The Mímir Project

Evaluating the Impact of Copyrighted Materials on Generative Large Language Models for Norwegian Languages

Version: 1.1
Date: October 24, 2024

Introduction

The Mímir Project is an initiative by the Norwegian government that aims to assess the significance and influence of copyrighted materials in the development and performance of generative large language models (LLMs) tailored to the Norwegian languages. This collaborative effort involves three leading institutions from different regions of the country: the National Library of Norway (NB), the University of Oslo (UiO), and the Norwegian University of Science and Technology (NTNU); each contributing unique expertise in language technology, corpus curation, model training, copyright law, and computational linguistics. Additionally, the project has been supported with computational resources provided by Sigma2. The ultimate goal of the project is to gather empirical evidence that will inform the formulation of a compensation scheme for authors whose works are utilized by these advanced artificial intelligence (AI) systems, ensuring that intellectual property rights are respected and adequately compensated.

Background

The advent of LLMs has revolutionized natural language processing (NLP), enabling machines to generate, understand, and interact using human languages with unprecedented accuracy and fluency. However, these models require vast amounts of textual data to train, often sourced from a wide array of digital repositories that include copyrighted materials. This raises critical legal and ethical questions regarding the use of such content without explicit permission from authors, as well as the economic implications for the creative industry.

In November of 2023, the right-holders organization of Norway contacted the government demanding compensation for the use of their material in the training of generative AI. In order to

support these negotiations with data, the government 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 for Science and Technology under the umbrella of the so-called Mímir Project, a name chosen after a figure in Norse mythology renowned for his knowledge and wisdom. The Mímir Project addresses these concerns by systematically evaluating the impact that copyrighted texts may have on the capabilities of LLMs for Norwegian. In January of 2024 the consortium was formed and responsibilities were assigned. A final report documenting the process and describing the conclusions had to be delivered by July 1st, 2024, leaving exactly 6 months to organize and curate data, build datasets, design experiments, train models, create evaluation benchmarks, evaluate models, and process and digest the results.

Methodology

The project's methodology involves a comprehensive analysis that spans several stages. Initially, a diverse corpus of 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, question-answering and sentiment analysis, the project seeks to quantify the specific contributions of copyrighted materials to the overall model quality.

To ensure robustness and reliability, the evaluation framework focuses on both foundational model generation and linguistically-inspired metrics. Quantitative measures include traditional NLP metrics like accuracy, F1, BLEU scores, and ROUGE scores, which provide objective 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.

Collections and Datasets

We adapted methods from the Norwegian Colossal Corpus (NCC)[1] for the building of datasets, saving time despite occasional limitations. Our data sources include the open internet (e.g., Wikipedia, crawls from the High Performance Language Technology project[2]), partner content (e.g., NRK, Amedia, Schibsted, TV2), and our internal safety store (e.g., copyrighted newspapers and books). Data is processed through a tailored cleaning procedure into a standardized format, ensuring uniform functionality, after which deduplication follows to ensure unique examples. Each data entry has enough metadata to guide language model training, balancing Norwegian text to prevent other languages from overshadowing it. This follows our experience with the previous NCC extended corpus. We maintain two types of corpora:

[1] The Norwegian Colossal Corpus: A Text Corpus for Training Large Norwegian Language Models (Kummervold et al., LREC 2022)
[2] https://hplt-project.org/

complete corpora for pretraining (see Table 1) and different subsets and configurations (delta corpora) for fine-tuning on restricted material (see Table 2). The complete corpora come in **base** and **extended** sizes, with the base not containing copyrighted materials under the protection of Norwegian law[3], while extended contains everything, thus being nearly double in size. All corpora are cleaned, deduplicated, and language-balanced. An extra dataset for so-called instruction fine-tuning was created which contains 5000 instruction pairs.

Complete Corpus	Documents	Words
base	60,182,586	40,122,626,817
extended	125,285,547	82,149,281,266

Table 1. Number of words and tokens per complete pre-training corpus

Delta Corpus	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 words and tokens per delta fine-tuning corpus, which only contain in-copyright materials.

