraph and an incomplete sentence with alternative sentence endings.
- **WinoGrande** ([Sakaguchi et al., 2021](#)) evaluates commonsense reasoning in a binary classification setting using pronoun resolution problems.
- **PIQA** ([Bisk et al., 2020](#)) focuses on physical commonsense reasoning, with examples involving manipulation of physical objects.
- **SIQA** ([Sap et al., 2019](#)) contains tasks on reasoning about social situations.
- **OpenBookQA** ([Mihaylov et al., 2018](#)) focuses on tasks requiring multi-step reasoning with additional context given.
- **ARC Easy and ARC Challenge** ([Clark et al., 2018](#)) are multiple choice benchmarks with science exam questions.
- **CommonsenseQA** ([Talmor et al., 2019](#)) consists of multiple-choice questions requiring prior world knowledge.
- **MMLU** ([Hendrycks et al., 2021a;b](#)) is a multiple-choice question answering dataset covering 57 distinct categories, such as US politics and anatomy.

These benchmarks were chosen by [Penedo et al. \(2024a\)](#) due to their compatibility with small model size, showing reliable, low-variance and monotonic signals even with limited training data, which is crucial for our experiments. For MMLU, we consider all 57 categories as one task and average our results accordingly. This ensures that models reaching high accuracy on MMLU but mediocre results on other benchmarks do not dominate the average results. See Appendix D for further discussion.

The selected benchmarks reflect the most common present-day expectations placed on LLMs: fluency, scientific and real-world knowledge, problem solving, and reasoning. However, this selection inevitably disadvantages certain register models. For example, models trained on the *Lyrical* register are likely to show poor performance, as none of the benchmarks

²Defined as white-space and punctuation separated segments

measure poetic capabilities that would be expected from a model trained on lyrical data. This does not mean that lyrical capabilities are unwanted or unnecessary for LLMs, but rather highlights the importance of benchmark diversity and the significance of selecting benchmarks that measure abilities relevant to specific use cases. In our case, although the benchmark selection disadvantages some registers and their expected performance more than others, using the full range of registers covering all kinds of web texts yields a complete picture of registers’ effects on LLM pretraining.

4 Experiments

Our experiments are divided into two parts: First, we examine the effects of register on model performance by training models for the 9 main-level registers, one hybrid class (*Instructive-Informational*), and two subregister classes (*News* and *Description*), presented in Table 1 under “Individual registers”. These experiments are meant to highlight the differences of registers as pretraining data.

Second, we investigate the effect of further mixing registers by selecting top-performing registers from our first experiment. This allows us to observe how adding or removing specific registers affects overall model performance and whether the strengths of individual registers are preserved in combined models. The selected top-scoring register combinations are presented in Table 1 under “Combinations”. When sampling data for register combinations, we ensure that no duplicate documents are included, even when combining the hybrid class *Instructive-Informational* (HI-IN) with the register *How-to-Instructions* (HI). We chose to use the subregister *Description* (dtp) instead of the main-level register *Informational Description* (IN) because *Informational Description* showed signs of performance degradation in later training steps, and as noted in Section 3.1, we had to exclude some documents from the *Informational Description* dataset. To maintain comparability, we train all models up to 100 billion tokens, although the larger amount of combined data would allow for longer training. Our aim is not to identify optimal register combinations, but rather to explore how register-based data curation might influence model performance on different tasks.

Some of our datasets contain fewer than 100 billion tokens, which necessitates repeating data during training. This raises questions about the validity of our results, as Hoffmann et al. (2022) argue that repeating data can lead to performance degradation. However, Taylor et al. (2022) report validation loss continues to improve up to 4 epochs for a 120 billion parameter model, and Muennighoff et al. (2023) find that for the GPT-2 (Radford et al., 2019) architecture with 2-8 billion parameters, multi-epoch training has negligible effects when limited to 4 or fewer epochs. Xue et al. (2023) further demonstrate that smaller models are less susceptible to the negative effects of repeated training examples in T5-architecture models (Raffel et al., 2020). Therefore, we are reasonably confident in the results from models trained on *Instructive-Informational* (1.4 epochs) and *Spoken* (3.1 epochs). The model trained on our smallest dataset, *Lyrical* (5 epochs), performs poorly even within the first epochs, suggesting its low performance is due to register characteristics and benchmark selection rather than data repetition. None of the combination models exceed 1 epoch of training, since we sample the 100 billion tokens from combinations of multiple datasets, together exceeding 100 billion tokens.

We additionally compare our register scheme to three other quality classifiers, FineWeb-edu classifier (Penedo et al., 2024a), the DCLM quality classifier (Li et al., 2024), and NVIDIA’s NemoCurator Quality Classifier DeBERTa (He et al., 2023). These experiments and their results are presented in Appendix C.

5 Results

5.1 Average performance

The results for individual register models, averaged over all benchmarks, are presented in Figure 1 and Table 2, and detailed in Appendix D. The evaluation shows striking differences in accuracy between the models, demonstrating that the register of the pretraining data

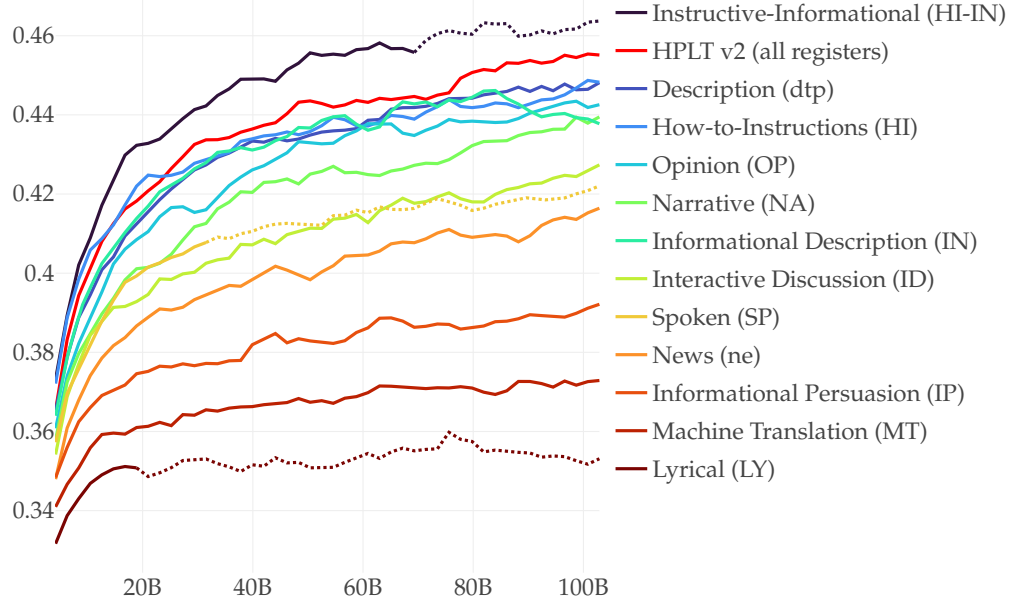


Figure 1: Individual register and HPLT v2 (deduplicated) model accuracies as a function of the number of training tokens. Rolling average over 6 billion tokens (3 adjacent checkpoints) applied for ease of reading overlapping lines, see Table 2 for numerical results for the final checkpoint. Dotted lines indicate training continuing over one epoch, legend in order of last checkpoint performance.

