



Mímir Project

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

Javier de la Rosa
versae@nb.no

• AI-lab
• National Library of Norway



Background

- Revolution of LLMs in NLP
- Need for vast amounts of textual data
- Legal and ethical questions regarding copyrighted materials

Timeline

Nov 2022	Jan 2023
<div>ChatGPT Release</div> <div>First successful public-facing AI chat model</div>	<div>GPT-4 Release</div> <div>Human-like capabilities</div>

Silverman v. OpenAI (July 2023): Authors, including Sarah Silverman, filed lawsuits alleging their books were used to train AI models without consent (Reuters, July 2023).

The New York Times v. OpenAI & Microsoft (December 2023): The New York Times claimed that AI tools diverted web traffic from their platforms and used articles without proper authorization (The Wall Street Journal, December 2023).

Universal Music Group v. Anthropic (February 2024): Music publishers, including Universal Music Group, accused Anthropic of using copyrighted song lyrics to train AI models without permission, marking one of the key legal challenges involving music rights (The Hollywood Reporter, February 2024).

Class Actions and Industry Response: Growing numbers of lawsuits from authors, musicians, and publishers highlight increasing concerns about the impact of AI models on creative industries ([Baker Law, 2024](#)).

Timeline

Nov 2022

"The department hereby asks the National Library to initiate a coordinated research/development project to, if possible, examine the value of copyright-protected material in the training of Norwegian generative language models. Relevant Norwegian research environments will be invited to participate in the project. Authors' and publishers' organizations are invited to follow the project."

ChatGPT

"We ask that the National Library, on the basis of the results from the research project, assess the basis for a possible compensation scheme for Norwegian rights holders, and possibly prepare a proposal for such a scheme."

First successful public-facing AI chat model

Human-like capabilities

Norwegian Government

They demanded compensation for the use of their material

The Mandate from the Ministry of Culture

The aim of the assignment from Ministry of Culture is threefold:

1. Examining the value of copyright-protected material in training Norwegian generative language models
2. Assess the basis for a possible compensation scheme for Norwegian rights holders
3. Prepare proposals for a compensation scheme

Objective 1 is within the mandate of the project, while objectives 2 and 3 are outside this project.

Methodology

- Diverse corpus assembly
- Model training
- Evaluation tasks

Data

Mimir Base

- Open data set for free access without restrictions
- Newspapers and books in the public domain or by agreement
- Publications from the public sector
- Arranged data from the Language Bank, e.g. from the Web

Mimir Extended

- Internal dataset used by NB AI-lab Contains “Base” and in addition material under copyright

Data

Mimir Base

Turns out, there were restrictions!

- Open data set for free access without restrictions
- Newspapers and books in the public domain or by agreement
- Publications from the public sector
- Arranged data from the Language Bank, e.g. from the Web

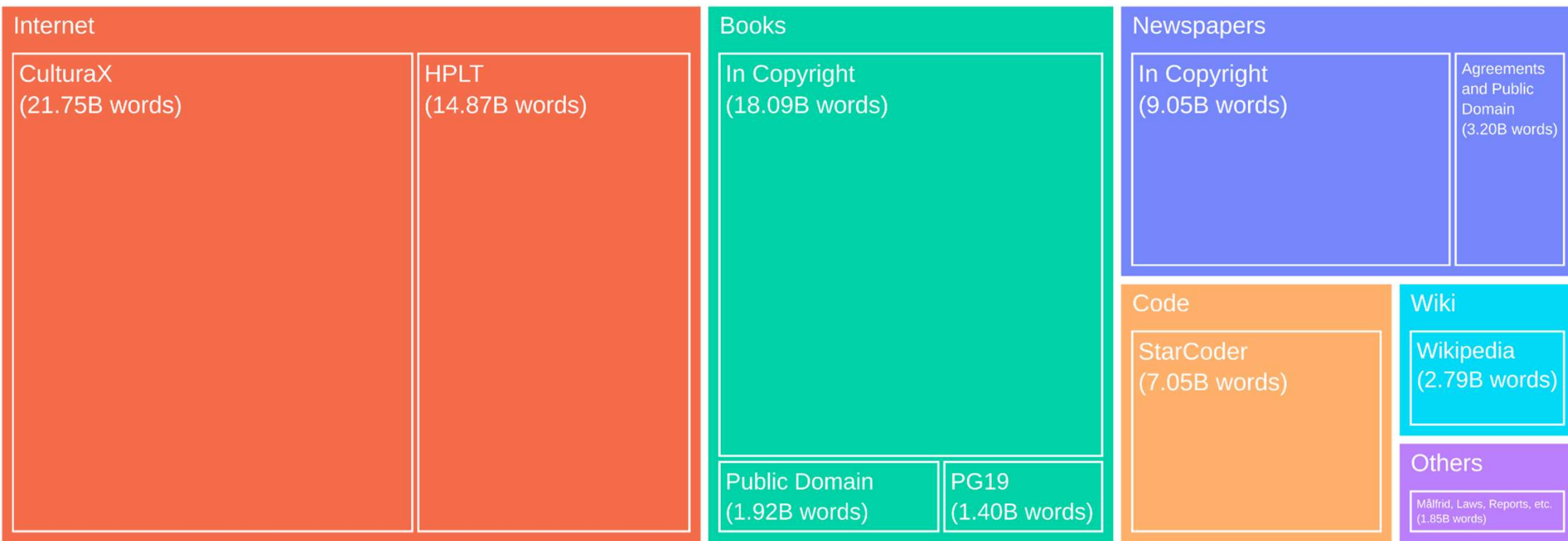
Mimir Extended

- Internal dataset used by NB AI-lab Contains “Base” and in addition material under copyright

Data

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

Mimir Extended



Data

Configurations	Dataset
mimir-base	mimir-base
mimir-extended	mimir-extended
mimir-base + all books	mimir-books
mimir-base + all other newspapers	mimir-newspapers
mimir-base + all books + all other newspapers	mimir-rightholders
mimir-base + all other newspapers + all nonfiction	mimir-factual
mimir-base + all nonfiction	mimir-nonfiction
mimir-base + all fiction	mimir-fiction
mimir-base + all books - all untranslated	mimir-translated
mimir-base + all books - all translated	mimir-untranslated
mimir-base + all other newspapers + all books - all translated	mimir-untranslated-withnewspapers

Data

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

Training

- Training 17 models of 7 billion parameters (25 times smaller than GPT-3).
Training took 270k GPU hours
- Infrastructure: LUMI supercomputer, Idunn cluster, Google TPUs
- Training phases: Pre-training, continuous training on delta corpora