Training

We trained 17 models of 7 billion parameters[4] following the Mistral architecture for a total of 270,000 GPU-hours. The infrastructure included the LUMI supercomputer, the Idunn cluster, and Google TPUs through their TPU Research Cloud program. Training was conducted in two

[3] Under Språkteknologiformål, the National Library has over the years entered into agreements with some newspapers for the use of their text for language technology purposes. Although somewhat different to each other, they all tend to consider the linguistic structure, i.e. the individual words and the way they are put together, of the utmost importance over its intellectual content.
[4] A model with 7 billion parameters is still 25 times smaller than OpenAI's GPT-3, the predecessor of ChatGPT.

phases. First, in order to measure the impact of copyrighted material as a whole and the impact that copyrighted materials might have in a realistic post-training scenario, we conducted pre-training on base and extended both from scratch and from the pre-existing weights (warm) of Mistral 7B v0.1. These 4 core models were trained for the same amount of total words (64,000 steps of ~2.5 million words) using identical setups. Second, to further isolate the effect of different ablations of the copyrighted materials, we continuously trained the base model from scratch for an extra 10,000 steps on each of the 9 delta corpora. The core models were also further-trained on the instruction corpus for 4 iterations to evaluate their performance on downstream tasks after fine-tuning.

Evaluation

Evaluating generative language models is far from a solved problem, in particular for Norwegian, where there were few existing resources at the outset of the project. Through a dedicated and intense effort in the context of the project, we compiled a set of 28 of the most common tasks in NLP, encompassing a range of different metrics to assess the performance of each of the models. These tasks can be grouped into 9 higher level skills:

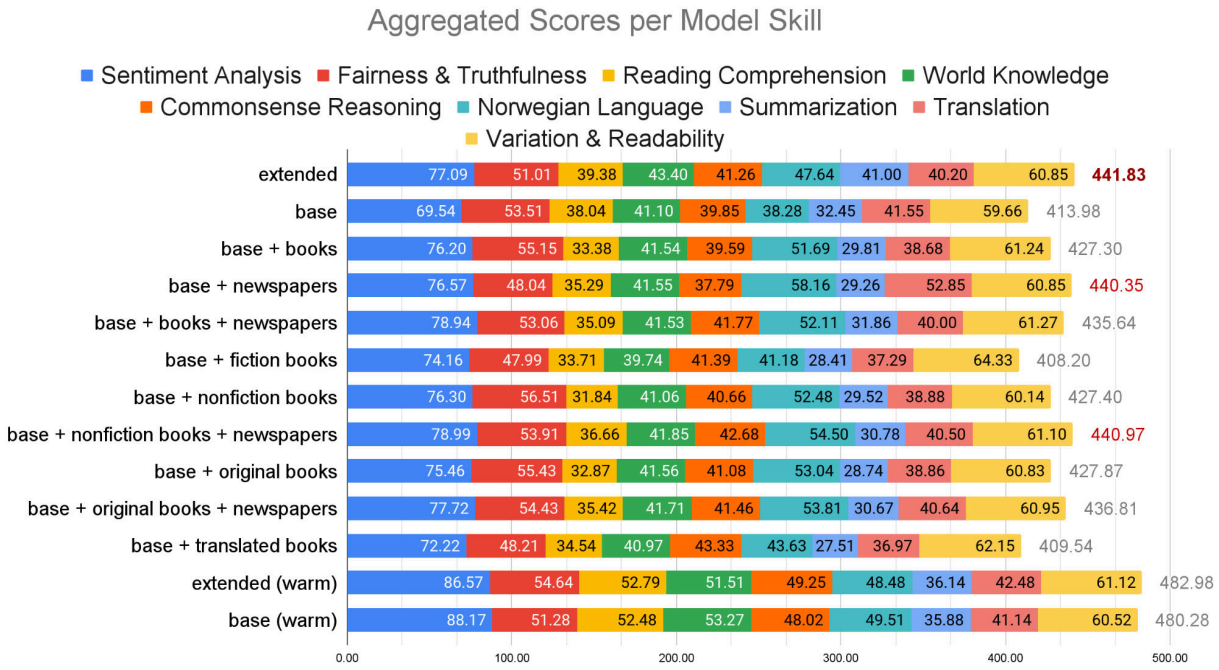
- **Sentiment Analysis**, which involves determining the emotional tone behind a series of words. It's used to identify the sentiment expressed in a piece of text, which could be positive, negative, or neutral. For example, in customer reviews or social media posts, sentiment analysis helps to gauge public opinion or satisfaction.
- **Fairness & Truthfulness**. Fairness in language models refers to the absence of bias in the model's predictions and outputs. Evaluating fairness ensures that the model does not favor or discriminate against particular groups based on attributes like race, gender, or ethnicity. Truthfulness involves the accuracy and reliability of the information produced by the model, ensuring it generates factual and verifiable content.
- **Reading Comprehension**, which measures a model's ability to understand and interpret text. It involves answering questions about a given passage, summarizing content, or explaining the meaning of specific phrases or sentences. This skill evaluates how well the model grasps the context and details in the text.
- **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 evaluation checks if the model can correctly answer questions or provide information based on real-world knowledge.
- **Commonsense Reasoning**, which involves the model's ability to make logical inferences based on everyday knowledge and understanding of the world. This skill tests whether the model can reason about situations that require practical, everyday knowledge that people take for granted.
- **Norwegian Language** evaluation focuses on the model's understanding and generation of text in Norwegian, specifically its grammar, structure, and sentence construction. This skill is important for assessing how well the model handles Norwegian languages and their specific syntactic rules.
- **Summarization**, which measures a model's ability 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.

