The Impact of Copyrighted Material on Large Language Models: A Norwegian Perspective

Javier de la Rosa[1] **Vladislav Mikhailov**[2] **Lemei Zhang**[3] **Freddy Wetjen**[1] **David Samuel**[2]
Peng Liu[3] **Rolv-Arild Braaten**[1] **Petter Mæhlum**[2] **Magnus Breder Birkenes**[1]
Andrey Kutuzov[2] **Tita Enstad**[1] **Hans Christian Farsethås**[2] **Svein Arne Brygfjeld**[1]
Jon Atle Gulla[3] **Stephan Oepen**[2] **Erik Velldal**[2] **Wilfred Østgulen**[1] **Lilja Øvrelid**[2]
Aslak Sira Myhre[1]

[1]National Library of Norway
[2]University of Oslo
[3]Norwegian University of Science and Technology

Correspondence: versae@nb.no

Abstract

The use of copyrighted materials in training language models raises critical legal and ethical questions. This paper presents a framework for and the results of empirically assessing the impact of publisher-controlled copyrighted corpora on the performance of generative large language models (LLMs) for Norwegian. When evaluated on a diverse set of tasks, we found that adding both books and newspapers to the data mixture of LLMs tend to improve their performance, while the addition of fiction works seems to be detrimental. Our experiments could inform the creation of a compensation scheme for authors whose works contribute to AI development.

1 Introduction

Generative language models have radically reshaped the landscape of natural language processing (NLP), enabling the development of systems that can generate and interact with human language at an unprecedented level. This includes Norwegian, for which several large language models (LLMs) have been trained and published in the recent years using different architectures and licensing choices (Kummervold et al., 2021; Kutuzov et al., 2021; Samuel et al., 2023, 2025; Liu et al., 2024).

However, the vast quantities of data required for training these models often include copyrighted materials, presenting novel challenges related to intellectual property rights and compensation. Additionally, prior research has highlighted significant concerns about dataset composition and quality in large-scale web-crawled datasets, emphasizing the need for more responsible data curation practices (Kreutzer et al., 2022; Artetxe et al., 2022; Penedo et al., 2024). Together, these challenges have led to numerous lawsuits across jurisdictions, fundamentally questioning the legitimacy of training models on copyrighted data without explicit permissions from content creators (Panettieri, 2024; Madigan, 2024; Weisenberger et al., 2024).[1]

The first wave of lawsuits emerged shortly after the public release of advanced generative AI models (see Appendix A). Content creators, including authors, visual artists, and musicians, began to express concerns about the unauthorized use of their work in training datasets. Multiple class-action lawsuits were filed in the United States, accusing prominent AI companies such as OpenAI and Meta Platforms of infringing on copyright laws by using copyrighted materials without obtaining explicit permissions. The authors argued that the unauthorized use of their works without any form of compensation or recognition undermines their intellectual property rights and jeopardizes their ability to earn a living from their creative endeavors. In Europe, a coalition of news publishers has taken legal action against Google and Meta Platforms, arguing that the use of journalistic content in training models without fair re-

[1]See Gervais et al. (2024) for an in-depth introduction on how LLMs are being interpreted in the legal domain.

Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), pages 544–560
March 3-4, 2025

muneration constitutes a breach of copyright and undermines the sustainability of high-quality journalism. Likewise, Norwegian rights-holder organizations representing publishing houses across the country, contacted the government in late 2023 expressing their concerns over the use of their material in training generative language models and demanding some sort of compensation were their contents to be used in the training of generative language models. As a result, the Ministry of Culture and Equality (*Kultur- og likestillingsdepartementet*) instructed the National Library to create a data-driven report they could use in order to make informed decisions in the elaboration of a compensation scheme for the authors. Led by the National Library of Norway, a consortium was formed together with the University of Oslo and the Norwegian University of Science and Technology under the umbrella of the so-called Mímir Project.[2]

In this context, and under the umbrella of Mímir, this paper describes a first attempt at empirically evaluating the impact of copyrighted content in the training of LLMs for Norwegian. We introduce a set of carefully curated datasets that are later used in the training of foundational, domain-tuned, and instruction-tuned models. We establish the proper training conditions to be able to compare models trained on the different datasets. A newly created benchmarking suite is used to evaluate the performance of each individual model and make the comparison meaningful. As a collaborative effort among several institutions, the results of our investigations set the basis to guide policymaking and proper compensation schemes for authors and right-holders in Norway (Nasjonalbiblioteket, 2024).

2 Methodology

The methodology involves a comprehensive analysis that spans several stages. Initially, a diverse corpus of primarily Norwegian language data is assembled, incorporating both copyrighted and non-copyrighted materials, plus materials commonly found on the Internet. This corpus serves as the foundation for training various LLMs, each with different configurations and access levels to copyrighted content. By comparing the performance of these models across a range of linguistic and natural language processing tasks, such as text generation, translation, summarization, question-answering, sentiment analysis and more, we seek to quantify the specific contributions of copyrighted materials to the overall model quality.

To ensure robustness and reliability, the evaluation framework focuses on generation capabilities, natural language understanding, and linguistically-inspired metrics. Quantitative measures include traditional NLP metrics like accuracy, F1, BLEU, and ROUGE, which provide assessments of model accuracy and fluency. Linguistic analysis, on the other hand, involves assessing the coherence, language variability, and contextual relevance of the generated outputs. This dual approach allows for a nuanced understanding of how copyrighted materials impact the performance and utility of LLMs.

3 Data Collection

With the objective of setting up a realistic training scenario where using Internet crawled sources is commonplace, we gathered publicly available text collections like Wikipedia, datasets from the HPLT (de Gibert et al., 2024) and CulturaX (Nguyen et al., 2024) projects, code in different programming languages from Lozhkov et al. (2024), governmental reports and publications published under open licenses, and books and newspapers articles in the public domain.

We then collaborated with the National Library of Norway and the rights-holder organizations to gain access to protected materials. Through the legal deposit act, the National Library of Norway has digitized almost all books in Norwegian and around 85% of the newspapers ever published in the country (Nasjonalbiblioteket, 2024). Where the quality of the digitized material was not enough (e.g., due to OCR processing), or was not been legally deposited (e.g., paywalled content), specific agreements were put in place to obtain the material from third party organizations such as the Norwegian Broadcasting Corporation (NRK), the TV channel TV2, and the newspaper conglomerates Amedia and Schibsted. In line with provisions that allow research on language technology and data mining (*Åndsverkloven*), and with the consent of the Norwegian right-holders, this study primarily relied on material legally deposited at, or under agreement with, the National Library of Norway. Specifically, we focus our study on the collection of publisher-controlled books and newspapers ar-

[2] A name chosen after a figure in Norse mythology renowned for his knowledge and wisdom.

Dataset	Documents	Words
base	60,182,586	40,125,975,241
extended	125,285,547	82,149,281,266

Table 1: Number of documents and words in each of the core datasets. Words refer to whitespace-separated sub-strings.

ticles.

3.1 Core Datasets

This mixture of data (see Figure 1 and Appendix C) allowed us to evaluate the impact of high-quality publisher-controlled copyright-protected corpora versus other sources commonly available on the Internet. The models trained on the copyrighted materials will not be made publicly available for further use and only serve the purpose of this study.