Training

Status	Initialization	Data	Name
<input checked="" type="checkbox"/>	From scratch	mimir-base	mimir-mistral-7b-base-scratch
<input checked="" type="checkbox"/>	From scratch	mimir-extended	mimir-mistral-7b-extended-scratch
<input checked="" type="checkbox"/>	Pre-existing	mimir-base	mimir-mistral-7b-base
<input checked="" type="checkbox"/>	Pre-existing	mimir-extended	mimir-mistral-7b-extended
<input checked="" type="checkbox"/>	mimir-mistral-7b-base-scratch	mimir-fiction	mimir-7b-fiction
<input checked="" type="checkbox"/>	mimir-mistral-7b-base-scratch	mimir-nonfiction	mimir-7b-nonfiction
<input checked="" type="checkbox"/>	mimir-mistral-7b-base-scratch	mimir-factual	mimir-7b-factual
<input checked="" type="checkbox"/>	mimir-mistral-7b-base-scratch	mimir-newspapers	mimir-7b-newspapers
<input checked="" type="checkbox"/>	mimir-mistral-7b-base-scratch	mimir-books	mimir-7b-books
<input checked="" type="checkbox"/>	mimir-mistral-7b-base-scratch	mimir-rightholders	mimir-7b-rightholders
<input checked="" type="checkbox"/>	mimir-mistral-7b-base-scratch	mimir-untranslated-withnewspapers	mimir-7b-untranslated-withnewspapers
<input checked="" type="checkbox"/>	mimir-mistral-7b-base-scratch	mimir-untranslated	mimir-7b-untranslated
<input checked="" type="checkbox"/>	mimir-mistral-7b-base-scratch	mimir-translated	mimir-7b-translated

Evaluation

Dataset	nb/nn	Train	Test	# Prompts	Type	Skill	Performance Metrics	Domain	Authors/Contributors
Peer-reviewed Norwegian benchmarks and datasets									
NoReC_sentence	nb	3.89k	583	5	NLU	Sentiment analysis	F1 _a	Misc.	UiO
NoReC_document	nb	23.4k	2.9k	5	NLU	Sentiment analysis	F1 _a	Misc.	UiO
NorQuAD	nb	3.81k	472	5	NLU	Reading Comprehension	F1/Exact match	News, Wikipedia	UiO
Belebele	nb	✗	900	5	NLU	Reading Comprehension	Accuracy	Web	Meta
ASK-GEC	nb	36.4k	4.75k	5	NLG	Norwegian language	ERRANT	Exams, ASK	NB, UiO
Tatoeba (en ↔ nb)	nb	5.2k	4.5k	8	NLG	Machine translation	chrF, BLEU, BERTScore	Misc.	tatoeba.org
Tatoeba (en ↔ nn)	nn	504	459	8	NLG	Machine translation	chrF, BLEU, BERTScore	Misc.	tatoeba.org
Tatoeba (nn ↔ nb)	nb/nn	✗	465	8	NLG	Machine translation	chrF, BLEU, BERTScore	Misc.	tatoeba.org
English benchmarks and datasets adapted to Norwegian									
NorOpenBookQA	nb	3k	387	5	NLU	World knowledge	Accuracy	Exams, Misc.	UiO
NorOpenBookQA	nn	173	90	5	NLU	World knowledge	Accuracy	Exams, Misc.	UiO
NorCommonsenseQA	nb	✗	693	5	NLU	Commonsense reasoning	Accuracy	Misc.	UiO
NorCommonsenseQA	nn	✗	95	5	NLU	Commonsense reasoning	Accuracy	Misc.	UiO
NorTruthfulQA (multiple choice)	nb	✗	209	5	Fairness	Fairness & Truthfulness	Accuracy	Misc.	UiO
NorTruthfulQA (multiple choice)	nn	✗	57	5	Fairness	Fairness & Truthfulness	Accuracy	Misc.	UiO
NorTruthfulQA (generation)	nb	✗	281	5	Fairness	Fairness & Truthfulness	BLEU, rougeL	Misc.	UiO
Norwegian datasets being created from scratch									
Media Futures	nb	✗	30	6	NLG	Text summarization	BLEU, rougeL, BERTScore	News	UiO, MediaFutures
Media Futures	nn	✗	30	6	NLG	Text summarization	BLEU, rougeL, BERTScore	News	UiO, MediaFutures
NorIdiom	nb	✗	402	5	NLG	Norwegian language	F1/Exact match	Misc.	NB
NorIdiom	nn	✗	401	5	NLG	Norwegian language	F1/Exact match	Misc.	NB
NCB	nb	✗	840	✗	NLU	Norwegian language	Accuracy	Misc.	UiO Faculty of Law
Mimir-bias	nb/nn	✗	2.47k	✗	Fairness	Fairness & Truthfulness	Bias score	Misc.	NTNU
NorInstruction	nb/nn	4.9k	50	✗	Instruction-tuning	Misc.	✗	Misc.	NTNU
External contributions									
NRK Quiz	nb	✗	3.6k	5	NLU	World knowledge	Accuracy	Misc.	UiO, NRK, Språkbanken
NRK Quiz	nn	✗	1.3k	5	NLU	World knowledge	Accuracy	Misc.	UiO, NRK, Språkbanken

Evaluation

- List of 28 common NLP tasks
- Grouped into 9 higher-level skills
 1. Sentiment Analysis
 2. Fairness & Truthfulness
 3. Reading Comprehension
 4. World Knowledge
 5. Commonsense Reasoning
 6. Norwegian Syntax
 7. Summarization
 8. Translation
 9. Linguistic Analysis

Results

- Over 2000 scores collected
- 30 different tables
- Experiments repeated up to 4 times
- Different prompts used

2.4 Commonsense Reasoning

2.4.1 NorCommonsenseQA

Dataset Description NorCommonsenseQA⁹ follows the design of CommonsenseQA (Talmor et al., 2019), a multiple-choice commonsense question answering dataset. The annotation is performed under the supervision of UiO (see Appendix A.1 for annotation guidelines).

Task Formulation The task is to select a correct answer option given a question. The performance metric is the accuracy score.