- **Translation**, which evaluates how accurately a language model can convert text from one language to another while preserving the meaning, tone, and context. This skill is important for multilingual applications and for users who need content accessible in multiple languages.
- **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.

Results

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 k-shot configuration[5] for each task and the best score for each of the different prompts used to elicit a proper continuation from the models. As shown in Figure 1, the total scores across all evaluated skills, averaged by task for each model, demonstrate the cumulative impact of including copyrighted materials in the training process. Models trained with a mix of copyrighted and non-copyrighted content generally exhibited superior performance compared to those trained exclusively on non-copyrighted data. This indicates that copyrighted materials, likely due to their higher quality and curated nature, tend to contribute positively to the models' overall efficacy. However, the performance difference for training on the base versus extended corpus is less pronounced for the warm-started models than for those trained from scratch.



[5] Few-shot or k-shot learning is a machine learning technique where models learn to recognize patterns and make predictions based on a very small number of annotated samples (k). In our benchmark, we experimented with values of k of 0 (no annotated samples at all), 1, 4, and 16.

Figure 1. Summary of the total scores (sum) across all skills averaged by task for each model. Best scores among not warm-started models in red, best overall not warm-started in bold red.

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 3. 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.

As shown in Table 3 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. These insights suggest that the base (warm-started) configuration is optimal for tasks requiring sentiment analysis and world knowledge, while the extended (warm-started) configuration is preferable for tasks involving detailed language analysis and comprehension. Furthermore, it indicates that we could leverage the existing metadata available at the National Library to tailor subsets of the copyrighted material and build models that excel at specific tasks. However, the difference between the warm models is very small, but further testing is required to assess whether the difference is statistically significant.