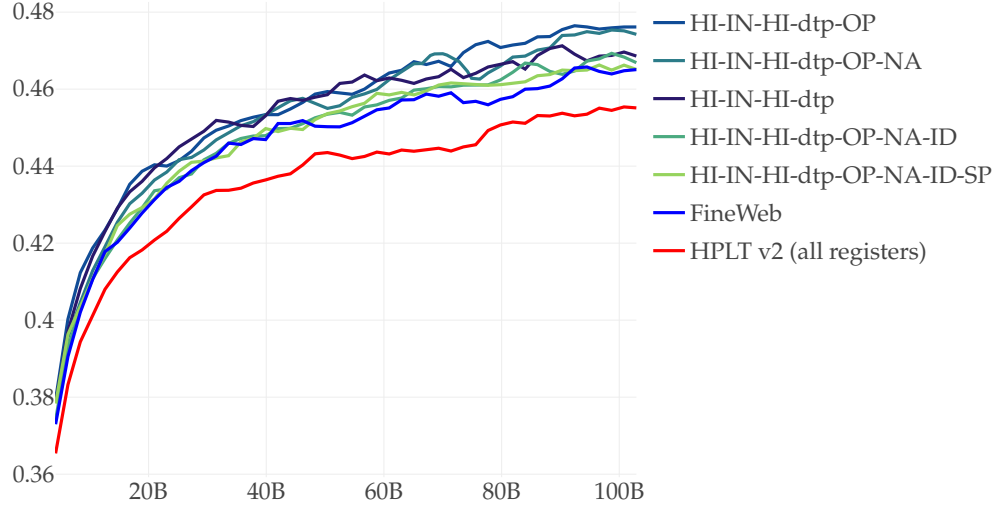


Figure 2: Combination model and baseline accuracies as a function of the number of training tokens. Rolling average over 6 billion tokens (3 adjacent checkpoints) applied for ease of reading overlapping lines, see Table 2 for numerical results for the final checkpoint. Legend in order of last checkpoint performance.

Pretraining dataset	Accuracy	Pretraining dataset	Accuracy
<i>HI-IN-HI-dtp-OP</i>	0.475	<i>Opinion</i> (OP)	0.447
<i>HI-IN-HI-dtp-OP-NA</i>	0.472	<i>Narrative</i> (NA)	0.441
<i>HI-IN-HI-dtp</i>	0.466	<i>Informational Description</i> (IN)	0.437
FineWeb	0.466	<i>Interactive Discussion</i> (ID)	0.431
<i>HI-IN-HI-dtp-OP-NA-ID-SP</i>	0.465	<i>Spoken</i> (SP)	0.422
<i>Instructive-Informational</i> (HI-IN)	0.465	<i>News</i> (ne)	0.418
<i>HI-IN-HI-dtp-OP-NA-ID</i>	0.464	<i>Informational Persuasion</i> (IP)	0.393
HPLT v2 (deduplicated, all registers)	0.457	<i>Machine Translation</i> (MT)	0.374
<i>Description</i> (dtp)	0.452	<i>Lyrical</i> (LY)	0.358
<i>How-to-Instructions</i> (HI)	0.447		

Table 2: Results by the last checkpoint performance in numerical format. Note that the values differ slightly from the Figures 1 and 2 due averaging over adjacent checkpoints in the figures.

has a substantial effect on model performance. We find that the model trained on the full HPLT v2 data outperforms models trained on individual registers, demonstrating that a mix of registers is important for LLM training. The performance ranking of individual registers generally aligns with intuitions about their content and what is typically considered high-quality pretraining data: registers like *Lyrical* (LY), *Machine Translation* (MT), and *Informational Persuasion* (IP) yield worse-performing models, while models trained on registers containing informational content and instructions, such as *Description* (dtp) and *How-to-Instructions* (HI), perform the best. Surprisingly, *Opinion* (OP) yields the 4th best performing model. Both *Informational Description* (IN) and *Narrative* (NA) models perform worse than *How-to-Instructions*, *Description* and *Opinion*. Following these, with a noticeable performance drop, are the models trained on *Interactive Discussion* (ID), *Spoken* (SP) and remarkably, the subregister *News* (ne), consisting of news articles, which performs among the worst.

As opposed to the main register models, the hybrid class model *Instructive-Informational* (HI-IN) outperforms the model trained on the full HPLT v2 data. This shows that hybrid documents do not simply result in averaged performance between the two hybridised registers, but can show better performance than either alone. This finding motivated our second experiment, investigating register combinations. The subregisters *News* (ne) (subregister of *Narrative*) and *Description* (dtp) (subregister of *Informational Description*) show divergent patterns when compared to their parent registers: *News* drastically underperforms its main-level register *Narrative*, while *Description* achieves higher accuracy than *Informational Description*. We analyse differences between main-subregister pairs in more detail in the following section.

As the results show, pretraining a model with data from only one register class does not provide improvements compared to training with the full dataset. This aligns with the established notion that pretraining data should have variability to yield capable models. The results for our combination model experiment are shown in Figure 2, Table 2, and again with more detail in Appendix D. The results clearly support the above idea: all combination models outperform the individual register models and the baseline HPLT v2 model. Additionally, the best models in our study (*HI-IN-HI-dtp-OP* and *HI-IN-HI-dtp-OP-NA*) reach higher average accuracy than the models trained on the HPLT v2 and FineWeb datasets. *HI-IN-HI-dtp* model shows much worse performance than *HI-IN-HI-dtp-OP*, which further includes *Opinion* (OP). This suggests that opinionated texts provide valuable pretraining data and are key to high model performance, at least as measured by the selected benchmarks. Further adding the *Narrative* (NA) dataset, as seen in the *HI-IN-HI-dtp-OP-NA* model, slightly worsens performance, though the difference is small. Significant drops occur when including *Interactive Discussion* (ID) and *Spoken* (SP) in models *HI-IN-HI-dtp-OP-NA-ID* and *HI-IN-HI-dtp-OP-NA-ID-SP*, indicating either that these registers contain qualities that harm the abilities measured on these benchmarks, or that they do not provide benefits that outweigh the reduced proportion of *Instructive-Informational* (HI-IN), *Description* (dtp), and *Opinion* (OP) in the training mix.

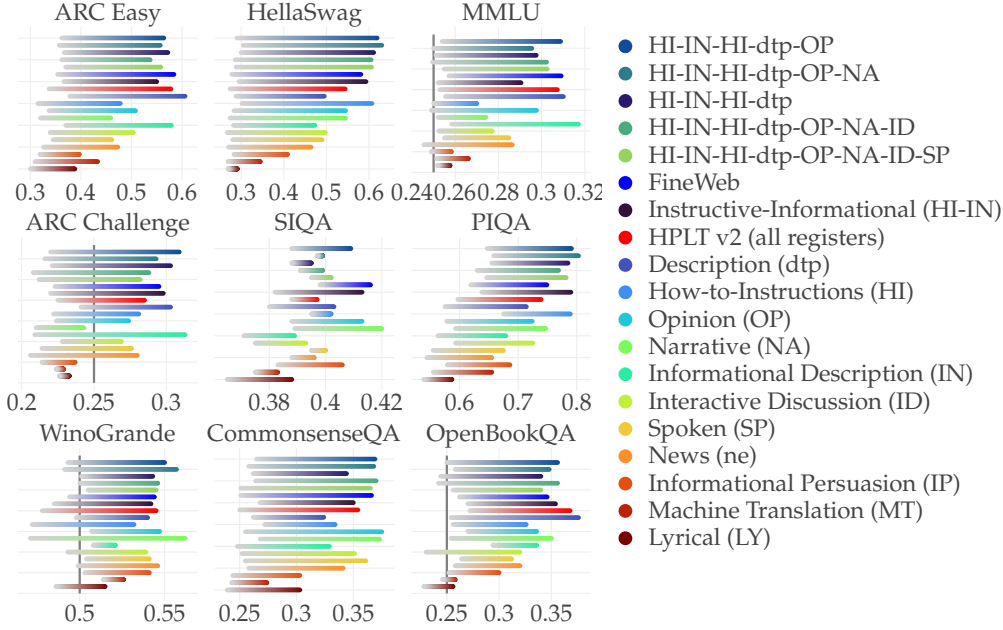


Figure 3: Change of accuracy from first to final checkpoint on individual benchmarks shown as a range, with grey indicating the first checkpoint and colours indicating the last checkpoint. The random-guess threshold is shown as a grey vertical line in cases where at least one model falls below it. Bars and legend shown in order of average accuracy.