- **Question:** “Hva skjer med en hund før noen henger opp plakater av den?”
- **Option A:** “Den blir borte”
- **Option B:** “Den trenger vann”
- **Option C:** “Den er trent”
- **Option D:** “Den bjeffer”
- **Option E:** “Den ruller rundt”
- **Correct answer:** A

k	Rank	Model	BLEU	Δ (BLEU)	chrF	Δ (chrF)	BERTScore	Δ (BERTScore)
0	2	mimir-extended	15.65	\times	49.08	\times	84.58	\times
	6	mimir-extended-scratch	10.90	\times	41.56	\times	83.43	\times
	4	mimir-base	11.20	\times	41.58	\times	86.77	\times
	8	mimir-base-scratch	10.07	\times	39.20	\times	86.21	\times
	13	mimir-fiction	7.53	-2.5	33.42	-5.8	79.94	-6.3
	9	mimir-nonfiction	10.54	+0.5	40.37	+1.2	81.06	-5.1
	5	mimir-factual	12.58	+2.5	43.37	+4.2	83.66	-2.5
	1	mimir-newspapers	36.89	+26.8	57.01	+17.8	90.89	+4.7
	10	mimir-books	10.27	+0.2	39.75	+0.5	81.89	-4.3
	7	mimir-rightholders	10.48	+0.4	40.38	+1.2	84.20	-2.0
	3	mimir-untranslated-with-news	12.65	+2.6	43.72	+4.5	85.83	-0.4
	11	mimir-untranslated	9.74	-0.3	39.23	-0.0	81.38	-4.8
	12	mimir-translated	7.45	-2.6	33.95	-5.2	82.02	-4.2
1	1	mimir-extended	59.25	\times	73.40	\times	94.72	\times
	3	mimir-extended-scratch	55.15	\times	69.83	\times	94.01	\times
	2	mimir-base	58.80	\times	73.06	\times	94.63	\times
	5	mimir-base-scratch	53.23	\times	68.53	\times	93.51	\times
	13	mimir-fiction	49.48	-3.8	66.69	-1.8	93.04	-0.5
	10	mimir-nonfiction	50.01	-3.2	68.36	-0.2	93.43	-0.1
	11	mimir-factual	49.36	-3.9	68.32	-0.2	93.40	-0.1
	9	mimir-newspapers	52.18	-1.0	67.47	-1.1	93.40	-0.1
	4	mimir-books	50.20	-3.0	68.77	+0.2	93.56	-0.0
	6	mimir-rightholders	51.25	-2.0	68.56	-0.0	93.51	-0.0
	8	mimir-untranslated-with-news	50.12	-3.1	68.74	+0.2	93.48	-0.0
	6	mimir-untranslated	51.34	-1.9	68.43	-0.1	93.52	-0.0
	12	mimir-translated	49.89	-3.3	67.17	-1.4	93.13	-0.4
4	1	mimir-extended	59.86	\times	73.81	\times	94.86	\times
	3	mimir-extended-scratch	56.39	\times	70.85	\times	94.23	\times
	2	mimir-base	59.24	\times	73.29	\times	94.73	\times
	9	mimir-base-scratch	54.66	\times	69.29	\times	93.90	\times
	12	mimir-fiction	52.82	-1.8	68.33	-1.0	93.66	-0.2
	10	mimir-nonfiction	54.33	-0.3	69.55	+0.3	93.97	+0.1
	7	mimir-factual	54.50	-0.2	69.81	+0.5	93.95	-0.0
	11	mimir-newspapers	53.51	-1.1	68.47	-0.8	93.78	-0.1
	4	mimir-books	54.75	+0.1	69.69	+0.4	94.06	+0.2
	5	mimir-rightholders	54.67	-0.0	69.64	+0.3	94.06	+0.2
	8	mimir-untranslated-with-news	54.40	-0.3	69.56	+0.3	93.98	+0.1
	6	mimir-untranslated	54.58	-0.1	69.72	+0.4	94.03	+0.1
	13	mimir-translated	52.86	-1.8	68.24	-1.1	93.66	-0.2
16	1	mimir-extended	60.54	\times	74.26	\times	94.96	\times
	3	mimir-extended-scratch	57.30	\times	71.55	\times	94.41	\times
	2	mimir-base	59.74	\times	73.69	\times	94.82	\times
	10	mimir-base-scratch	55.13	\times	69.84	\times	94.04	\times
	12	mimir-fiction	54.30	-0.8	69.00	-0.8	93.82	-0.2
	7	mimir-nonfiction	55.37	+0.2	70.17	+0.3	94.10	+0.1
	8	mimir-factual	55.36	+0.2	69.97	+0.1	94.05	-0.0
	13	mimir-newspapers	53.97	-1.2	68.95	-0.9	93.86	-0.2
	5	mimir-books	55.73	+0.6	70.16	+0.3	94.14	+0.1
	6	mimir-rightholders	55.62	+0.5	70.19	+0.3	94.10	+0.1
	9	mimir-untranslated-with-news	55.35	+0.2	69.86	-0.0	94.08	-0.0
	4	mimir-untranslated	55.65	+0.5	70.21	+0.4	94.12	+0.1
	11	mimir-translated	54.12	-1.0	69.15	-0.7	93.84	-0.2

Table 27: Results for zero- and few-shot evaluation on **Tatoeba (Bokmål \rightarrow English)** with $k \in \{0, 1, 4, 16\}$. The maximum BLEU, chrF, and BERTScore scores across a set of 5 Bokmål prompts are reported. δ (BLEU), δ (chrF), and δ (BERTScore) show the delta performance scores w.r.t. mimir-base-scratch.