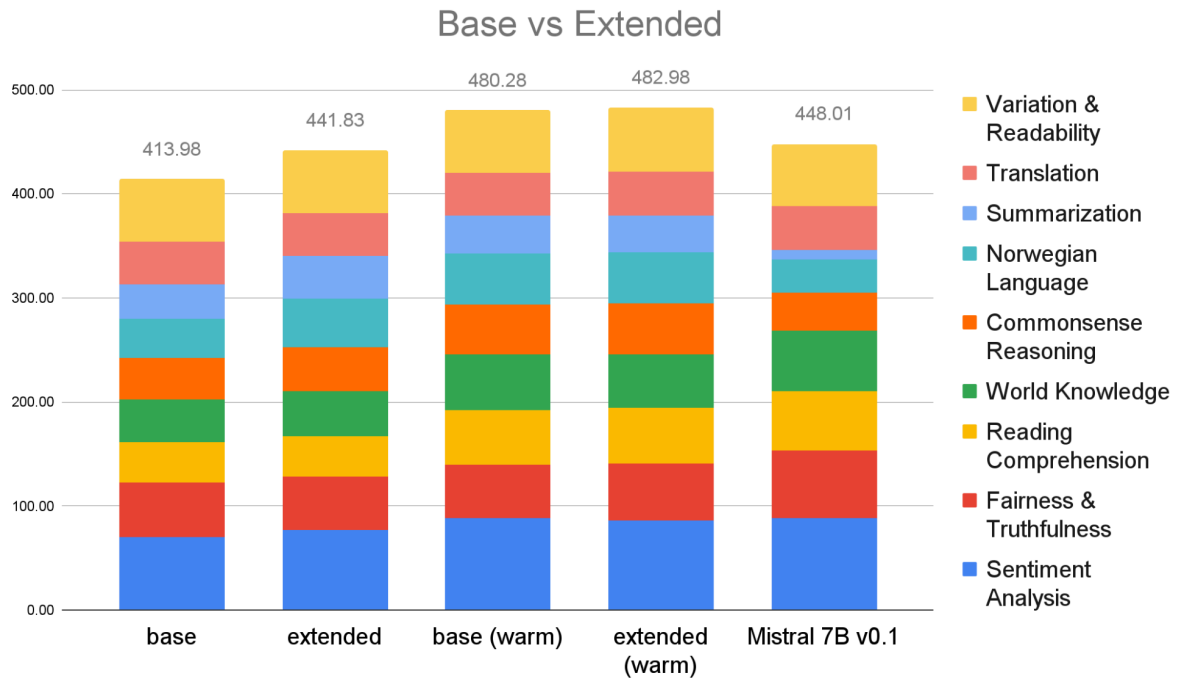


Figure 2. Cumulative score of the core pre-trained models under different regimes. Including original Mistral 7B v0.1 for reference.

For the delta configurations alone, Figure 3 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 (lower half of Figure 3), the addition of fiction books makes the model excel at generating more diverse texts while underperforming all others in grammar and punctuation; we call this the “Jon Fosse paradox”. These results highlight the varying impacts of different training data combinations on model performance.

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 pre-training data, while further training on instructions gives a consistent boost in performance.