5.2 Per task

The results for each benchmark separately are shown in Figure 3 and detailed further in Appendix E. The differences between benchmarks are remarkable, both in the performance ranking of models and in the way accuracy trends with respect to the number of training tokens. This shows that registers have strengths in different tasks. Most benchmarks display the monotonicity condition that influenced the selection of Penedo et al. (2024a), as they tend towards increased accuracy with more training tokens. A clear outlier is the SIQA benchmark, where increases in accuracy are very small, with the largest increase seen in *Narrative* (NA) and *Instructive-Informational* (HI-IN) models, both increasing approximately by 0.03 during the training.

Generally, the trends observed in Section 5.1 hold. The models with the lowest average accuracy—*Lyrical* (LY), *Machine Translation* (MT) and *Informational Persuasion* (IP)—perform poorly on all benchmarks, and, notably, worse than a random guess on ARC Challenge. The model for subregister *News* (ne), despite weak average performance, improves on the MMLU and WinoGrande benchmarks, while *Spoken* (SP) enhances results on CommonsenseQA. The differences between benchmarks become more pronounced with models trained on *Narrative* (NA), *Informational Description* (IN), *Opinion* (OP), *How-to-Instructions* (HI), and *Interactive Discussion* (ID), which display great variation in ranking across tasks, highlighting their strengths. The *Narrative* model performs well on SIQA and WinoGrande but barely surpasses the random baseline on ARC Challenge. *Informational Description* trumps other models on MMLU and both ARC benchmarks, but ranks low to mediocre elsewhere. The model trained on the *How-to-Instructions* dataset outperforms almost all other models on PIQA and HellaSwag. This might be because PIQA measures reasoning in physical tasks, which aligns with the instructions related to physics included in *How-to-Instructions*. On the other hand, *How-to-Instructions* ranks among the worst on MMLU and CommonsenseQA. The *Opinion* model excels in WinoGrande, SIQA and CommonsenseQA. This signal also appears in the combination models’ performance on CommonsenseQA: including *Opinion* in the training data greatly increases accuracy on this benchmark, unlike the *HI-IN-HI-dtp* model, which excludes *Opinion*.

The hybrid *Instructive-Informational* (HI-IN) model generally performs between individual *How-to-Instructions* (HI) and *Informational Description* (IN) models, and is notably closer in performance to whichever yields better accuracy. Exceptions to this pattern are OpenBookQA, CommonsenseQA, and WinoGrande, where the *Instructive-Informational* model shows better accuracy than either *How-to-Instructions* or *Informational Description* separately. Specifically in the case of WinoGrande, this is surprising, as both *How-to-Instructions* and *Informational Description* struggle with this benchmark. From this result, we conclude that hybrids may contain qualities not found in individual registers and require further study. The *News* (ne) model trails behind *Narrative* (NA), except on the MMLU and ARC Challenge benchmarks. This is likely due to news articles covering topics related to MMLU tasks, such as politics, business, science, and the environment. Similarly, the other main-subregister pair, *Description* (dtp) and *Informational Description* (IN), shows differences in performance across benchmarks, with the *Description*-model outshining *Informational Description* in ARC Easy, HellaSwag, PIQA, WinoGrande, and OpenbookQA. *Informational Description*, on the other hand, excels in ARC Challenge and MMLU. Specifically for the WinoGrande benchmark, the performance difference might be explained by the *Description* register containing more texts focused on things and people, thus resulting in better pronoun-resolution abilities.

The combination models lead in performance on most benchmarks. On MMLU specifically, we observe a huge drop between the *HI-IN-HI-dtp-OP* model and others: including *Opinion* in the pretraining data increases performance, while removing *Opinion* or further adding *Narrative* greatly disrupts model performance on this benchmark. Notably, the models *HI-IN-HI-dtp-OP* and *HI-IN-HI-dtp-OP-NA* are trained with registers that individually perform well on a wide range of benchmarks: *Instructive-Informational* on HellaSwag and ARC Easy, *How-to-Instructions* on HellaSwag and PIQA, *Description* on ARC Easy, ARC Challenge, and OpenBookQA, *Opinion* on CommonsenseQA, and *Narrative* on SIQA and WinoGrande.

6 Conclusion

In this paper, we presented the first study investigating the effect of linguistic register on the performance of small generative LLMs. Our findings show that register is an important explainer of LLM performance, and we were able to show surprising relationships between the register of pretraining data and model accuracy on different benchmarks. While using the whole dataset yielded better performance than limiting the pretraining data to any individual register class, we found specific combinations of registers that outperform the full uncuration dataset. These findings can be used to increase pretraining data transparency and to optimise data selection methods, potentially addressing specific performance gaps by incorporating appropriate registers.

Limitations and future work

As a first study in the intersection of two broad lines of research – the linguistic registers and LLM training – our work leaves a number of questions incompletely answered. Our experiments are conducted in English only. While register classification has shown multilingual prowess, our results might not generalise to other languages, and we leave multilingual analysis for future work. The benchmarks used, though standard in studies of this type, do not fully measure all model capabilities, as discussed regarding the *Lyrical* class. Evaluating with more diverse benchmarks would provide more comprehensive results and might reveal strengths and drawbacks we were unable to uncover.

Bias from data curation methods is known to propagate to trained models (Rae et al., 2022; Gururangan et al., 2022), and register classification is not exempt from this issue. To mitigate bias, we used the *deduplicated* version of the HPLT v2 data. Although this version has been processed with Trafilatúra, Penedo et al. (2024a) note that Trafilatúra can leave “undesirable” content in the dataset, potentially affecting some registers more than others. The possibility of benchmark contamination also exists; register classification could systematically assign leaked benchmark material to specific classes, potentially explaining some performance differences.

Acknowledgments

This project has received funding from the *Finnish Doctoral Program Network in Artificial Intelligence*, AI-DOC under decision number VN/3137/2024-OKM-6, the European Union’s *Horizon Europe research and innovation programme* under grant agreement No. 101070350, the *Digital Europe Programme* under grant agreement No. 101195233, and the Research Council of Finland under grant No. 362459. The contents of this publication are the sole responsibility of its authors and do not necessarily reflect the opinion of the European Union. We wish to acknowledge CSC - IT Center for Science, Finland, for computational resources.

Ethics Statement

Training large language models is resource-demanding. This fact guided our decisions: we used small model sizes and trained only up to the point where clear conclusions can be drawn. In total, we used approximately 11,000 GPU hours per model. See Appendix B for calculation of the total computational cost. To further mitigate the environmental impact of training, we used LUMI supercomputer, which is among the most environmentally friendly supercomputers in the world³. The models trained for this study are planned to be used in future research.

Releasing our models publicly for reproducibility opens up the possibility of our models being used in contexts other than research. Although we do not evaluate the possible toxicity of the generations they produce, the model sizes are small, and they are trained only up to 100 billion tokens, which makes their generative capabilities weak and thus likely unsuitable for applications outside the scope of this study.

Reproducibility statement

We have prioritised reproducibility in our work and make all artefacts created as part of this study, including code, model training configurations, and the trained model checkpoints, available on <https://huggingface.co/TurkuNLP/register-ablations> under open licenses. Our work only makes use of data openly available from <https://hplt-project.org/datasets/v2.0>, and our model training setup is detailed in Appendix B.

References

- Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, Colin Raffel, Shiyu Chang, Tatsunori Hashimoto, and William Yang Wang. A survey on data selection for language models, 2024. URL <https://arxiv.org/abs/2402.16827>.
- Ahmad Ansarifar, Hesamoddin Shahriari, Shelley Staples, and Mohammad Ghazanfari. Linguistic variation in two written academic sub-registers. *Revista Española de Lingüística Aplicada/Spanish Journal of Applied Linguistics*, 38(1):162–191, 2025. ISSN 0213-2028. doi: <https://doi.org/10.1075/resla.22052.ans>. URL <https://www.jbe-platform.com/content/journals/10.1075/resla.22052.ans>.
- Adrien Barbaresi. Trafilatura: A web scraping library and command-line tool for text discovery and extraction. In Heng Ji, Jong C. Park, and Rui Xia (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pp. 122–131, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-demo.15. URL <https://aclanthology.org/2021.acl-demo.15/>.
- Marco Baroni and Adam Kilgarriff. Large linguistically-processed web corpora for multiple languages. In *Demonstrations*, pp. 87–90, 2006. URL <https://aclanthology.org/E06-2001/>.