We followed the recipe of the Norwegian Colossal Corpus (NCC) by Kummervold et al. (2022), adapting and updating it with new up-to-date contents, re-OCRing some materials, enriching their metadata, and ensuring uniform format and functionality across datasets. The preparation involved cleaning, deduplication, metadata tagging, and language balancing to maintain consistent representation of Norwegian, preventing other languages from overshadowing it. The corpus was divided into two main datasets: a **base** dataset excluding publisher-controlled copyright-protected books and newspapers,[3] and an **extended** dataset that included all collected texts, thus including all of **base** (see Table 1).

We decided to include texts from other Scandinavian (Swedish, Danish, and Icelandic) and English sources to boost the performance of the resulting language models via cross-lingual transfer (Conneau et al., 2020b; Xue et al., 2021). To ensure that languages other than Norwegian, and primarily coming via Internet crawling, were balanced, we adapted the perplexity-based sampling strategy from De la Rosa et al. (2022) to maintain a high quality in the selected data. Instead of sampling a fixed number of documents, parameters for a Gaussian curve were calculated from 500,000-1M random documents per source, utilizing Wikipedia-based Kneser-Ney language models from Wenzek et al. (2019) and Conneau et al. (2020a).[4] We also modified the perplexity calculation to account for normalized text. These parameters then guided dataset sub-sampling to target ratios per language, reducing foreign language content while maintaining quality (Appendix B).

It is also important to notice than in order to maintain the language distributions for foreign languages with respect to the amount of Norwegian texts, the total number of documents in foreign languages in the **extended** dataset is consequently higher and slightly different (due to the sampling strategies) than that of **base**; we keep the same ratios (see Appendix C).

Subset	Documents	Words
books	492,281	18,122,699,498
newspapers	46,764,024	9,001,803,515
books + newspapers	47,256,305	26,078,915,554
fiction books	117,319	5,287,109,366
nonfiction books	359,979	12,384,323,012
nonfiction books + newspapers	42,083,532	20,340,539,068
original books	392,887	13,352,261,605
original books + newspapers	47,156,911	22,354,065,120
translated books	96,258	4,695,814,506

Table 2: Number of documents and words (comma separated) in each subset of the publisher-controlled corpora.

3.2 Domain Specific Subsets

The publisher-controlled copyright-protected materials present in the **extended** dataset were further divided into groups attending to different criteria. These subsets were carefully designed to test the effect of adding them to the training sets for LLMs. We split the books into fiction vs nonfiction, and original works in Norwegian vs translations. While most books in the collection had metadata information regarding the original language in which a work was written in, genre labels were more scarce. To overcome this limitation, we built a Doc2Vec model (Le and Mikolov, 2014) that classified fiction vs nonfiction with 98% accuracy and used it to annotate books for which this information was missing.[5] As shown in Table 2, we then built domain specific subsets to investigate 1) the effect of books vs newspapers vs books + newspapers, 2) the effect of factuality by adding only fiction words, only nonfiction works,

[3] Except for newspapers that fall under the Language Technology Use (*Språkteknologiformål*), as they were already included in other datasets such as NCC.

[4] Built with KenLM (Heafield, 2011).

[5] https://huggingface.co/mimir-project/literary-form-classifier

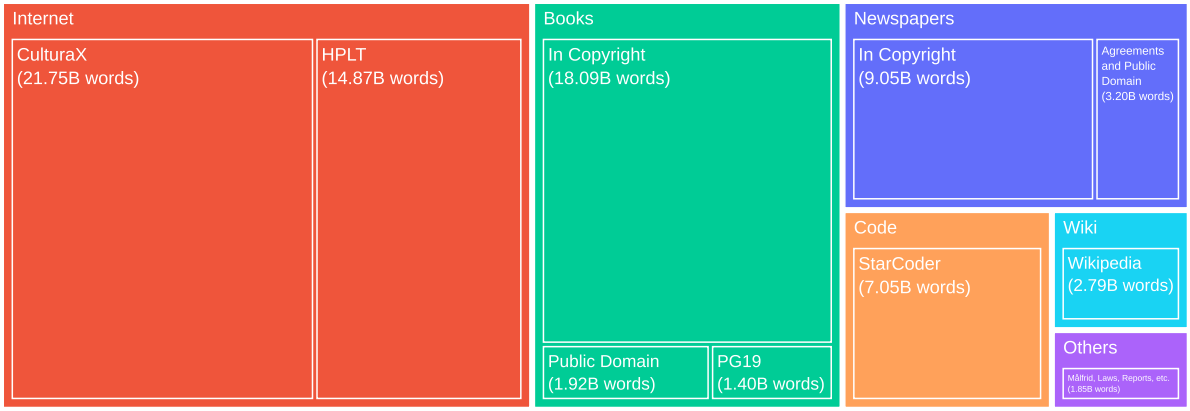

Figure 1: Treemap with the final number of words (comma separated) contributed by each source after cleaning and deduplication.

and nonfiction works + newspapers, and 3) the effect of adding content written originally in Norwegian, such as original books or original books + newspapers, vs translated books.

3.3 Instruction-tuning Datasets

To align the models more closely with human objectives and assess whether instruction tuning with limited high-quality data can enhance the performance of our pre-trained models across various tasks, we built upon prior work and collected nearly 5,000 instructions annotated by research assistants.[6] The instructions were formatted as *(instruction, input, output)* triplets, where *instruction* refers to the directive provided by humans for the model, *input* is an optional field containing task-related information, and *output* denotes the desired response that follows the given instruction.

The instruction tuning dataset combines three key categories –Reading Comprehension, Norwegian Culture, and Words and Expressions– with diverse domains to enhance model performance. The domains include Literature, Commonsense, Geography, Language, History, Sports, Entertainment, Food, Politics, Science, Art, Music, and Culture. The variety of the instructions seeks to improve the model's ability to understand complex texts, provide culturally relevant responses, and handle language nuances, resulting in more versatile, knowledgeable, and context-aware LLMs.

4 Model Training

The training phase involved multiple models, each based on the Mistral architecture (Jiang et al., 2023). The training was conducted in the following stages.

1. To measure the overall impact of publisher-controlled copyrighted corpora and its effect in realistic scenarios, we conducted pre-training on the **base** and **extended** datasets, both from scratch and using the existing weights (warm) of the pre-trained model Mistral 7B v0.1.[7] These four *core models* were trained on the same amount of data, 64,000 steps of 4 million sub-word tokens each. using identical sets of hyperparameters (see Table 7 in Appendix D). This roughly translates to 3 epochs for the **base** dataset and 2 for the **extended** dataset, which according to Muennighoff et al. (2023) is still far from saturating the available data.
2. To further isolate the effect of different ablations of the publisher-controlled copyright-protected corpora, we continuously fine tuned the model trained on **base** from scratch for an extra 10,000 steps on each of the 9 domain specific subsets.
3. The core models were also fine tuned on the instruction data for 4 iterations to evaluate their performance on downstream tasks.