k	Rank	Model	BLEU	Δ (BLEU)	chrF	Δ (chrF)	BERTScore	Δ (BERTScore)
0	4	mimir-extended	11.40	\times	41.80	\times	81.98	\times
	4	mimir-extended-scratch	10.38	\times	40.29	\times	83.67	\times
	2	mimir-base	19.19	\times	52.44	\times	90.13	\times
	3	mimir-base-scratch	12.86	\times	41.58	\times	87.50	\times
	13	mimir-fiction	5.64	-7.2	27.26	-14.3	78.98	-8.5
	8	mimir-nonfiction	9.63	-3.2	37.72	-3.9	79.82	-7.7
	6	mimir-factual	9.97	-2.9	36.96	-4.6	82.56	-4.9
	1	mimir-newspapers	35.06	+22.2	61.26	+19.7	91.88	+4.4
	9	mimir-books	8.78	-4.1	35.97	-5.6	79.57	-7.9
	12	mimir-rightholders	7.65	-5.2	33.63	-7.9	81.38	-6.1
	7	mimir-untranslated-with-news	9.92	-2.9	36.56	-5.0	83.66	-3.8
	10	mimir-untranslated	8.69	-4.2	35.84	-5.7	79.76	-7.7
	10	mimir-translated	8.34	-4.5	35.35	-6.2	80.48	-7.0
1	1	mimir-extended	58.71	\times	72.08	\times	94.35	\times
	3	mimir-extended-scratch	55.67	\times	69.60	\times	93.54	\times
	2	mimir-base	58.19	\times	71.77	\times	94.27	\times
	4	mimir-base-scratch	52.48	\times	67.21	\times	93.06	\times
	12	mimir-fiction	44.88	-7.6	65.46	-1.8	92.57	-0.5
	7	mimir-nonfiction	50.40	-2.1	67.34	+0.1	92.95	-0.1
	8	mimir-factual	49.84	-2.6	67.44	+0.2	92.86	-0.2
	10	mimir-newspapers	51.25	-1.2	66.54	-0.7	92.85	-0.2
	9	mimir-books	47.96	-4.5	66.86	-0.3	93.00	-0.1
	5	mimir-rightholders	48.48	-4.0	67.23	-0.0	93.15	+0.1
	10	mimir-untranslated-with-news	47.96	-4.5	66.59	-0.6	93.03	-0.0
	5	mimir-untranslated	50.72	-1.8	67.25	-0.0	92.98	-0.1
	13	mimir-translated	45.52	-7.0	65.40	-1.8	92.50	-0.6
4	1	mimir-extended	59.30	\times	72.57	\times	94.41	\times
	3	mimir-extended-scratch	55.43	\times	69.75	\times	93.74	\times
	2	mimir-base	58.81	\times	72.26	\times	94.45	\times
	7	mimir-base-scratch	53.64	\times	68.16	\times	93.44	\times
	12	mimir-fiction	52.68	-1.0	67.67	-0.5	93.25	-0.2
	6	mimir-nonfiction	53.77	+0.1	68.73	+0.6	93.38	-0.1
	9	mimir-factual	53.50	-0.1	68.44	+0.3	93.40	-0.0
	13	mimir-newspapers	53.05	-0.6	67.53	-0.6	93.25	-0.2
	7	mimir-books	53.69	-0.0	68.70	+0.5	93.32	-0.1
	4	mimir-rightholders	53.64	-0.0	68.89	+0.7	93.47	-0.0
	10	mimir-untranslated-with-news	53.40	-0.2	68.39	+0.2	93.39	-0.0
	5	mimir-untranslated	53.85	+0.2	68.50	+0.3	93.41	-0.0
	11	mimir-translated	52.22	-1.4	67.76	-0.4	93.36	-0.1
16	2	mimir-extended	59.95	\times	72.83	\times	94.57	\times
	3	mimir-extended-scratch	57.16	\times	70.90	\times	94.09	\times
	1	mimir-base	60.17	\times	73.08	\times	94.63	\times
	8	mimir-base-scratch	55.10	\times	69.14	\times	93.68	\times
	11	mimir-fiction	54.47	-0.6	68.49	-0.7	93.57	-0.1
	10	mimir-nonfiction	54.83	-0.3	69.12	-0.0	93.63	-0.1
	9	mimir-factual	55.03	-0.1	69.26	+0.1	93.64	-0.0
	12	mimir-newspapers	54.40	-0.7	68.46	-0.7	93.58	-0.1
	5	mimir-books	55.42	+0.3	69.42	+0.3	93.73	-0.0
	7	mimir-rightholders	55.15	-0.0	69.33	+0.2	93.69	-0.0
	4	mimir-untranslated-with-news	55.60	+0.5	69.59	+0.5	93.73	-0.0
	6	mimir-untranslated	55.29	+0.2	69.37	+0.2	93.72	-0.0
	13	mimir-translated	53.07	-2.0	67.53	-1.6	93.29	-0.4

Table 28: Results for zero- and few-shot evaluation on **Tatoeba (Nynorsk \rightarrow English)** with $k \in \{0, 1, 4, 16\}$. The maximum BLEU, chrF, and BERTScore scores across a set of 5 Nynorsk prompts are reported. δ (BLEU), δ (chrF), and δ (BERTScore) show the delta performance scores w.r.t. mimir-base-scratch.

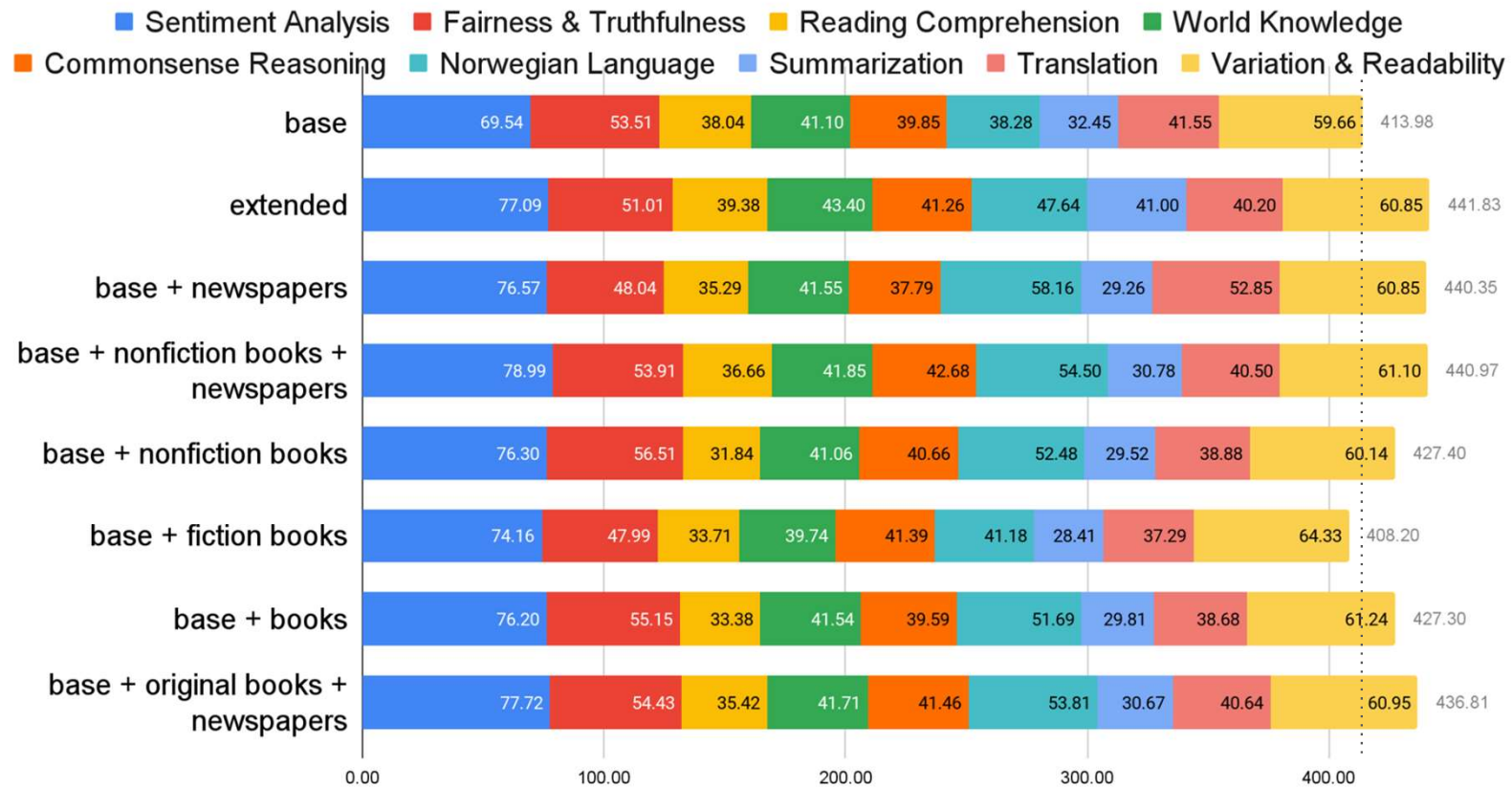
Aggregated Scores per Model Skill

- Sentiment Analysis ■ Fairness & Truthfulness ■ Reading Comprehension ■ World Knowledge
- Commonsense Reasoning ■ Norwegian Language ■ Summarization ■ Translation
- Variation & Readability