³<https://lumi-supercomputer.eu/sustainable-future/>

- Douglas Biber. *Variation across Speech and Writing*. Cambridge University Press, 1988.
- Douglas Biber. *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge University Press, Cambridge, 1995.
- Douglas Biber. Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory*, 8(1):9–37, 2012. doi: doi:10.1515/cllt-2012-0002. URL <https://doi.org/10.1515/cllt-2012-0002>.
- Douglas Biber. Interview with Douglas Biber. *Journal of English Linguistics*, 41(4): 359–379, 2013. doi: 10.1177/0075424213502237. URL <https://doi.org/10.1177/0075424213502237>.
- Douglas Biber and Susan Conrad. *Register, Genre, and Style*. Cambridge University Press, Cambridge, 2019.
- Douglas Biber and Jesse Egbert. *Register Variation Online*. Cambridge University Press, 2018.
- Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020.
- Rishi Bommasani, Kevin Klyman, Shayne Longpre, Sayash Kapoor, Nestor Maslej, Betty Xiong, Daniel Zhang, and Percy Liang. The foundation model transparency index, 2023. URL <https://arxiv.org/abs/2310.12941>.
- Andrei Z. Broder, Moses Charikar, Alan M. Frieze, and Michael Mitzenmacher. Min-wise independent permutations (extended abstract). In *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*, STOC '98, pp. 327–336, New York, NY, USA, 1998. Association for Computing Machinery. ISBN 0897919629. doi: 10.1145/276698.276781. URL <https://doi.org/10.1145/276698.276781>.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.
- Laurie Burchell, Alexandra Birch, Nikolay Bogoychev, and Kenneth Heafield. An open dataset and model for language identification. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 865–879, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.75. URL <https://aclanthology.org/2023.acl-short.75/>.
- Laurie Burchell, Ona de Gibert, Nikolay Arefyev, Mikko Aulamo, Marta Bañón, Pinzhen Chen, Mariia Fedorova, Liane Guillou, Barry Haddow, Jan Hajič, Jindřich Helcl, Erik Henriksson, Mateusz Klimaszewski, Ville Komulainen, Andrey Kutuzov, Joona Kytöniemi, Veronika Laippala, Petter Mæhlum, Bhavitvya Malik, Farrokh Mehryary, Vladislav Mikhailov, Nikita Moghe, Amanda Myntti, Dayyán O'Brien, Stephan Oepen, Proyag Pal, Jousia Piha, Sampo Pyysalo, Gema Ramírez-Sánchez, David Samuel, Pavel Stepachev, Jörg Tiedemann, Dušan Variš, Tereza Vojtěchová, and Jaume Zaragoza-Bernabeu. An expanded massive multilingual dataset for high-performance language technologies, 2025. URL <https://arxiv.org/abs/2503.10267>.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018. URL <https://arxiv.org/abs/1803.05457>.

- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL <https://aclanthology.org/2020.acl-main.747/>.
- Javier de la Rosa, Vladislav Mikhailov, Lemei Zhang, Freddy Wetjen, David Samuel, Peng Liu, Rolv-Arild Braaten, Petter Mæhlum, Magnus Breder Birkenes, Andrey Kutuzov, Tita Enstad, Hans Christian Farsethås, Svein Arne Brygfeld, Jon Atle Gulla, Stephan Oepen, Erik Velldal, Wilfred Østgulen, Lilja Øvrelid, and Aslak Sira Myhre. The impact of copyrighted material on large language models: A norwegian perspective, 2025. URL <https://arxiv.org/abs/2412.09460>.
- Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1286–1305, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.98. URL <https://aclanthology.org/2021.emnlp-main.98/>.
- Jesse Egbert, Douglas Biber, and Mark Davies. Developing a user-based method of web register classification. *Journal of the Association for Information Science and Technology*, 66, 02 2015. doi: 10.1002/asi.23308.
- Clémentine Fourrier, Nathan Habib, Thomas Wolf, and Lewis Tunstall. Lighteval: A lightweight framework for llm evaluation, 2023. URL <https://github.com/huggingface/lighteval>.
- Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.
- Suchin Gururangan, Dallas Card, Sarah Dreier, Emily Gade, Leroy Wang, Zeyu Wang, Luke Zettlemoyer, and Noah A. Smith. Whose language counts as high quality? measuring language ideologies in text data selection. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2562–2580, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.165. URL <https://aclanthology.org/2022.emnlp-main.165/>.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, 2023. URL <https://arxiv.org/abs/2111.09543>.
- Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021a.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021b.
- Erik Henriksson, Amanda Mynntti, Saara Hellström, Anni Eskelinen, Selcen Erten-Johansson, and Veronika Laippala. Automatic register identification for the open web using multilingual deep learning, 2024. URL <https://arxiv.org/abs/2406.19892>.
- Erik Henriksson, Otto Tarkka, and Filip Ginter. Finerweb-10bt: Refining web data with llm-based line-level filtering, 2025. URL <https://arxiv.org/abs/2501.07314>.

- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack W. Rae, and Laurent Sifre. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22*, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. URL <https://arxiv.org/abs/2001.08361>.
- Adam Kilgarrieff and Gregory Grefenstette. Introduction to the special issue on the web as corpus. *Computational Linguistics*, 29(3):333–348, 2003. doi: 10.1162/089120103322711569. URL <https://aclanthology.org/J03-3001/>.
- Taja Kuzman, Igor Mozetič, and Nikola Ljubešić. Automatic genre identification for robust enrichment of massive text collections: Investigation of classification methods in the era of large language models. *Machine Learning and Knowledge Extraction*, 5(3):1149–1175, Sep 2023a. ISSN 2504-4990. doi: 10.3390/make5030059. URL <http://dx.doi.org/10.3390/make5030059>.
- Taja Kuzman, Peter Rupnik, and Nikola Ljubešić. Get to know your parallel data: Performing English variety and genre classification over MaCoCu corpora. In Yves Scherrer, Tommi Jauhiainen, Nikola Ljubešić, Preslav Nakov, Jörg Tiedemann, and Marcos Zampieri (eds.), *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pp. 91–103, Dubrovnik, Croatia, May 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.vardial-1.9. URL <https://aclanthology.org/2023.vardial-1.9/>.
- Veronika Laippala, Samuel Rönqvist, Miika Oinonen, Aki Juhani Kyröläinen, Anna Salmela, Douglas Biber, Jesse Egbert, and Sampo Pyysalo. Register identification from the unrestricted open web using the corpus of online registers of english. *Language Resources and Evaluation*, 57(3):1045–1079, September 2023. ISSN 1574-020X. doi: 10.1007/s10579-022-09624-1. 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).
- Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8424–8445, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.577. URL <https://aclanthology.org/2022.acl-long.577/>.
- 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, Yonatan Bitton, Marianna Nezhurina, Amro Abbas, Cheng-Yu Hsieh, Dhruva Ghosh, Josh Gardner, Maciej Kilian, Hanlin Zhang, Rulin Shao, Sarah Pratt, Sunny Sanyal, Gabriel Ilharco, Giannis Daras, Kalyani Marathe, Aaron Gokaslan, Jieyu Zhang, Khyathi Chandu, Thao Nguyen, Igor Vasiljevic, Sham Kakade, Shuran Song, Sujay Sanghavi, Fartash Faghri, Sewoong Oh, Luke Zettlemoyer, Kyle Lo, Alaaeldin El-Nouby, Hadi Pouransari, Alexander Toshev, Stephanie Wang, Dirk Groeneveld, Luca Soldaini, Pang Wei Koh, Jenia Jitsev, Thomas Kollar, Alexandros G. Dimakis, Yair Carmon, Achal Dave, Ludwig Schmidt, and Vaishaal Shankar. Datacomp-lm: In search of the next generation of training sets for language models, 2024. URL <https://arxiv.org/abs/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. A pre-trainer’s guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3245–3276, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.179. URL <https://aclanthology.org/2024.naacl-long.179/>.
- Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *EMNLP*, 2018.
- Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. Scaling data-constrained language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS ’23, Red Hook, NY, USA, 2023. Curran Associates Inc.
- Amanda Myntti, Liina Repo, Elian Freyermuth, Antti Kanner, Veronika Laippala, and Erik Henriksson. Intersecting register and genre: Understanding the contents of web-crawled corpora. In Mika Hämäläinen, Emily Öhman, So Miyagawa, Khalid Alnajjar, and Yuri Bizzoni (eds.), *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pp. 386–397, Miami, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.nlp4dh-1.38. URL <https://aclanthology.org/2024.nlp4dh-1.38/>.
- Malte Ostendorff, Pedro Ortiz Suarez, Lucas Fonseca Lage, and Georg Rehm. LLM-datasets: An open framework for pretraining datasets of large language models. In *First Conference on Language Modeling*, 2024. URL <https://openreview.net/forum?id=5RdIMlGLXL>.
- Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon llm: outperforming curated corpora with web data only. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS ’23, Red Hook, NY, USA, 2023. Curran Associates Inc.
- Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The fineweb datasets: Decanting the web for the finest text data at scale. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 30811–30849. Curran Associates, Inc., 2024a. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/370df50ccfd8bde18f8f9c2d9151bda-Paper-Datasets_and_Benchmarks_Track.pdf.
- Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Martin Jaggi, Leandro von Werra, and Thomas Wolf. Fineweb2: A sparkling update with 1000s of languages, December 2024b. URL <https://huggingface.co/datasets/HuggingFaceFW/fineweb-2>.
- Philipp Petrenz and Bonnie Webber. Squibs: Stable classification of text genres. *Computational Linguistics*, 37(2):385–393, June 2011. doi: 10.1162/COLLa.00052. URL <https://aclanthology.org/J11-2004/>.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI*, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf. Accessed: 2024-11-15.
- 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, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia

- Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d’Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. Scaling language models: Methods, analysis & insights from training gopher, 2022. URL <https://arxiv.org/abs/2112.11446>.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1), January 2020. ISSN 1532-4435.
- Liina Repo, Valtteri Skantsi, Samuel Rönqvist, Saara Hellström, Miika Oinonen, Anna Salmela, Douglas Biber, Jesse Egbert, Sampo Pyysalo, and Veronika Laippala. Beyond the english web: Zero-shot cross-lingual and lightweight monolingual classification of registers. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, 2021. URL <https://aclanthology.org/2021.eacl-srw.24.pdf>.
- Samuel Rönqvist, Valtteri Skantsi, Miika Oinonen, and Veronika Laippala. Multilingual and zero-shot is closing in on monolingual web register classification. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pp. 157–165, Reykjavik, Iceland (Online), 2021. Linköping University Electronic Press, Sweden. URL <https://aclanthology.org/2021.nodalida-main.16>.
- Noveen Sachdeva, Benjamin Coleman, Wang-Cheng Kang, Jianmo Ni, Lichan Hong, Ed H. Chi, James Caverlee, Julian McAuley, and Derek Zhiyuan Cheng. How to train data-efficient llms, 2024. URL <https://arxiv.org/abs/2402.09668>.
- Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: an adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106, August 2021. ISSN 0001-0782. doi: 10.1145/3474381. URL <https://doi.org/10.1145/3474381>.
- Marina Santini. *Automatic identification of genre in web pages*. PhD thesis, University of Brighton, 2007. Unpublished doctoral dissertation.
- Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQa: Commonsense reasoning about social interactions. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4463–4473, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1454. URL <https://aclanthology.org/D19-1454/>.
- Serge Sharoff. Know thy corpus! robust methods for digital curation of web corpora. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 2453–2460, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL <https://aclanthology.org/2020.lrec-1.298/>.
- Serge Sharoff, Zhili Wu, and Katja Markert. The web library of babel: evaluating genre collections. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias (eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*, Valletta,

- Malta, May 2010. European Language Resources Association (ELRA). URL <https://aclanthology.org/L10-1011/>.
- Dan Su, Kezhi Kong, Ying Lin, Joseph Jennings, Brandon Norick, Markus Kliegl, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. Nemotron-cc: Transforming common crawl into a refined long-horizon pretraining dataset, 2024. URL <https://arxiv.org/abs/2412.02595>.
- Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In Jill Burstein, Christy Doran, and Tamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421. URL <https://aclanthology.org/N19-1421/>.
- Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science, 2022. URL <https://arxiv.org/abs/2211.09085>.
- Kushal Tirumala, Daniel Simig, Armen Aghajanyan, and Ari Morcos. D4: Improving llm pretraining via document de-duplication and diversification. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 53983–53995. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a8f8cbd7f7a5fb2c837e578c75e5b615-Paper-Datasets_and_Benchmarks.pdf.
- Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. Position: will we run out of data? limits of llm scaling based on human-generated data. In *Proceedings of the 41st International Conference on Machine Learning, ICML’24. JMLR.org*, 2024.
- Maurice Weber, Daniel Y Fu, Quentin Gregory Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, Ben Athiwaratkun, Rahul Chalamala, Kezhen Chen, Max Ryabinin, Tri Dao, Percy Liang, Christopher Re, Irina Rish, and Ce Zhang. Redpajama: an open dataset for training large language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL <https://openreview.net/forum?id=lnuXaRpwv>.
- Fuzhao Xue, Yao Fu, Wangchunshu Zhou, Zangwei Zheng, and Yang You. To repeat or not to repeat: Insights from scaling llm under token-crisis. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 59304–59322. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/b9e472cd579c83e2f6aa3459f46aac28-Paper-Conference.pdf.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

A Examples of texts in register classes

Below we show examples of texts classified as selected registers. These texts were chosen randomly but with a focus towards brevity (with some texts additionally truncated), general topics, and non-toxic language use. We also omit new-line characters.

How-to-Instructions (HI)

Home & DIY How To Cut Your Own Hair At Home With countless non-essential businesses closing due to the government-implemented lockdowns, many people have been forced to be creative and resourceful in order to get by. Case in point, not everyone of us are lucky enough to be quarantined at home with a professional hairdresser. So we do the best we can – which often means cutting our own hair. [...]

Interactive Discussion (ID)

I tried last night to compile a standalone which I've compiled approximately 1,000,000 times before, and have been met with a long series of very lengthy crash reports. As far as I can tell, something in the compiler is trying to install every framework on my computer, and I sometimes wind up with 1 gig+ .mxr files. Naturally, little dinky tests compile fine, and I can compile all of the modules that make up my patch, but not the whole thing at once. I did a super-clean Max reinstall, no third-party objects aside a couple of Lobjects which are up to date, and I'm stumped. The only difference I can think of since I compiled this a few weeks ago is an upgrade to 4.5. Anyone have any light to throw on this? Im happy to send crashlogs, but they're big. 10.5.2, Max 4.6, Jitter 1.6.3 thanks, M

Machine Translation (MT)

Incredible Interior Design Sketches Good Looking Interior Designer Remarkable Interior Design Exceptional Interior Design Sketches Cozy Interior Design Sketches Chic Interior Design Sketches Cool Interior Design Sketches

News (ne)

Wal-Mart has signed a storm water settlement agreement with the Connecticut Department of Environmental Protection concerning alleged violations at 20 Wal-Mart stores and two SAM'S CLUB locations in the state. Under the agreement, Wal-Mart will pay \$600,000 in civil penalties for violations alleged to have taken place between 1996 and 2003. Wal-Mart also will contribute \$550,000 to two different supplemental environmental projects—\$500,000 to assist municipalities in addressing storm water issues, and \$50,000 for environmental projects in the Connecticut River Watershed. [...]