Overall, we trained 17 models (7 billion parameters each) using a total of 270,000 GPU-hours. Model training specifications are shown in Table 3. The infrastructure for training included the LUMI supercomputer, Idunn cluster, and Google

[6] https://huggingface.co/datasets/mimir-project/mimir-instruction

[7] https://huggingface.co/mistralai/Mistral-7B-v0.1

Model	Initialization	GPU/hours	Accelerator
Core Models			
base	From scratch	50K	AMD MI250X
extended	From scratch	50K	AMD MI250X
base (warm)	Mistral 7B v0.1	13.8K	NVIDIA H100
extended (warm)	Mistral 7B v0.1	55.6K	AMD MI250X
Domain Tuned Models			
base + fiction books	base	7.5K	AMD MI250X
base + nonfiction books	base	7.5K	AMD MI250X
base + nonfiction books + newspapers	base	7.5K	AMD MI250X
base + newspapers	base	4.8K	Google TPUv4
base + books	base	4.8K	Google TPUv4
base + books + newspapers	base	4.8K	Google TPUv4
base + original books + newspapers	base	9.1K	AMD MI250X
base + original books	base	9.1K	AMD MI250X
base + translated books	base	9.1K	AMD MI250X
Instruction Fine Tuned Models			
base *instruct*	base	14.2	NVIDIA H100
extended *instruct*	extended	14.2	NVIDIA H100
base (warm) *instruct*	base (warm)	14.2	NVIDIA H100
extended (warm) *instruct*	extended (warm)	14.2	NVIDIA H100

Table 3: Model training specifications, where *Model* represents the model identifier and the data used for training, *Initialization* represents the base model used for training, *GPU/hours* indicates the total GPU hours required for model training, and *Accelerator* represents the type of accelerator used.

TPUs through the Tensor Research Cloud program[8]. Besides, we trained two tokenizers with the **base** and **extended** datasets separately, both with the same vocabulary size of $32,768$. After an initial test of the fertility of the tokenizers,[9] we found the difference between them was only 0.0013. Therefore, we decided to use the same tokenizer trained with the **base** dataset for all the models.

5 Evaluation Framework

In our empirical evaluation experiments, we utilize `NorEval`,[10] a publicly available framework for evaluating Norwegian generative LLMs built on `lm-evaluation-harness` (Gao et al., 2024). We consider 28 tasks, which test model's various Norwegian language understanding and generation abilities. `NorEval` covers both Norwegian language varieties (Bokmål and Nynorsk) and provides a set of 4–6 prompts for each downstream task. The tasks can be grouped into nine higher level **skills**:

1. **Sentiment Analysis**, here defined as binary polarity classification on both the sentence- and document-level based on the existing `NoReC` datasets of professional reviews (Velldal et al., 2018; Øvrelid et al., 2020).
2. **Fairness & Truthfulness**. Fairness refers to the absence of bias in the predictions and outputs of a model. Evaluating fairness ensures that the model does not favor or discriminate against particular groups based on attributes like race, gender, or ethnicity. This skill was evaluated using a newly-created dataset,[11] which covers a wide range of bias types, including race, religion, gender, geography, occupation, age etc. Truthfulness involves the accuracy and reliability of the information produced by the model, ensuring it generates factual and verifiable content. This skill was evaluated using `NorTruthfulQA` (Mikhailov et al., 2025), which assesses whether a model is truthful in selecting and generating answers to questions that involve common human misconceptions.[12]
3. **Reading Comprehension**, which measures the ability of a model to understand and interpret text. It involves answering questions

[8] To assess the deviation introduced by differences in training infrastructures and platforms across the participating institutions, each team trained a control model with 1.5B parameters based on the Llama 2 architecture. The training setups were identical, utilizing the **base** dataset. After comparing the validation loss curves from each team, we found that the curves were almost the same, with a deviation of less than 0.05 in terms of the final convergence validation loss.

[9] Fertility expresses the fragmentation rate of a tokenizer and is $\#tokens/\#words$ in one corpus.

[10] https://github.com/ltgoslo/noreval

[11] https://huggingface.co/datasets/mimir-project/mimir-bias

[12] https://huggingface.co/datasets/ltg/nortruthfulqa_mc and https://huggingface.co/datasets/ltg/nortruthfulqa_gen

about a given passage, summarizing content, or explaining the meaning of specific phrases or sentences. This skill estimates how well the model grasps the context and details in the text. It was evaluated using the existing extractive question-answering `NorQUAD` dataset (Ivanova et al., 2023) and multiple-choice question-answering Belebele dataset (Bandarkar et al., 2024).

4. **World Knowledge**, which assesses the extent of factual information a language model has about the world. This includes historical events, geographical data, scientific facts, cultural knowledge, and more. The model should correctly answer questions or provide information based on real-world knowledge. This skill was evaluated using the `NorOpenBookQA` and `NRK-Quiz-QA` by Mikhailov et al. (2025).[13]

5. **Commonsense Reasoning**, which involves the ability of a model to make logical inferences based on everyday knowledge and understanding of the world. The model should reason about situations that require practical, everyday knowledge that people take for granted. This skill was evaluated using `NorCommonsenseQA` (Mikhailov et al., 2025),[14] which consists of multiple-choice commonsense question answer-pairs which adapts the corresponding English CommonsenseQA dataset (Talmor et al., 2019) to Norwegian.

6. **Norwegian Language** evaluation focuses on the ability of a model to understand and generate text in Norwegian, specifically its grammar, structure, and sentence construction. This skill is important for assessing how well the model handles Norwegian and their specific syntactic rules. It was evaluated using the existing `NCB` (Farsethås and Tjøstheim, 2024) and `ASK-GEC` (Jentoft, 2023) datasets, and the newly-created `NorIdiom` dataset.[15]

7. **Summarization**, which measures the ability of a model to condense longer pieces of text into shorter, coherent summaries that capture the main points. This skill is crucial for applications where users need a quick understanding of large volumes of information, such as news articles or research papers. It was evaluated using the `NorSumm` dataset (Touileb et al., 2025).[16]

8. **Translation**, which assesses how accurately a language model can convert a text from one language to another while preserving the meaning, tone, and context. It was evaluated using the existing `Tatoeba` dataset (Tiedemann, 2020). The following six language pairs are considered: Bokmål ↔ Nynorsk, Bokmål ↔ English, and English ↔ Nynorsk.

9. **Variation and Readability**, which consists of measuring the lexical diversity of a model by looking at the amount of redundancy in the text it produces and at the readability of these texts measured by average sentence length and the proportion of long words. As such, this skill evaluation did not require any specific benchmarking datasets.

We follow the standard in-context learning evaluation design for pretrained decoder-only language models (e.g., Brown et al., 2020; Touvron et al., 2023), which includes zero-shot and few-shot evaluation. In this paper, for the sake of simplicity, we selected the most common metrics per task and aggregated scores using a simple cumulative sum per higher-level skill. In order to aggregate results into an overall score, with the caveats of aggregating metrics of different nature, scores were extracted for the best available $\{0, 1, 4, 16\}$-shot configuration for each task and the best score for each of the prompts. Metrics were normalized to exhibit the same higher-is-better behavior in a range of 0 to 100.