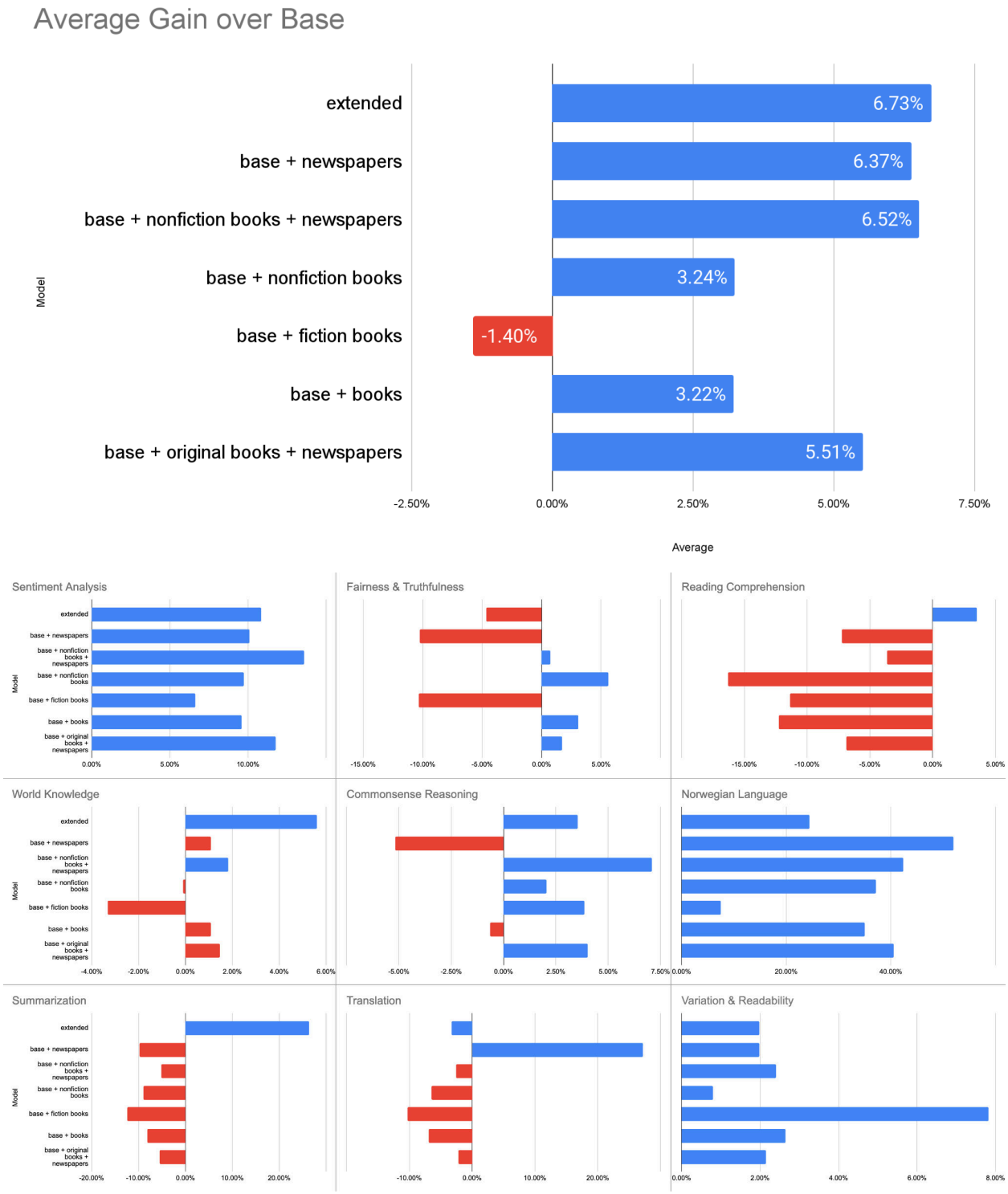


Figure 3. Average percentage gains overall and for each higher level skill over the performance of the base model. Negative results indicate a decrease in performance over base, a positive result a gain.