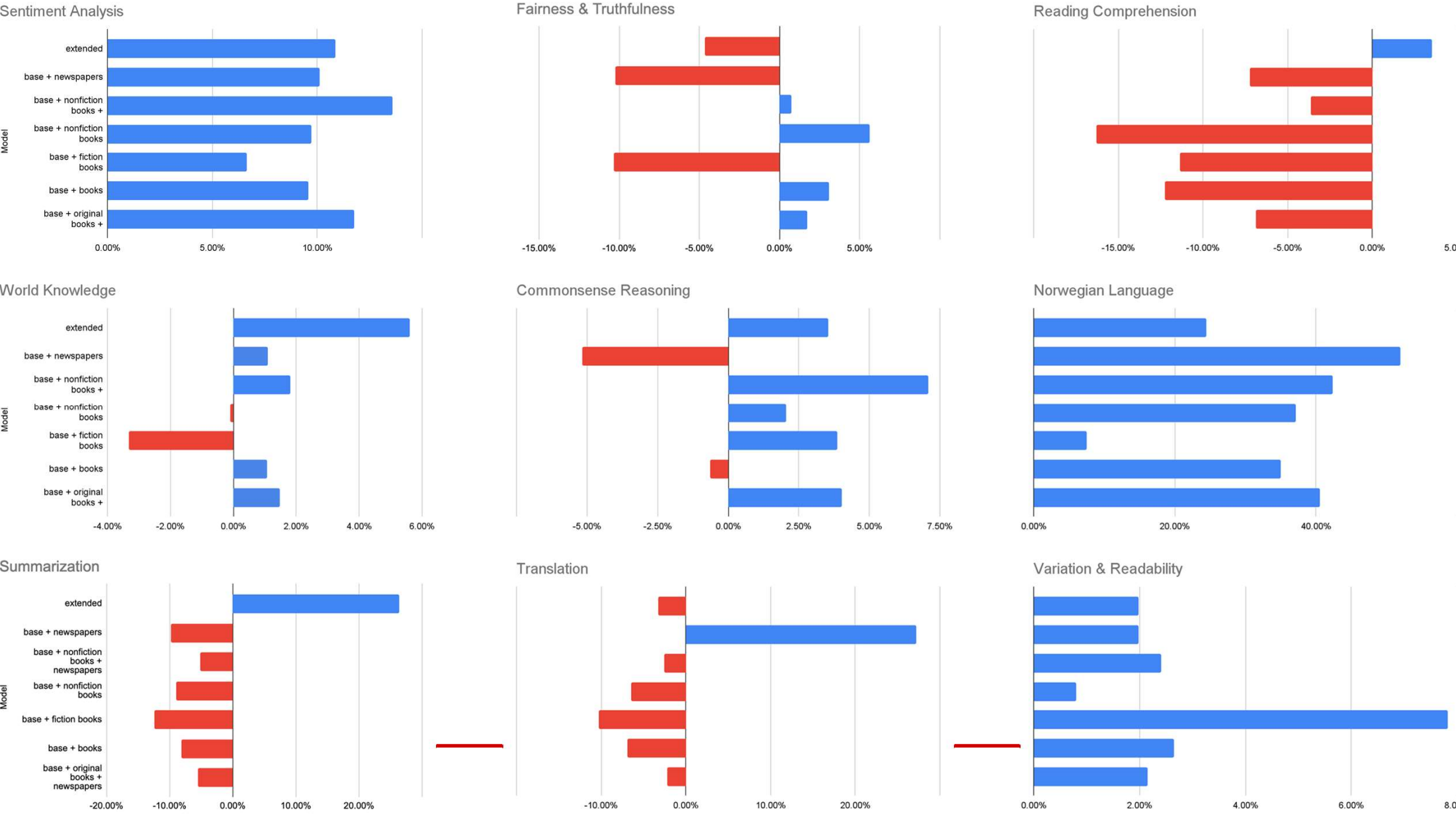


Evaluation: Ablations from Scratch

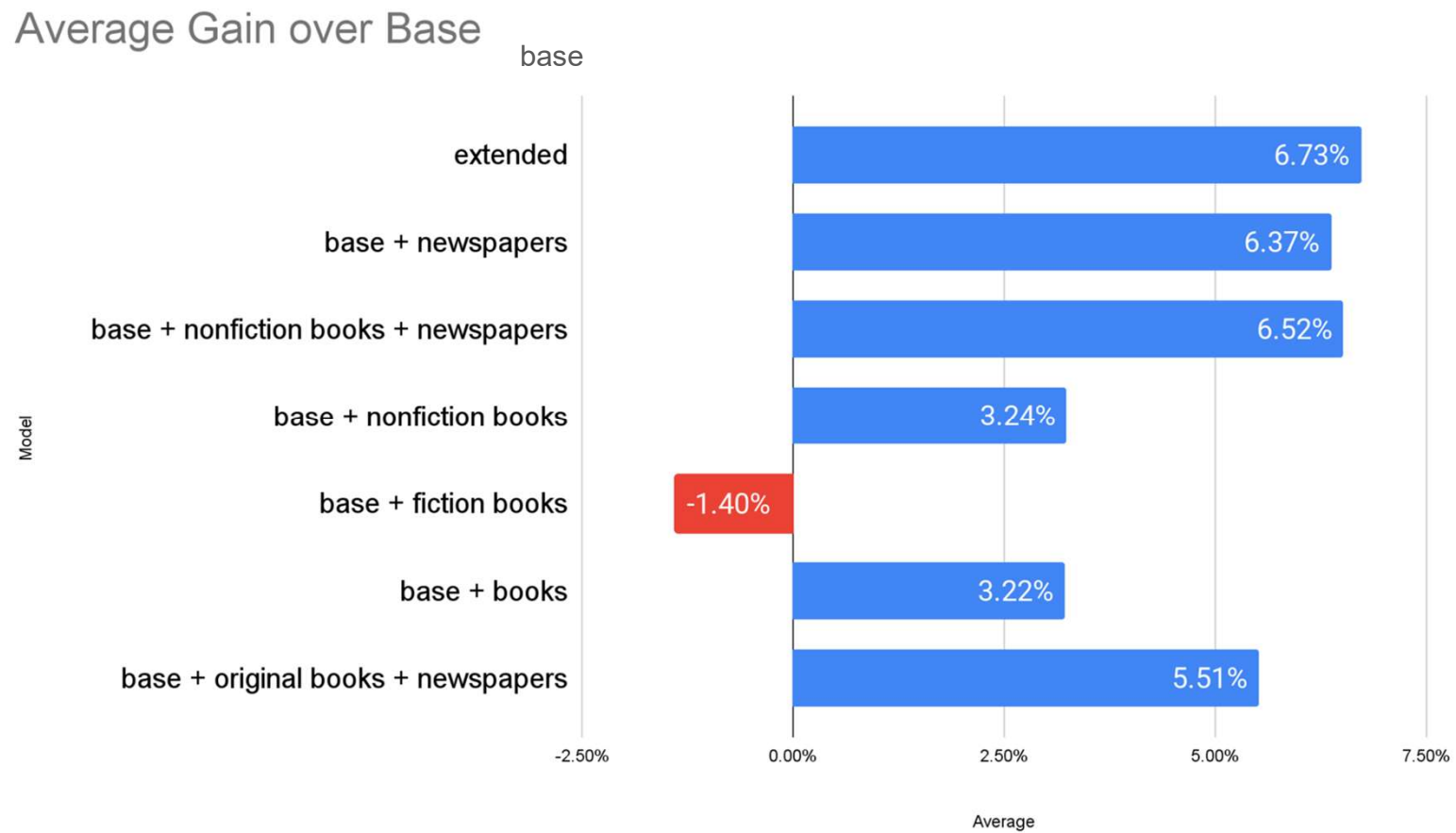
Data Ablations over Base