6 Results

The evaluation of the trained models demonstrated that incorporating publisher-controlled copyright-protected corpora provided a measurable boost in performance across a range of NLP tasks. To illustrate the overall performance differences, Figure 2 shows the total scores across all skills, averaged by task for each model. Non-aggregated scores

[13] https://huggingface.co/datasets/ltg/noropenbookqa and https://huggingface.co/datasets/ltg/nrk_quiz_qa

[14] https://huggingface.co/datasets/ltg/norcommonsenseqa

[15] https://huggingface.co/datasets/mimir-project/noridiom

[16] https://huggingface.co/datasets/SamiaT/NorSumm

Figure 2: Total summed scores across all skills averaged by task for each model. Best scores among from-scratch models underlined, best overall from-scratch in **bold**. Dashed line at the **base** score.

for all tasks, prompts, and models are available at the Mímir repository.[17]

6.1 Core Models

As shown in Table 4 and Figure 2, the performance analysis of core models across various tasks reveals distinct strengths for different configurations. The **base** (warm-started) configuration consistently excels in Sentiment Analysis, World Knowledge, and Norwegian Language. In contrast, the extended (warm-started) configuration leads in Fairness & Truthfulness, Reading Comprehension, Commonsense Reasoning, Translation, and Variation & Readability, indicating its robust performance for language-intensive tasks. The **base** configuration generally lags behind others, scoring the lowest across multiple tasks. Meanwhile, the **extended** configuration performs well, particularly in Summarization. Furthermore, it indicates that we could leverage the existing metadata available at the National Library to tailor subsets of the publisher-controlled copyrighted corpora and build models that excel at specific tasks. However, the difference between the **base** and **extended** warm-started models is very small.

[17] https://github.com/mimir-project/mimir-evaluation

Model	SA	FT	RC	WK	RC	NL	S	T	VR
extended	3	2	3	3	2	2	1	3	2
base	4	3	4	4	3	4	3	4	3
extended (warm)	2	3	1	2	1	1	2	1	1
base (warm)	1	1	2	1	1	3	2	2	4

Table 4: Results for ranking the core models on all tasks by skill via (i) finding the best k-shot configuration for each task and (ii) aggregating metric-wise rankings. SA=Sentiment Analysis. FT=Fairness & Truthfulness. RC=Reading Comprehension. NL=Norwegian Language. WK=World Knowledge. CR=Commonsense Reasoning. S=Summarization. T=Translation. VR=Variation & Readability. Lower is better.

Further testing is required to assess whether this difference is still statistically significant.[18]

While warm-started models generally outperformed models trained from scratch, there was reduced sensitivity to the presence of copyrighted materials. This suggests that the pre-existing weights, which were primarily trained on English data, diminished the impact of adding high-quality Norwegian copyrighted texts (see also Section 7).

[18] Detailed scores available in Appendix F Table 8.

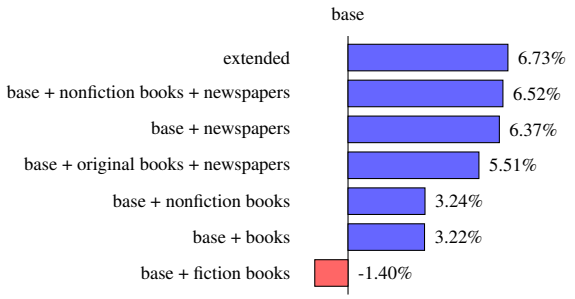

Figure 3: Average percentage gains over the performance of the base model. Negative results indicate a decrease in performance over base, positive results a gain.

6.2 Domain-tuned Models

To further explore the specific effects of different types of training data, we analyzed the gains in performance by focusing on different sub-corpora, such as newspapers, books, and mixed datasets. Figure 3 provides an overview of the average percentage gains for models trained on various data configurations compared to the **base** model. It shows that the **extended** model exhibits the highest average gain at 6.73%, indicating substantial overall improvement. The addition of nonfiction books and newspapers follows with a 6.52% gain, and the addition of only newspapers shows a 6.37% improvement. Other configurations, such as adding original books and newspapers or nonfiction books, also demonstrate positive gains of 5.51% and 3.22%, respectively. Conversely, the addition of fiction books is the only one to show a negative performance, with a decrease of 1.40%. Interestingly, when decomposed by skill, the addition of fiction books makes the model excel at generating more diverse texts (see Figure 5 in Appendix E).

6.3 Instruction-tuned Models

Lastly, as shown in Figure 4, when the core models are further fine-tuned on data to follow instructions, the gains across models are all consistent, showing that the core advantage lies in the pretraining data, while further training on instructions gives a consistent boost in performance. Instruction tuning also seems to reduce the gap between the **base** and **extended** configurations, suggesting that publisher-controlled copyrighted corpora might become less relevant as supervised fine-tuning datasets increase in size in the post-training phases of LLMs. Interestingly, adding Norwegian instruction data on top of the **extended** model seems enough to improve over the performance of Mistral 7B v0.1.

7 Discussion

Our findings underline the value of copyrighted materials in improving the performance of generative language models, particularly for specialized NLP tasks in Norwegian. The inclusion of these curated publisher-controlled texts provide a substantial advantage in terms of language richness, coherence, and context-specific understanding. However, these advantages are significantly less evident in models that are warm-started using weights pre-trained on other languages, primarily English. We see two possible reasons for this:

1. The *amount* of training data matters more than its quality or licensing status. Warm-started models are effectively trained on more data than the 'from-scratch' models, and at some point adding even more data brings diminishing returns (with a given model size).
2. Publisher-controlled copyrighted Norwegian data is indeed beneficial for LLMs, but the original models used for warm-starting *were presumably already pre-trained on datasets that may share similarities with this data*. Due to the lack of transparency regarding the exact composition of training datasets in models like Mistral, concerns about potential data contamination remain relevant. This overlap could explain why continuous pre-training on similar content did not yield the expected benefits for the warm-started extended models (Li et al., 2024; Dong et al., 2024; Xu et al., 2024; Samuel et al., 2024).

7.1 Ethical and Legal Considerations

The use of copyrighted materials in model training raises significant ethical and legal questions. The observed improvements in model quality must be balanced against the rights of content creators, who have not consented to the use of their work. This highlights the need for guidelines and compensation mechanisms that recognize the value of copyrighted materials in LLM development.

7.2 Implications for Policy

The empirical evidence gathered in our research is crucial for informing copyright policy in the digital age. Policymakers can use these findings to

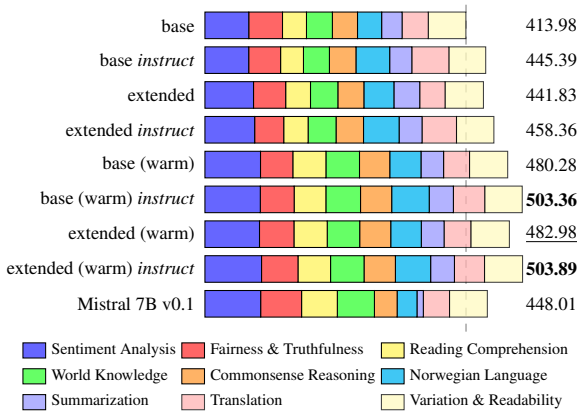

Figure 4: Total scores (sum) of all averaged scores per skill for the core models and their instruct versions, with original Mistral 7B v0.1 for reference. Dashed line at the **base** score. Best scores in **bold**, second best underlined.

establish frameworks that ensure creators are adequately compensated, balancing the needs of LLM innovation with the rights of authors and publishers. This is particularly relevant in light of ongoing lawsuits against major AI companies.