Opinion (OP)

Once again a very nice stop for the night. Staff has been very well trained, all are friendly and helpful with one exception/the young man who checked us in. Breakfast was so appreciated after 12 nights at a Holiday Inn Express which was ok but does not have your choices. Lynne does an especially good job/appreciated all her help. They... More

B Model training setup

Models were trained following the setup of [Penedo et al. \(2024a\)](#). We used the GPT-2 tokeniser, Llama architecture and the same training settings: 1.71B parameters (1.82B with embeddings), sequence length of 2048 tokens, and a global batch size of ~2 million tokens.

Models were trained for 50,000 iterations, which amounts to 100 billion tokens. Training was done on LUMI supercomputer on 16 nodes, each with a single 64-core CPU and 4 MI250x GPUs with dual-GCD (graphics compute die). We used NVIDIA’s Megatron-LM⁴ training framework instead of HuggingFace’s nanotron⁵ framework used by Penedo et al. (2024a).

Training each model took approximately 84 hours, amounting to 10,752 GPU hours per model with the node setup described above, with average performance of 25 teraflops per second. Evaluation was carried out on a single GPU, and with 19 models, each with 50 checkpoints requiring approximately 20 minutes to evaluate, this added 300 GPU hours to the total computational cost. Other processing tasks, such as sampling and tokenisation, required no GPU resources and used a comparatively negligible amount of CPU hours.

C Comparison to other data curation schemes

As discussed in Section 2, many different training data selection schemes and tools have been introduced in the literature. To quantify the connection and possible similarities between our register scheme and three other publicly available quality classification tools, FineWeb-edu classifier (Penedo et al., 2024a), the DCLM quality classifier (Li et al., 2024), and NVIDIA’s NemoCurator Quality Classifier DeBERTa (He et al., 2023), we classify a sample of over 50 000 documents from our pretraining material with all four classifiers. We compare the selected tool to our register classification scheme by conducting a χ^2 -test to evaluate dependence and also report metrics pertaining to the possible overlap.

The FineWeb-edu classifier assigns documents an ordinal label ranging from 0 to 5, the NemoCurator a label 0, 1, or 2, and the DCLM classifier uses a binary label (0,1). For registers, we use the classes in Table 1 under “Individual registers”. Obtained results from a χ^2 -test (considering FineWeb-edu and NemoCurator’s labels nominal) show that all quality classification tools are dependent on the register classification results with $p < 0.01$. This dependence between the classification schemes is to be expected, as all schemes are used as a proxy for “quality” in LLM pretraining. To evaluate the strength of this connection, we calculated Cramer’s V, which yielded the following results:

- Registers – DCLM classifier: 0.136
- Registers – FineWeb-edu classifier: 0.221
- Registers – NemoCurator classifier: 0.419

In the context of machine learning, these values indicate a low to moderate connection between the classifiers, with the NemoCurator showing the greatest association, which can be considered somewhat substantial. To account for the two ordinal classification schemes, FineWeb-edu and NemoCurator, we also calculated the η^2 -metric, which measures the amount of impact a variable has on the variance of another. These results were 0.173 and 0.273, for FineWeb-edu and NemoCurator, respectively, which similarly show a strong connection between the schemes but do not imply direct overlap.

Although the above analysis shows associations between the studied schemes, the register classes coincide with all classes of the three tools, with the exception of FineWeb-edu’s highest quality class 5, which only occurs in conjunction with register classes *Instructive-Informational* and *Informational Description*. We visualise this in Figures 4, 5, and 6. This means that our register scheme can be seen as mutually reinforcing when combined with other established quality labelling schemes. We specifically see from our results that the 3 tools we compared to the register scheme often assign high-quality labels to documents from registers *Informational Description*, *Description*, and the case of NemoCurator, *News*, but neglect classes that we found beneficial in this study, such as *How-to-Instructions* and *Opinion*. This highlights the value of using the non-binary but linguistically motivated and versatile scheme that the registers offer for curating data.

⁴<https://github.com/NVIDIA/Megatron-LM>

⁵<https://github.com/huggingface/nanotron>

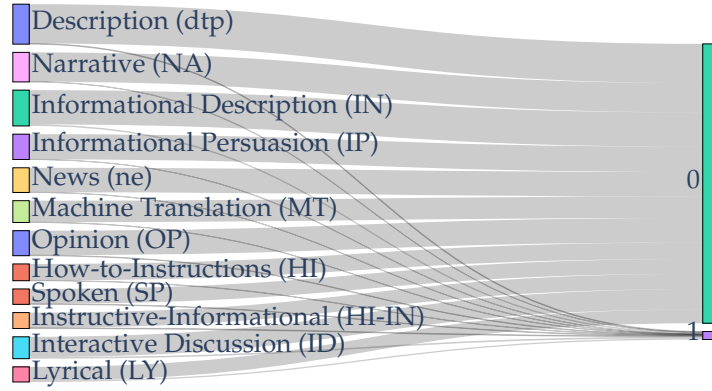


Figure 4: Visualised cross-classification for our register scheme and the DCLM quality classifier. DCLM label 0 corresponds to low quality and label 1 to high quality.

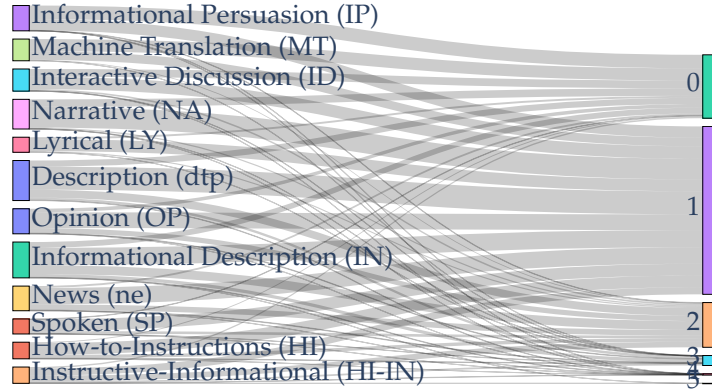


Figure 5: Visualised cross-classification for our register scheme and the FineWeb-edu classifier. FineWeb-edu label 0 corresponds to the lowest quality and label 5 to the highest quality.

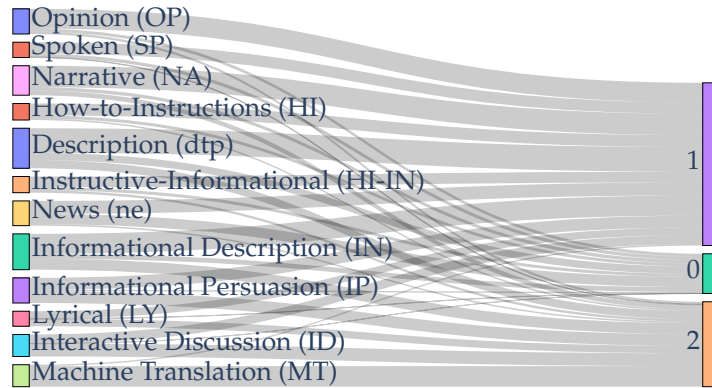


Figure 6: Visualised cross-classification for our register scheme and the NVIDIA NemoCurator classifier. NemoCurator label 0 corresponds to the highest quality and label 2 to the lowest quality.

D Numerical results

We present the results in numerical format in Table 3. As stated in Section 3.3, we reweigh the accuracy outputted by LightEval, `acc_norm`, as the calculation of `acc_norm` is heavily biased against MLLU. MMLU consists of 57 tasks, which are considered separate benchmarks by LightEval. This is why values of `acc_norm` are lower than “Accuracy” in the table: average performance of our models on MMLU is just under 0.3.