Evaluation: Ablations from Scratch

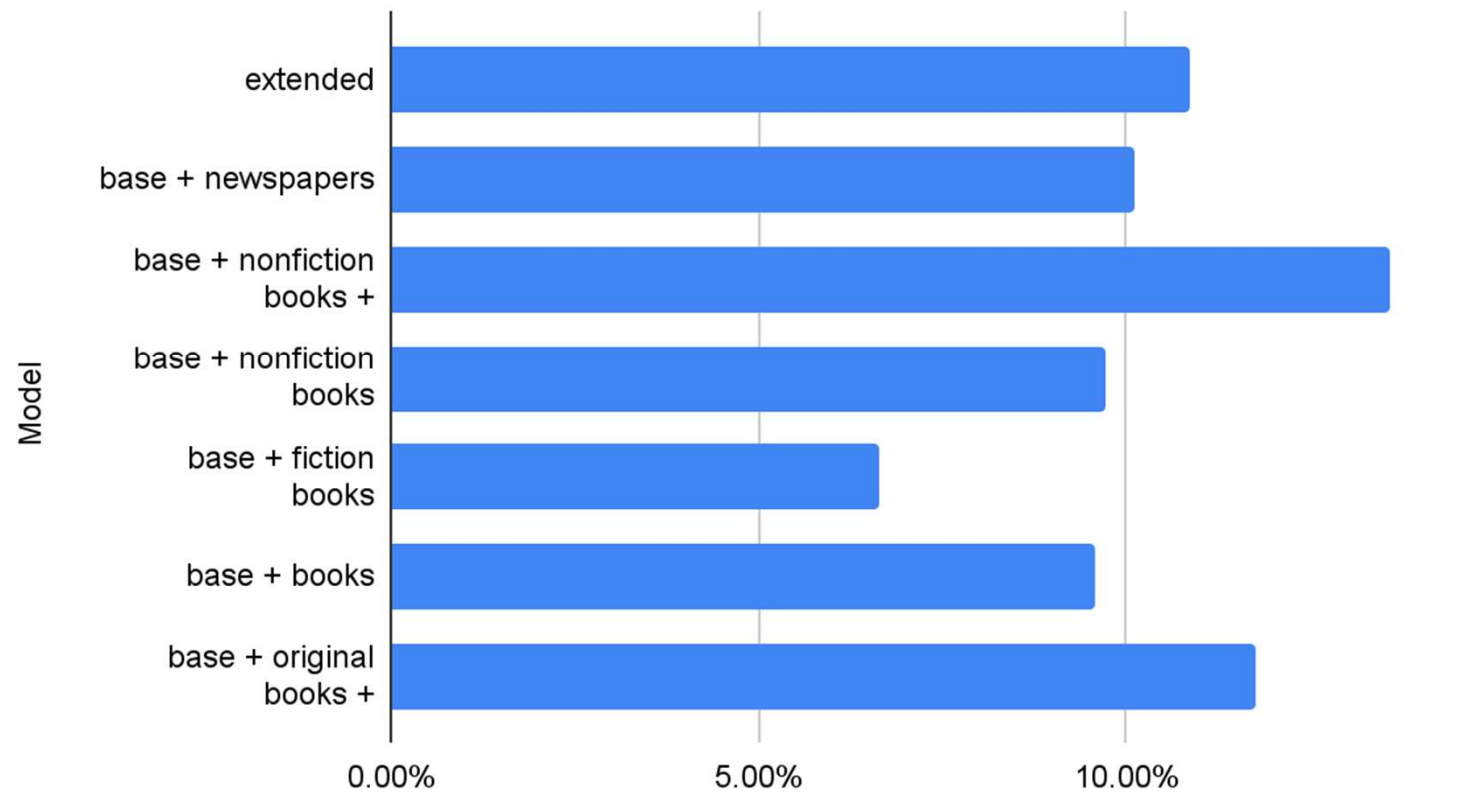


Evaluation: Ablations from Scratch

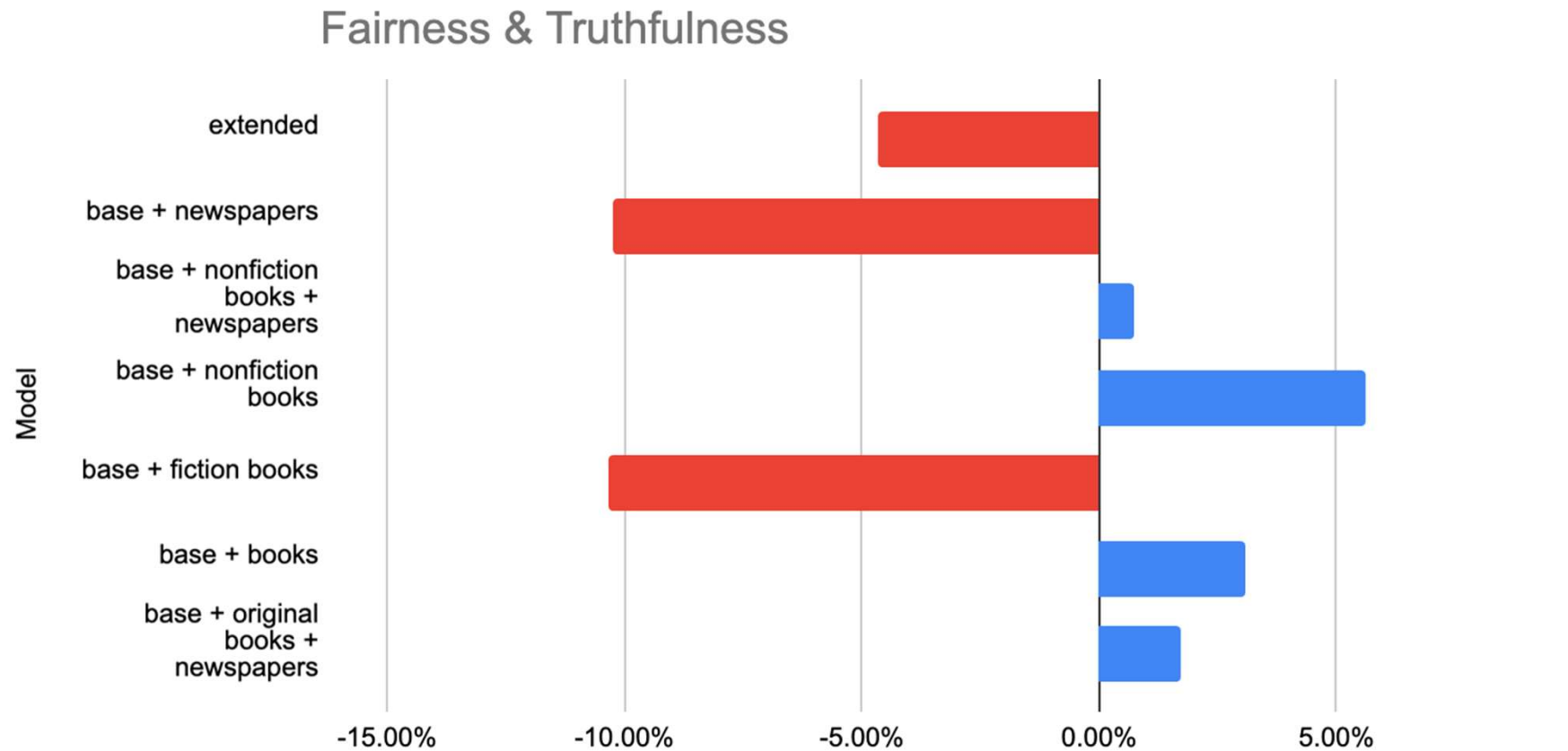