8 Conclusion

Our study represents a pioneering effort to quantify the impact of copyrighted materials on LLMs for Norwegian. Our results indicate that high-quality publisher-controlled copyrighted content significantly enhances model performance, especially for complex NLP tasks. However, these benefits bring forth ethical and legal challenges that must be addressed to ensure a sustainable and fair approach to LLM development. By providing empirical evidence, we aim to contribute to the ongoing discourse on the role of copyright in AI and inform future policies that support both innovation and the rights of content creators.

9 Future Work

Future work should focus on testing models at various scales and different pre-trained open weights to better understand how dataset composition affects performance. By experimenting with models of different sizes, we could identify any scaling thresholds where the impact of copyrighted material varies significantly. In retrospect, one notable flaw in the experimental design is the lack of fully traceable and transparent models, such as OLMo (Groeneveld et al., 2024), which provide detailed documentation of their training data and processes. Without utilizing models with verifiable data provenance, it becomes challenging to accurately assess how specific dataset compositions, including copyrighted or genre-specific materials, influence model behavior and performance for warm-started models. Incorporating traceable models would improve the reproducibility and reliability of findings, ensuring that conclusions drawn about the impact of various text genres are well-founded.

Additionally, the observed effects of fiction on model performance highlight the need to 1) examine how different types of fiction –such as fantasy or historical fiction– impact tasks like Sentiment Analysis and Commonsense Reasoning, and 2) design new and adequate benchmarks for evaluating the contribution of fiction in Norwegian LLMs for tasks such as creative writing, plot understanding, or descriptive language use. This investigation could clarify the role of fiction in model training and help refine data curation strategies.

Lastly, exploring genre-specific influences more deeply, including essays, technical writing, and narrative nonfiction, may reveal distinct benefits or biases tied to each genre. Analyzing these nuances, even in a diachronic manner, will guide balanced genre representation in datasets and support the development of better performing models.

10 Distribution

The **base** dataset and models were intended to be freely distributed, as all materials included were granted redistribution permissions under different agreements. After we communicated the results of our investigations to the different partners, some right-holders demanded a reinterpretation of the agreements (primarily the Language Technology Use, *Språkteknologiformål*), in the light of the results and this new era of generative AI. This prevented us from sharing publicly the exact models trained within the Mímir project, but instead we built a subset of **base**, which we are calling **core**, excluding the affected newspapers (around 1B words) and trained models both from scratch and from Mistral 7B v0.1. Their performance is on par with their **base** counterparts. We are also releasing these models under a permissive license.[19]

[19] https://huggingface.co/mimir-project/mimir-mistral-7b-core-scratch

Acknowledgments

We extend our sincere gratitude to Hans Eide from Sigma2 for facilitating access to the LUMI supercomputer, enabling the computationally intensive tasks integral to this study. Additionally, we thank Google for providing compute resources via the Tensor Research Cloud program, which significantly supported our model training efforts.

This project would not have been possible without the trust and collaboration of the Ministry of Culture and Equality, which empowered the National Library of Norway to spearhead this endeavor, with the invaluable contributions of the Norwegian University of Science and Technology (NTNU) and the University of Oslo (UiO), whose expertise and insights were instrumental throughout the process. We are grateful for their vision and faith in the potential of this research.

We are also deeply appreciative of Olaus Bergstrøm and the entire legal team at the National Library of Norway for their guidance on the legal dimensions of this research. Their expertise was invaluable in navigating the complexities of copyright law and ensuring compliance with the unique considerations surrounding the materials used in this project.

A special thanks goes to the representatives of the Norwegian rights-holder organizations, who not only agreed to the use of their materials for this project but were steadfast in their support of the initiative. Their cooperation and encouragement have been vital in ensuring the project's success and advancing research on the intersection of copyright and AI development.

References

Mikel Artetxe, Itziar Aldabe, Rodrigo Agerri, Olatz Perez-de Viñaspre, and Aitor Soroa. 2022. Does corpus quality really matter for low-resource languages? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7383–7390, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 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. Association for Computational Linguistics.

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. 2020. Language models are few-shot learners.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.

Javier De la Rosa, Eduardo González Ponferrada, Paulo Villegas, Pablo González de Prado Salas, Manu Romero, and María Grandury. 2022. BERTIN: Efficient pre-training of a Spanish language model using perplexity sampling. *Procesamiento de Lenguaje Natural*, 68:13–23.

Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, Bin Gu, Mengfei Yang, and Ge Li. 2024. Generalization or memorization: Data contamination and trustworthy evaluation for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*.

Hans Christian Farsethås and Joakim Tjøstheim. 2024. Norwegian comma benchmark. https://huggingface.co/datasets/hcfa/ncb.

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, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. A framework for few-shot language model evaluation.

and https://huggingface.co/mimir-project/mimir-mistral-7b-core

Daniel Gervais, Noam Shemtov, Haralambos Marmanis, and Catherine Rowland. 2024. The heart of the matter: Copyright, AI training, and LLMs.

Ona de Gibert, Graeme Nail, Nikolay Arefyev, Marta Bañón, Jelmer van der Linde, Shaoxiong Ji, Jaume Zaragoza-Bernabeu, Mikko Aulamo, Gema Ramírez-Sánchez, Andrey Kutuzov, Sampo Pyysalo, Stephan Oepen, and Jörg Tiedemann. 2024. A new massive multilingual dataset for high-performance language technologies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1116–1128, Torino, Italia. ELRA and ICCL.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, William Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah Smith, and Hannaneh Hajishirzi. 2024. OLMo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15789–15809, Bangkok, Thailand. Association for Computational Linguistics.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the sixth workshop on statistical machine translation*, pages 187–197.

Sardana Ivanova, Fredrik Andreassen, Matias Jentoft, Sondre Wold, and Lilja Øvrelid. 2023. NorQuAD: Norwegian question answering dataset. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 159–168, Tórshavn, Faroe Islands. University of Tartu Library.

Matias Jentoft. 2023. Grammatical error correction with byte-level language models. Master's thesis, University of Oslo.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Per E Kummervold, Javier de la Rosa, Freddy Wetjen, and Svein Arne Brygfjeld. 2021. Operationalizing a national digital library: The case for a Norwegian transformer model. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 20–29, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Per E Kummervold, Freddy Wetjen, and Javier de la Rosa. 2022. The Norwegian Colossal Corpus: A text corpus for training large Norwegian language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3852–3860, Marseille, France. European Language Resources Association.

Andrey Kutuzov, Jeremy Barnes, Erik Velldal, Lilja Øvrelid, and Stephan Oepen. 2021. Large-scale contextualised language modelling for Norwegian. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 30–40, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, page II–1188–II–1196. JMLR.org.

Yucheng Li, Frank Guerin, and Chenghua Lin. 2024. An open source data contamination report for large language models. In *Proceedings of the 2nd Workshop on Mathematical and Empirical Understanding of Foundation Models at ICLR 2024*.