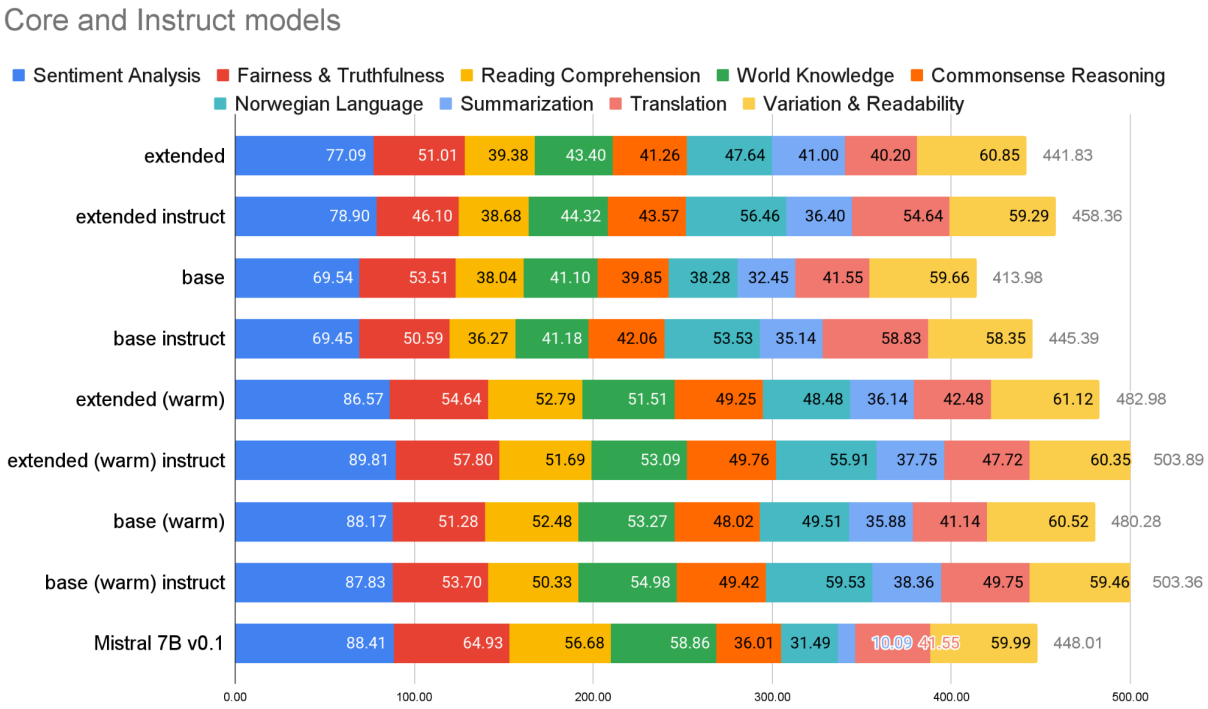


Figure 4. Total scores (sum) of all averaged scores per skill for the core models and their instruct versions. Including original Mistral 7B v0.1 for reference.

Discussion

One of the central hypotheses of the Mímir Project is that copyrighted materials, due to their high quality and curated nature, enhance the linguistic richness and diversity captured by LLMs. The project did not aim to produce the best performing model for Norwegian, but to identify specific areas where the inclusion of restricted material leads to substantial improvements. The key observations follow:

1. There is empirical evidence supporting the thesis that copyrighted material improves model performance. To a large extent, this effect seems to be mediated by non-fictional content.
2. Warm-starting from a pre-trained model like Mistral appears to give overall best results, reducing the impact of copyrighted materials on performance. However, while the warm-started models seem to perform best, these models are not fully auditable, as their pre-training data is unknown. Also, for some tasks, warm-starting from a predominantly English-trained model was found to be detrimental.
3. Instruction fine-tuning consistently increases the performance of all core models, regardless of their pre-training data.

Limitations and Future Work

The Mímir Project was completed within six months, covering all stages from ideation and consortium formation to evaluation and report writing. Consequently, time constraints significantly influenced every decision. With more time, we could have designed the

experiments differently. For instance, we opted to create delta datasets and conduct domain-specific training to properly isolate their effects. The delta datasets were upsampled to match the word count during training, resulting in over five iterations for smaller datasets like fiction or translated books, but only about two for larger datasets like newspapers. It remains uncertain if downsampling would show the same trends. Since this method amplifies the effects of different deltas, another alternative approach would be to extract the different deltas from the extended corpus one at a time and train on each resulting dataset from scratch. Comparing these strategies would require approximately five times the computational resources.

Other interesting questions remain unanswered, such as the effect of the pre-existing data mixture of the warm-started models, the model architecture, or their size. Regarding evaluation, there has been considerable effort put into developing novel evaluation resources for Norwegian in the project. Even so, we still lack proper ways to assess the creative aspects of language models for Norwegian, oftentimes requiring human evaluation, potentially biasing our evaluation suite towards more commercially valuable and automation-prone tasks.

Final Remarks

Investigating the impact of copyrighted material in LLMs is a novel and underexplored research avenue. As such, the outcomes of the Mímir Project are expected to have ramifications for both national and international contexts. For Norway, the project provides a critical evidence base that can inform national copyright policies and support the creation of a compensation scheme tailored to the digital age. Internationally, the findings contribute to the ongoing discourse on AI ethics and copyright law, offering a model that other countries can adapt and implement.

In conclusion, the Mímir Project represents a pioneering effort to empirically assess the impact of copyrighted materials on generative LLMs for Norwegian languages. By bridging the gap between technology and intellectual property rights, the project aims to foster a more equitable and innovative future for AI development. The insights gained from this initiative, along with the release of the generated artifacts (datasets, models, and benchmarks), will not only enhance our understanding of LLMs but also pave the way for more sustainable and fair practices in the use of copyrighted content in AI training.

Signatories

National Library of Norway: Javier de la Rosa, Freddy Wetjen, Rolv-Arild Braaten, Magnus Breeder, Tita Enstad, Wilfred Østgulen, Svein Arne Brygfjeld, Aslak Sira Myhre.
University of Oslo: Vladislav Mikhailov, David Samuel, Petter Mæhlum, Liljia Øvrelid, Andrey Kutuzov, Erik Velldal, Petter Mæhlum, Stephan Oepen.
Norwegian University of Science and Technology: Peng Liu, Lemei Zhang, Jon Atle Gulla.