Evaluation: Ablations from Scratch

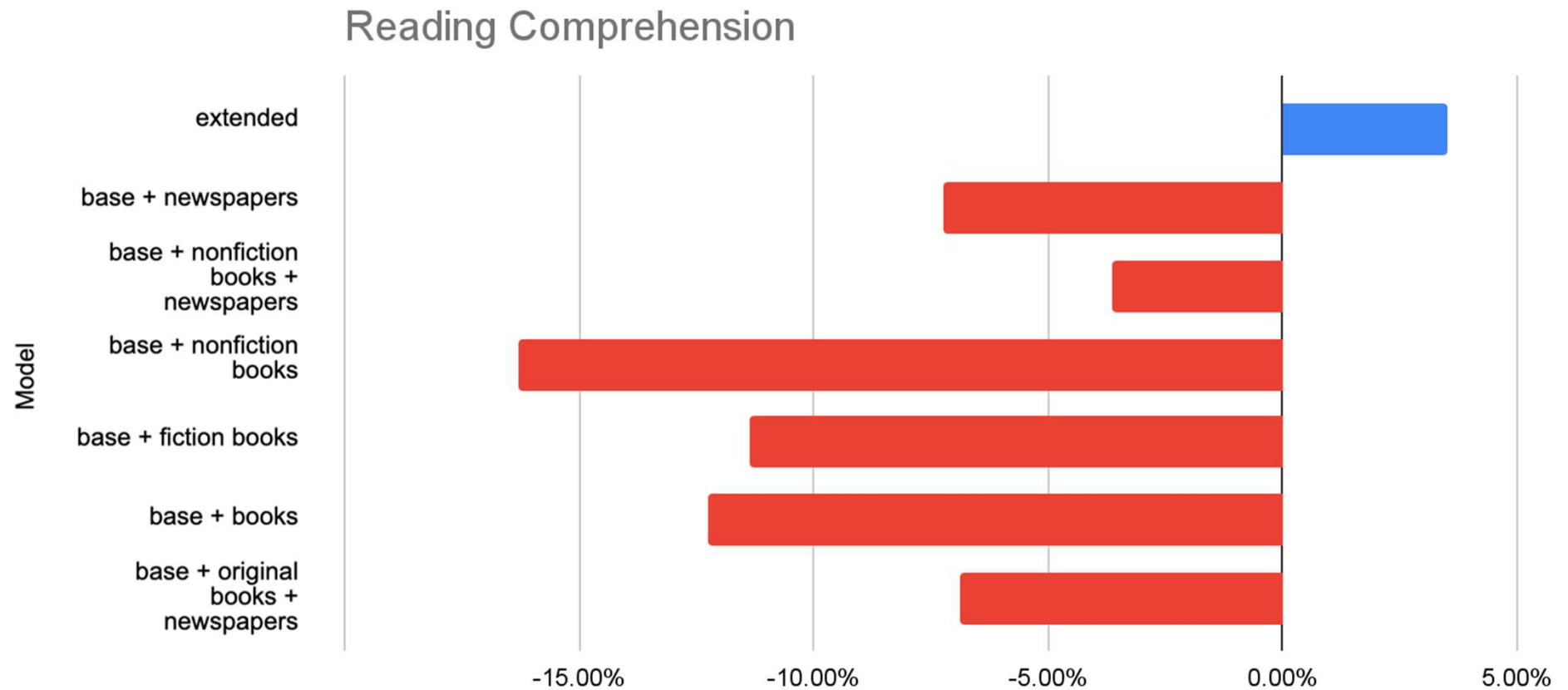
Sentiment Analysis



Evaluation: Ablations from Scratch