Peng Liu, Lemei Zhang, Terje Farup, Even W. Lauvrak, Jon Espen Ingvaldsen, Simen Eide, Jon Atle Gulla, and Zhirong Yang. 2024. NLEBench+NorGLM: A comprehensive empirical analysis and benchmark

dataset for generative language models in Norwegian. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5543–5560, Miami, Florida, USA. Association for Computational Linguistics.

Anton Lozhkov, Raymond Li, Loubna Ben Allal, Federico Cassano, Joel Lamy-Poirier, Nouamane Tazi, Ao Tang, Dmytro Pykhtar, Jiawei Liu, Yuxiang Wei, Tianyang Liu, Max Tian, Denis Kocetkov, Arthur Zucker, Younes Belkada, Zijian Wang, Qian Liu, Dmitry Abulkhanov, Indraneil Paul, Zhuang Li, Wen-Ding Li, Megan Risdal, Jia Li, Jian Zhu, Terry Yue Zhuo, Evgenii Zheltonozhskii, Nii Osae Osae Dade, Wenhao Yu, Lucas Krauß, Naman Jain, Yixuan Su, Xuanli He, Manan Dey, Edoardo Abati, Yekun Chai, Niklas Muennighoff, Xiangru Tang, Muhtasham Oblokulov, Christopher Akiki, Marc Marone, Chenghao Mou, Mayank Mishra, Alex Gu, Binyuan Hui, Tri Dao, Armel Zebaze, Olivier Dehaene, Nicolas Patry, Canwen Xu, Julian McAuley, Han Hu, Torsten Scholak, Sebastien Paquet, Jennifer Robinson, Carolyn Jane Anderson, Nicolas Chapados, Mostofa Patwary, Nima Tajbakhsh, Yacine Jernite, Carlos Muñoz Ferrandis, Lingming Zhang, Sean Hughes, Thomas Wolf, Arjun Guha, Leandro von Werra, and Harm de Vries. 2024. StarCoder 2 and The Stack v2: The next generation.

Kevin Madigan. 2024. Mid-year review: AI lawsuit developments in 2024. Accessed: 2024-10-07.

Vladislav Mikhailov, Petter Mæhlum, Victoria Ovedie Chruickshank Langø, Erik Velldal, and Lilja Øvrelid. 2025. A collection of question answering datasets for Norwegian. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, Tallinn, Estonia.

Niklas Muennighoff, Alexander M Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. 2023. Scaling data-constrained language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 50358–50376.

Nasjonalbiblioteket. 2024. Mímir-prosjektet: Evaluering av virkningen av opphavsrettsbeskyttet materiale på generative store språkmodeller for norske språk. Technical Report. Accessed: 2024-10-26.

Nasjonalbiblioteket. 2024. Årsrapportar. Accessed: 2024-10-26.

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

Lilja Øvrelid, Petter Mæhlum, Jeremy Barnes, and Erik Velldal. 2020. A fine-grained sentiment dataset for Norwegian. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5025–5033, Marseille, France. European Language Resources Association.

Joe Panettieri. 2024. Generative AI lawsuits timeline: Legal cases vs. OpenAI, Microsoft, Anthropic, Nvidia, Intel and more. Accessed: 2024-10-07.

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

David Samuel, Andrey Kutuzov, Samia Touileb, Erik Velldal, Lilja Øvrelid, Egil Rønningstad, Elina Sigdel, and Anna Palatkina. 2023. NorBench – a benchmark for Norwegian language models. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 618–633, Tórshavn, Faroe Islands. University of Tartu Library.

David Samuel, Vladislav Mikhailov, Erik Velldal, Lilja Øvrelid, Lucas Georges Gabriel Charpentier, and Andrey Kutuzov. 2025. Small Languages, Big Models: A Study of Continual Training on Languages of Norway. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, Tallinn, Estonia.

Vinay Samuel, Yue Zhou, and Henry Peng Zou. 2024. Towards data contamination detection for modern large language models: Limitations, inconsistencies, and oracle challenges. In *Proceedings of the 29th International Conference on Computational Linguistics (COLING 2024)*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Jörg Tiedemann. 2020. The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Samia Touileb, Vladislav Mikhailov, Marie Ingeborg Kroka, Øvrelid Lilja, and Erik Velldal. 2025. Benchmarking abstractive summarisation: A dataset

of human-authored summaries of Norwegian news articles. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, Tallin, Estonia.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Erik Velldal, Lilja Øvrelid, Eivind Alexander Bergem, Cathrine Stadsnes, Samia Touileb, and Fredrik Jørgensen. 2018. NoReC: The Norwegian review corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Theresa M. Weisenberger, Diana C. Milton, Harrison A. Enright, and Jiwon Kim. 2024. Case tracker: Artificial intelligence, copyrights, and class actions. Accessed: 2024-10-07.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzm'an, Armand Joulin, and Edouard Grave. 2019. CCNet: Extracting high quality monolingual datasets from web crawl data. In *International Conference on Language Resources and Evaluation*.

Cheng Xu, Shuhao Guan, Derek Greene, and M-Tahar Kechadi. 2024. Benchmark data contamination of large language models: A survey. In *Proceedings of the 1st Workshop on Data Contamination (CONDA) at ACL 2024*.

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

Appendices

A Legal Cases

- **Bartz v. Anthropic PBC**, No. 3:24-cv-05417 (N.D. Cal. Aug. 19, 2024)
- **The Ctr. for Investigative Reporting v. OpenAI, Inc.**, No. 1:24-cv-04872 (S.D.N.Y. Jun. 27, 2024)
- **UMG Recordings, Inc. v. Uncharted Labs, LLC**, No. 1:24-cv-04777 (S.D.N.Y. Jun. 24, 2024)
- **UMG Recordings, Inc. v. Suo, Inc.**, No. 1:24-cv-11611 (D. Mass. Jun. 24, 2024)
- **J. L. v. Alphabet, Inc.**, No. 3:23-cv-03440, 2024 WL 3282528 (N.D. Cal. June 6, 2024)
- **In re OpenAI ChatGPT Litigation**, No. 3:23-cv-03223, 2024 WL 2044625 (N.D. Cal. May 7, 2024)
- **Makkai v. Databricks, Inc.**, No. 4:24-cv-02653 (N.D. Cal. May 2, 2024)
- **Dubus v. NVIDIA Corp.**, No. 3:24-cv-02655 (N.D. Cal. May 2, 2024)
- **Daily News LP v. Microsoft Corp.**, No. 1:24-cv-03285 (S.D.N.Y. Apr. 30, 2024)
- **Zhang v. Google LLC**, No. 3:24-cv-02531 (N.D. Cal. Apr. 26, 2024)
- **Nazemian v. NVIDIA Corp.**, No. 4:24-cv-01454 (N.D. Cal. Mar. 8, 2024)
- **The Intercept Media, Inc. v. OpenAI, Inc.**, No. 1:24-cv-01515 (S.D.N.Y. Feb. 28, 2024)
- **Raw Story Media, Inc. v. OpenAI Inc.**, No. 1:24-cv-01514 (S.D.N.Y. Feb. 28, 2024)
- **Tremblay v. OpenAI, Inc.**, No. 3:23-cv-03233, 2024 WL 557720 (N.D. Cal. Feb. 12, 2024)
- **Universal Music Group v. Anthropic** (February 2024)
- **The New York Times v. OpenAI & Microsoft** (December 2023)
- **Kadrey v. Meta Platforms, Inc.**, No. 3:23-cv-03417, 2023 WL 10673221 (N.D. Cal. Dec. 1, 2023)
- **Alter v. OpenAI Inc.**, No. 1:23-cv-10211 (S.D.N.Y. Nov. 21, 2023)
- **Andersen v. Stability AI Ltd.**, No. 3:23-cv-00201, 2023 WL 7132064 (N.D. Cal. Oct. 30, 2023)
- **Concord Music Group, Inc. v. Anthropic PBC**, No. 3:23-cv-01092 (M.D. Tenn. Oct. 18, 2023)
- **Huckabee v. Meta Platforms, Inc.**, No. 3:23-cv-06663 (N.D. Cal. Oct. 17, 2023)
- **Authors Guild v. OpenAI Inc.**, No. 1:23-cv-08292 (S.D.N.Y. Sept. 18, 2023)
- **Silverman v. OpenAI Inc.**, No. 3:23-cv-03416 (N.D. Cal. July 7, 2023).
- **Thaler v. Perlmutter**, 687 F. Supp. 3d 140 (D.D.C. 2023)
- **Doe 1 v. Github, Inc.**, No. 4:22-cv-06823, 2023 WL 3449131 (N.D. Cal. May 11, 2023)
- **Getty Images (US), Inc. v. Stability AI, Inc.**, No. 1:23-cv-00135 (D. Del. Feb. 3, 2023)