Pretraining dataset	Step	Tokens	Accuracy	<code>acc_norm</code>	<code>stderr</code>
FineWeb	1000	2 B	0.348450	0.269707	0.030520
	50000	105 B	0.465567	0.330767	0.031929
<i>How-to-Instructions</i> (HI)	1000	2 B	0.351006	0.264036	0.030214
	50000	105 B	0.446558	0.294467	0.030880
<i>Instructive-Informational</i> (HI-IN)	1000	2 B	0.352893	0.266001	0.030249
	50000	105 B	0.464502	0.314607	0.031506
<i>HI-IN-HI-dtp</i>	1000	2 B	0.354339	0.265349	0.030291
	50000	105 B	0.465616	0.320815	0.031642
<i>HI-IN-HI-dtp-OP</i>	1000	2 B	0.352104	0.267529	0.030357
	50000	105 B	0.475093	0.331857	0.032058
<i>HI-IN-HI-dtp-OP-NA</i>	1000	2 B	0.354445	0.264463	0.030214
	50000	105 B	0.472279	0.319987	0.031603
<i>HI-IN-HI-dtp-OP-NA-ID</i>	1000	2 B	0.348062	0.263200	0.030155
	50000	105 B	0.463591	0.324652	0.031859
<i>HI-IN-HI-dtp-OP-NA-ID-SP</i>	1000	2 B	0.353098	0.268480	0.030369
	50000	105 B	0.464738	0.325085	0.031820
HPLT v2	1000	2 B	0.341536	0.265107	0.030274
	50000	105 B	0.457484	0.328140	0.032064
<i>Interactive Discussion</i> (ID)	1000	2 B	0.336598	0.264045	0.030252
	50000	105 B	0.430686	0.298429	0.031214
<i>Informational Description</i> (IN)	1000	2 B	0.344821	0.270357	0.030517
	50000	105 B	0.437347	0.333771	0.032217
<i>Informational Persuasion</i> (IP)	1000	2 B	0.336226	0.261104	0.030146
	50000	105 B	0.393026	0.276895	0.030480
<i>Lyrical</i> (LY)	1000	2 B	0.322654	0.261681	0.030248
	50000	105 B	0.357626	0.271479	0.030604
<i>Machine Translation</i> (MT)	1000	2 B	0.331829	0.262588	0.030269
	50000	105 B	0.374349	0.281125	0.030821
<i>Narrative</i> (NA)	1000	2 B	0.336791	0.263844	0.030236
	50000	105 B	0.441469	0.297377	0.031149
<i>Opinion</i> (OP)	1000	2 B	0.343446	0.262086	0.030109
	50000	105 B	0.446629	0.318284	0.031720
<i>Spoken</i> (SP)	1000	2 B	0.340778	0.266882	0.030336
	50000	105 B	0.422314	0.303859	0.031381
<i>Description</i> (dtp)	1000	2 B	0.347940	0.268266	0.030463
	50000	105 B	0.452347	0.329816	0.032027
<i>News</i> (ne)	1000	2 B	0.332720	0.257563	0.029931
	50000	105 B	0.417933	0.304612	0.031457

Table 3: Accuracies for all models, first and last checkpoint. “Accuracy” denotes accuracy weighted by benchmark, which we present in our figures. Reweighting was done to mitigate the dominance of MMLU in the results, see Section 3.3. `acc_norm` and `stderr` stand for average accuracy and standard error as given by LightEval.

E Performance by benchmark and register

Figures 7, 8, and 9 present the results over all checkpoints, with Figure 7 showing results by register, others by benchmark. To augment readability, we again applied a rolling average. In Figure 7, the values are normalised with respect to the average benchmark score on the final checkpoint; for example, the *Description* (dtp) model outperforms the average of all registers by 0.1 points on the final checkpoint on the *ARC Easy* benchmark.

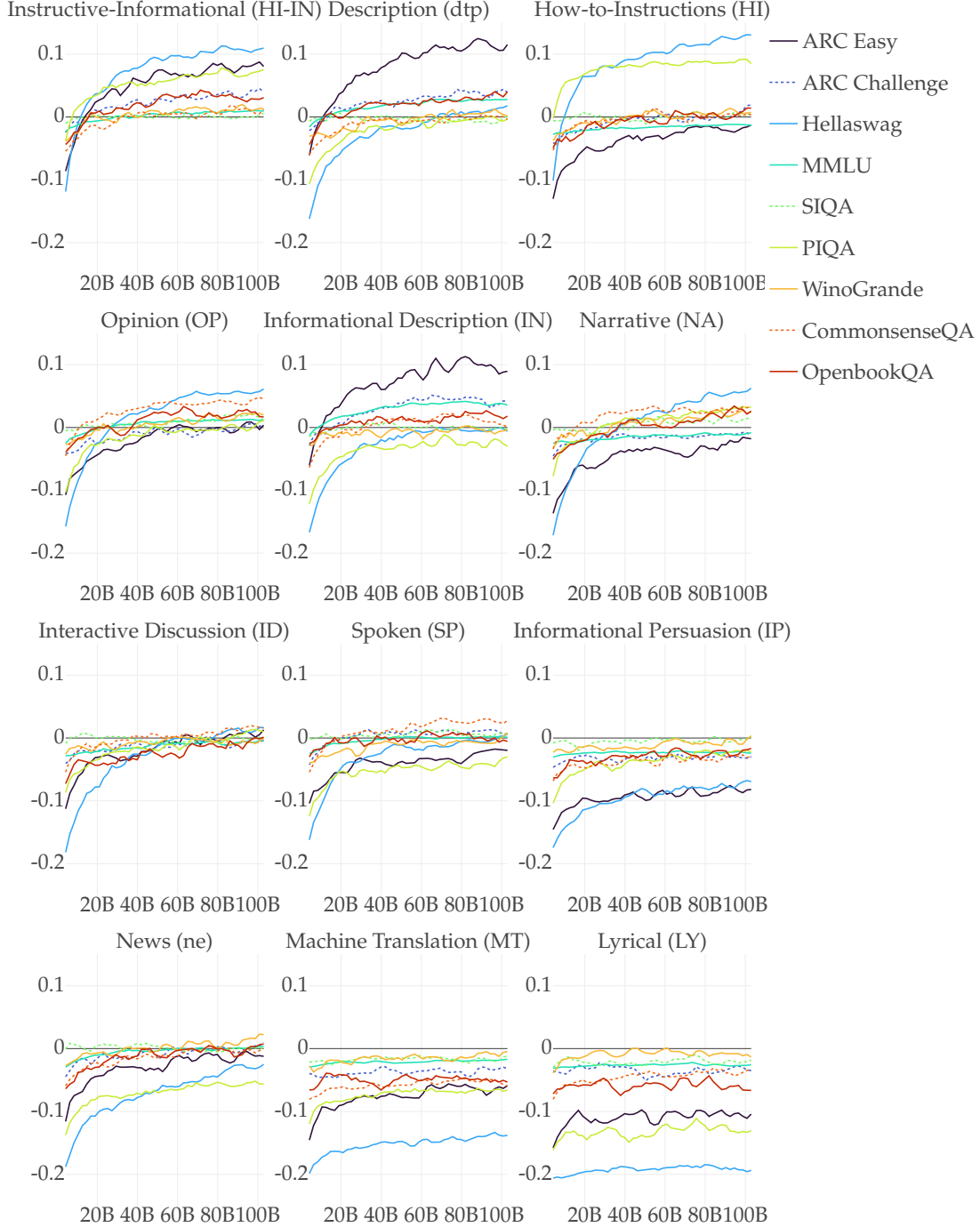


Figure 7: Models' performance by register. Results are scaled by the average last checkpoint performance to highlight the models' differences. Dotted lines to increase readability.