B Sampling

We built three custom perplexity models for specific Norwegian domains that proved too divergent from Wikipedia: books, newspapers, and government documents. These perplexity models were used to score each document in the datasets. Based on their perplexity scores, the documents were further divided into three segments corresponding to their quartile distribution. Documents with scores below the first quartile were classified as "good", those between Q_1 and Q_3 as "medium", and those above Q_3 were considered "bad". The documents in each segment were randomized. While the intention was to train all models on progressively better data, starting from "bad" segment, then "medium" and finally the "good" segment, we never got around to test whether this approach would result in better performing models.

Moreover, from the clean and deduplicated sources, we sub-sampled each non-Norwegian language at an specific sampling ratio until achieving the proportion of documents shown in Figure 5. Pseudo-code for the algorithm used to subsample is shown in Algorithm 1.[20] We also discovered that a good amount of documents were misclassified by the fastText language identifier (Joulin et al., 2016).

Language	Sampling ratio	Final ratio
Bokmål	100.00%	35.74%
Danish	43.00%	8.01%
English	81.00%	4.53%
Icelandic	100.00%	1.31%
Nynorsk	100.00%	2.02%
Swedish	15.40%	4.46%
Code	62.00%	4.53%

Table 5: Percentage of documents kept from the clean and deduplicated sources and the final proportion of documents in each language present in the final dataset. Code was considered its own language when sampling.

C Sources

Source	Raw	Clean	extended	base
Books	3.7B	2.5B	1.9B	1.9B
CulturaX	52.7B	52.1B	21.8B	16.9B
Digimanus	9.6M	4.6M	3.4M	3.3M
Evaluerings- rapport	76.7M	68.6M	61.2M	61.5M
HPTL v1.2	35.5B	34.1B	14.9B	11.3B
LovData	57.1M	57.1M	53.7M	54.8M
Målfrid	7.5B	1.9B	1.7B	1.7B
Newspapers	4.6B	3.6B	3.2B	3.3B
Parlamint	170.3M	84.4M	83.4M	83.3M
PG19	2.0B	1.9B	1.4B	428.6M
StarCoder	19.7B	9.8B	7.1B	3.4B
Wikipedia	4.0B	3.9B	2.8B	996.2M
Books (restricted)	21.7B	20.0B	18.1B	0
Newspapers (restricted)	14.3B	9.8B	9.1B	0
Total	**166.1B**	**139.8B**	**82.1B**	**40.1B**

Table 6: Number of words (comma separated) per source at the start of the data pipeline (raw count), after cleaning, and in the **extended** and **base** datasets.

[20] https://huggingface.co/mimir-project/mimir-perplexity

Algorithm 1 Sub-sampling Dataset Based on Perplexity Distribution

Input: Dataset D with perplexity distribution, target sampling ratio R
Output: Sub-sampled dataset D'
procedure SUBSAMPLE(D, R)
    Compute the quartile values q_1 and q_3 from the perplexity distribution of D
    Define an initial Gaussian PDF with mean $\mu = (q_1 + q_3)/2$ and standard deviation σ such that q_1 and q_3 align with the corresponding positions in the Gaussian curve
    Compute the histogram H of perplexity values from D
    Combine H with the Gaussian weights to estimate the initial sampling ratio R_0
    Compute the normalization factor N such that $R_0 \times N = R$
    while Error in central quartile probabilities exceeds tolerance **do**
        Adjust the parameters μ and σ of the Gaussian curve to minimize the error in the desired probabilities within the central quartiles $[q_1, q_3]$
        Update the normalization factor N to match the target ratio R
    end while
    for each sample s in D **do**
        Compute the perplexity p_s of sample s
        Estimate the probability $P(s)$ of retaining sample s based on the normalized Gaussian PDF
        if $P(s) \geq$ random threshold **then**
            Retain s in the sub-sampled dataset D'
        end if
    end for
end procedure
return D'

D Hyperparameters

Hyperparameter	Core Models	Domain-Tuned Models	Instruction-tuned Models
Model size	7B	7B	7B
Hidden layers	32	32	32
Attention heads	32	32	32
Hidden size	4096	4096	4096
Intermediate size	14336	14336	14336
Max position embeddings	2048	2048	2048
Key-value heads	8	8	8
Sliding window	4096	4096	4096
Precision	`bfloat16`	`bfloat16`	`bfloat16`
Optimizer	AdamW	AdamW	AdamW
Optimizer parameters	$\beta_1 = 0.9, \beta_2 = 0.95, \epsilon = 10^{-8}$	$\beta_1 = 0.9, \beta_2 = 0.95, \epsilon = 10^{-8}$	$\beta_1 = 0.9, \beta_2 = 0.95, \epsilon = 10^{-8}$
Global batch size	4M (2048×2048) tokens	4M (2048×2048) tokens	512 seqs
Initial/final learning rate	3.0×10^{-4} / 3.0×10^{-5}	3.0×10^{-5} / 3.0×10^{-6}	3.0×10^{-6} / 3.0×10^{-7}
Vocabulary size	32768	32768	32768
Training steps	64k	10k	4 epochs
Dropout	0	0	0
Warm-up steps	2000	200	20
Weight decay	0.1	0.1	0.1
Checkpoints	Every 1000 steps	Every 1000 steps	Every 1 epoch
Shuffle	Shuffle after each epoch	Shuffle after each epoch	Shuffle after each epoch

Table 7: Hyperparameters for the Mímir model set.

E Percentage Gains

Figure 5 illustrates the percentage gains of each domain-tuned model with respect to the performance of the **base** model, per higher level skill. Training on different materials shows distinct trade-offs: newspaper data excels at Translation (27.20% gain) and Norwegian Language (51.92%), while fiction books improve Variation & Readability (7.83%). Combining books and newspapers often yields balanced improvements, though most configurations struggle with Reading Comprehension and Translation. The **extended** configuration, which supplements books and newspapers with Internet data, shows strong all-around performance, particularly in Summarization (26.37%) and World Knowledge (5.60%).

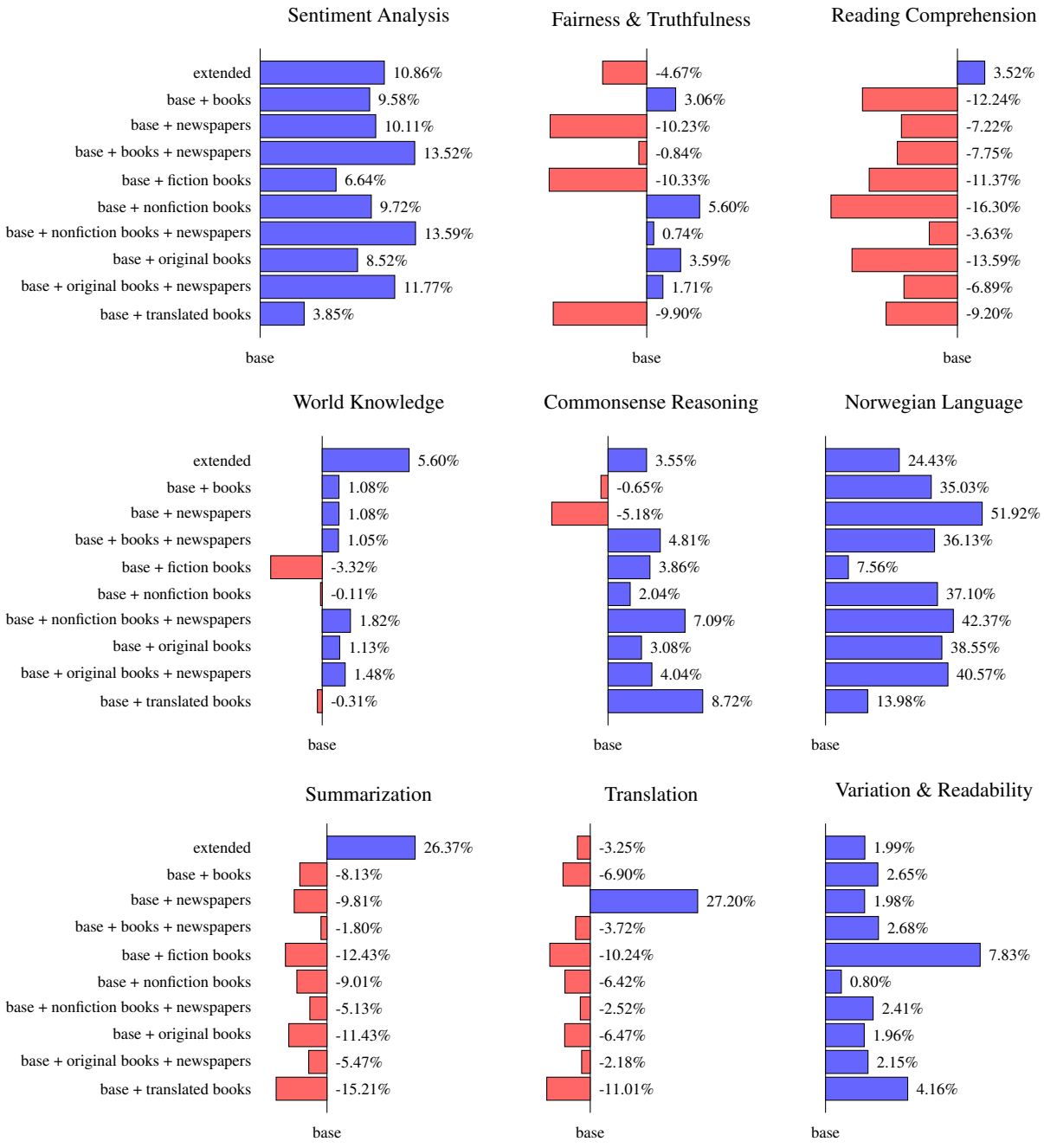

Figure 5: Percentage gains over the performance of the **base** model per skill.

F Evaluation Scores

Model	SA	FT	RC	WK	CR	NL	S	T	VR	Score
Core Models										
base	69.54	53.51	38.04	41.10	39.85	38.28	32.45	41.55	59.66	413.98
extended	77.09	51.01	39.38	43.40	41.26	47.64	**41.00**	40.20	60.85	441.83
base (warm)	88.17	51.28	52.48	53.27	48.02	49.51	35.88	41.14	60.52	480.28
extended (warm)	86.57	54.64	52.79	51.51	49.25	48.48	36.14	42.48	61.12	482.98
Domain Tuned Models										
base + books	76.20	55.15	33.38	41.54	39.59	51.69	29.81	38.68	61.24	427.30
base + newspapers	76.57	48.04	35.29	41.55	37.79	58.16	29.26	52.85	60.85	440.35
base + books + newspapers	78.94	53.06	35.09	41.53	41.77	52.11	31.86	40.00	61.27	435.64
base + fiction books	74.16	47.99	33.71	39.74	41.39	41.18	28.41	37.29	**64.33**	408.20
base + nonfiction books	76.30	56.51	31.84	41.06	40.66	52.48	29.52	38.88	60.14	427.40
base + nonfiction books + newspapers	78.99	53.91	36.66	41.85	42.68	54.50	30.78	40.50	61.10	440.97
base + original books	75.46	55.43	32.87	41.56	41.08	53.04	28.74	38.86	60.83	427.87
base + original books + newspapers	77.72	54.43	35.42	41.71	41.46	53.81	30.67	40.64	60.95	436.81
base + translated books	72.22	48.21	34.54	40.97	43.33	43.63	27.51	36.97	62.15	409.54
Instruction Fine Tuned Models										
base (warm) *instruct*	87.83	53.70	50.33	54.98	49.42	**59.53**	38.36	49.75	59.46	503.36
extended (warm) *instruct*	**89.81**	57.80	51.69	53.09	**49.76**	55.91	37.75	47.72	60.35	**503.89**
base *instruct*	69.45	50.59	36.27	41.18	42.06	53.53	35.14	**58.83**	58.35	445.39
extended *instruct*	78.90	46.10	38.68	44.32	43.57	56.46	36.40	54.64	59.29	458.36
Mistral 7B v0.1	88.41	**64.93**	**56.68**	**58.86**	36.01	31.49	10.09	41.55	59.99	448.01

Table 8: Detailed scores across all skills for each model configuration. Abbreviations: SA = Sentiment Analysis, FT = Fairness & Truthfulness, RC = Reading Comprehension, WK = World Knowledge, CR = Commonsense Reasoning, NL = Norwegian Language, S = Summarization, T = Translation, VR = Variation & Readability. Best overall scores per skill in **bold**. Best score per skill and model group underlined. Mistral 7B v0.1 also added for reference.