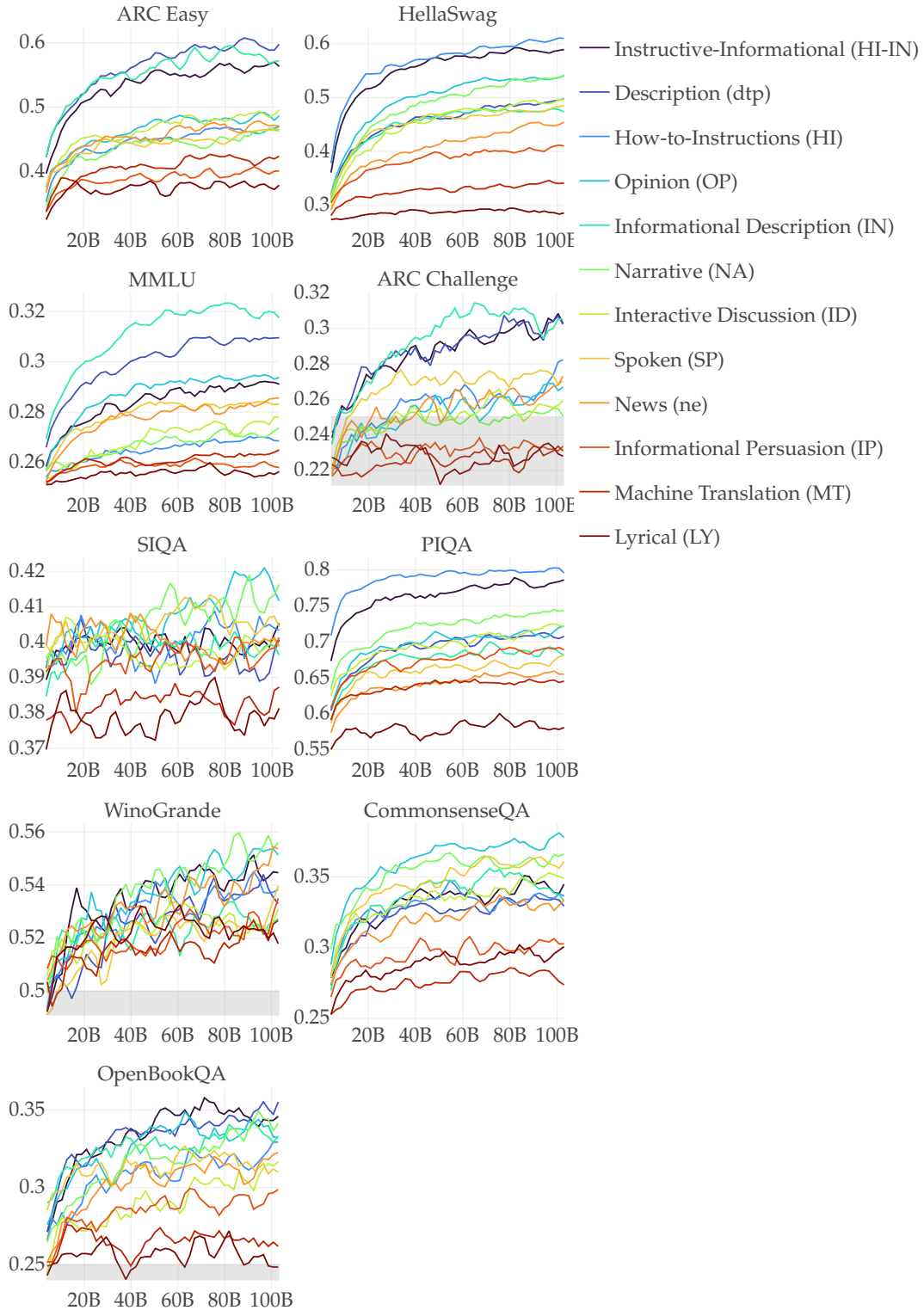


Figure 8: Accuracies of register models on individual benchmarks. Random-guess thresholds shown in grey when at least one model falls under it. Rolling average over 6 billion tokens applied for ease of reading overlapping lines.

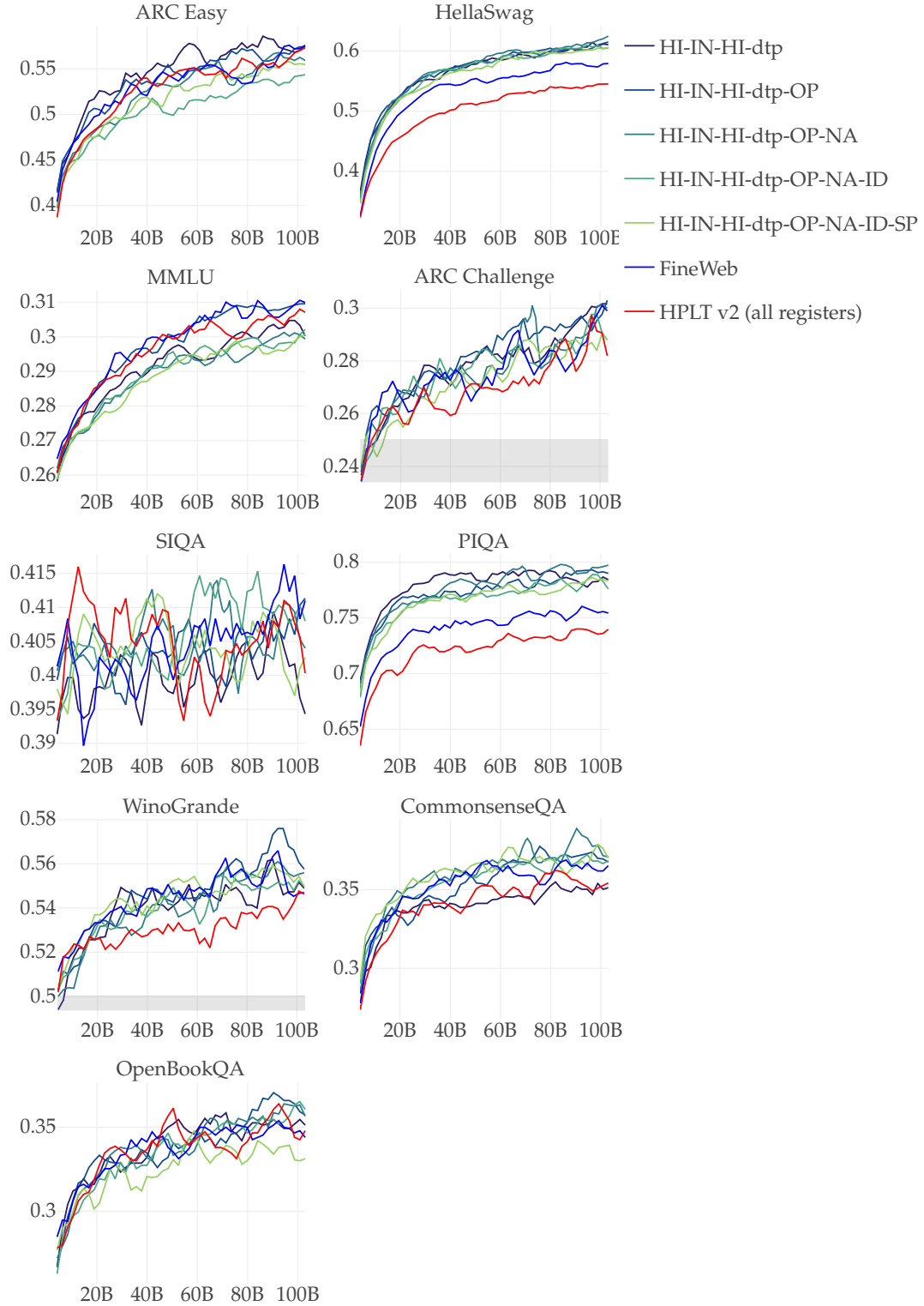


Figure 9: Accuracies of combination models on individual benchmarks. Random-guess thresholds shown in grey when at least one model falls under it. Rolling average over 6 billion tokens applied for ease of reading overlapping lines.