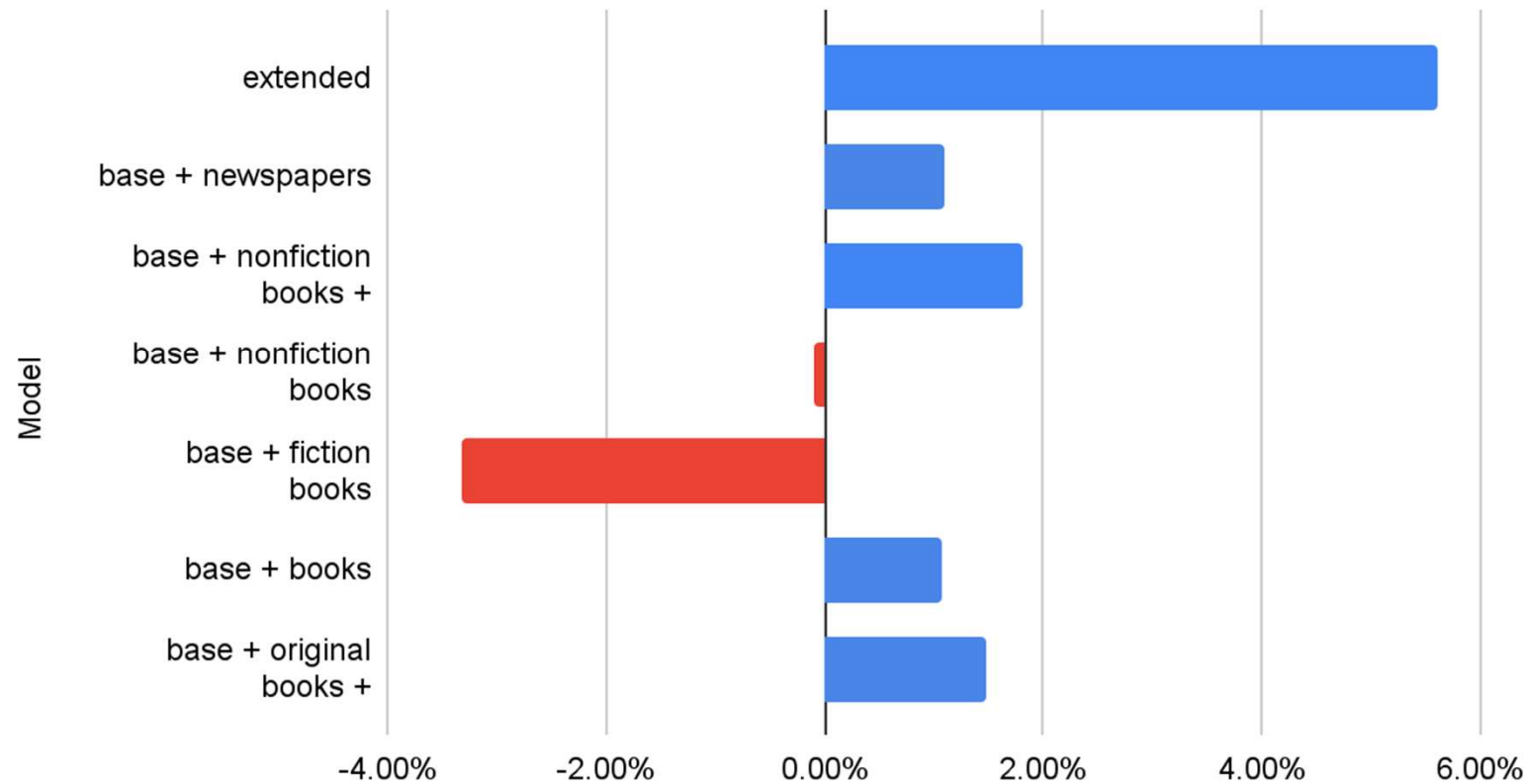


Evaluation: Ablations from Scratch

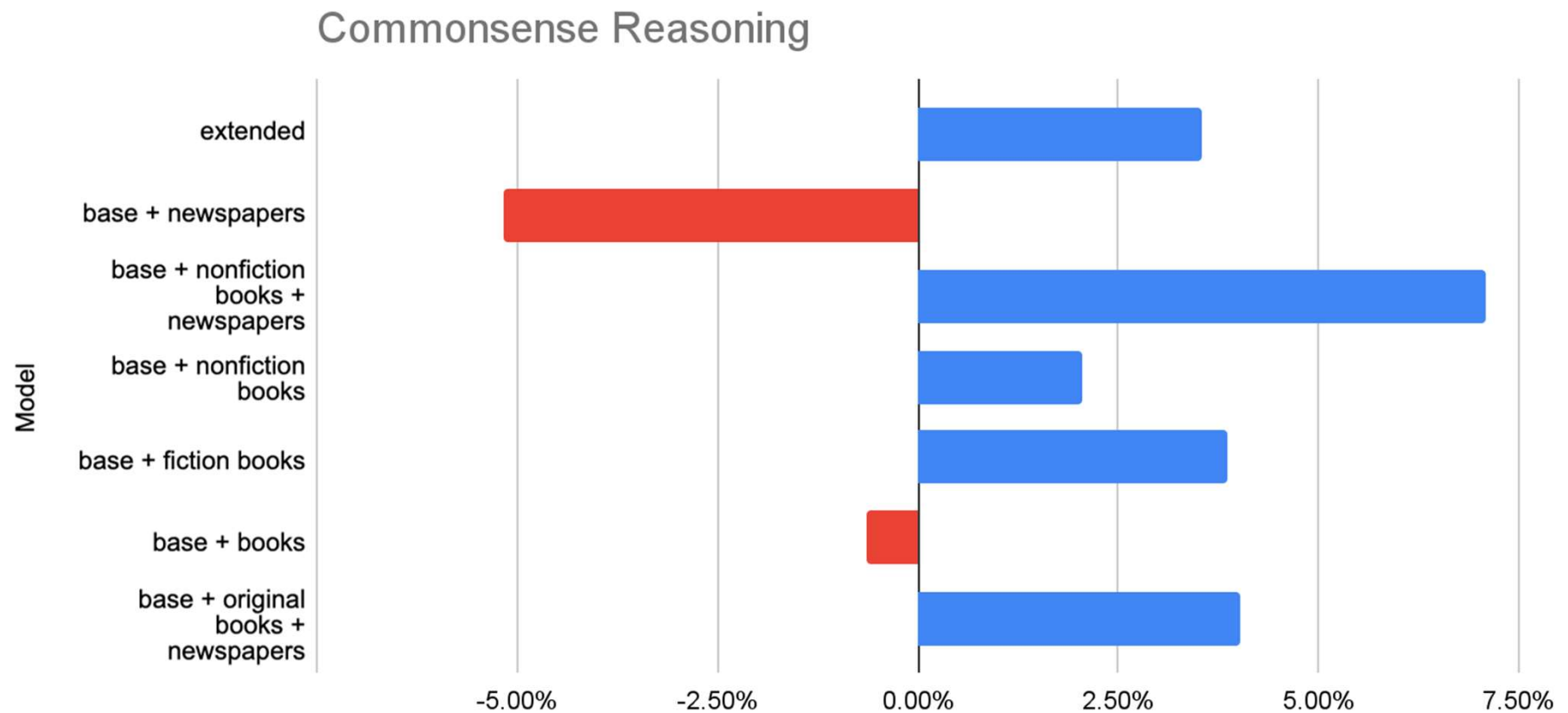


Evaluation: Ablations from Scratch

World Knowledge

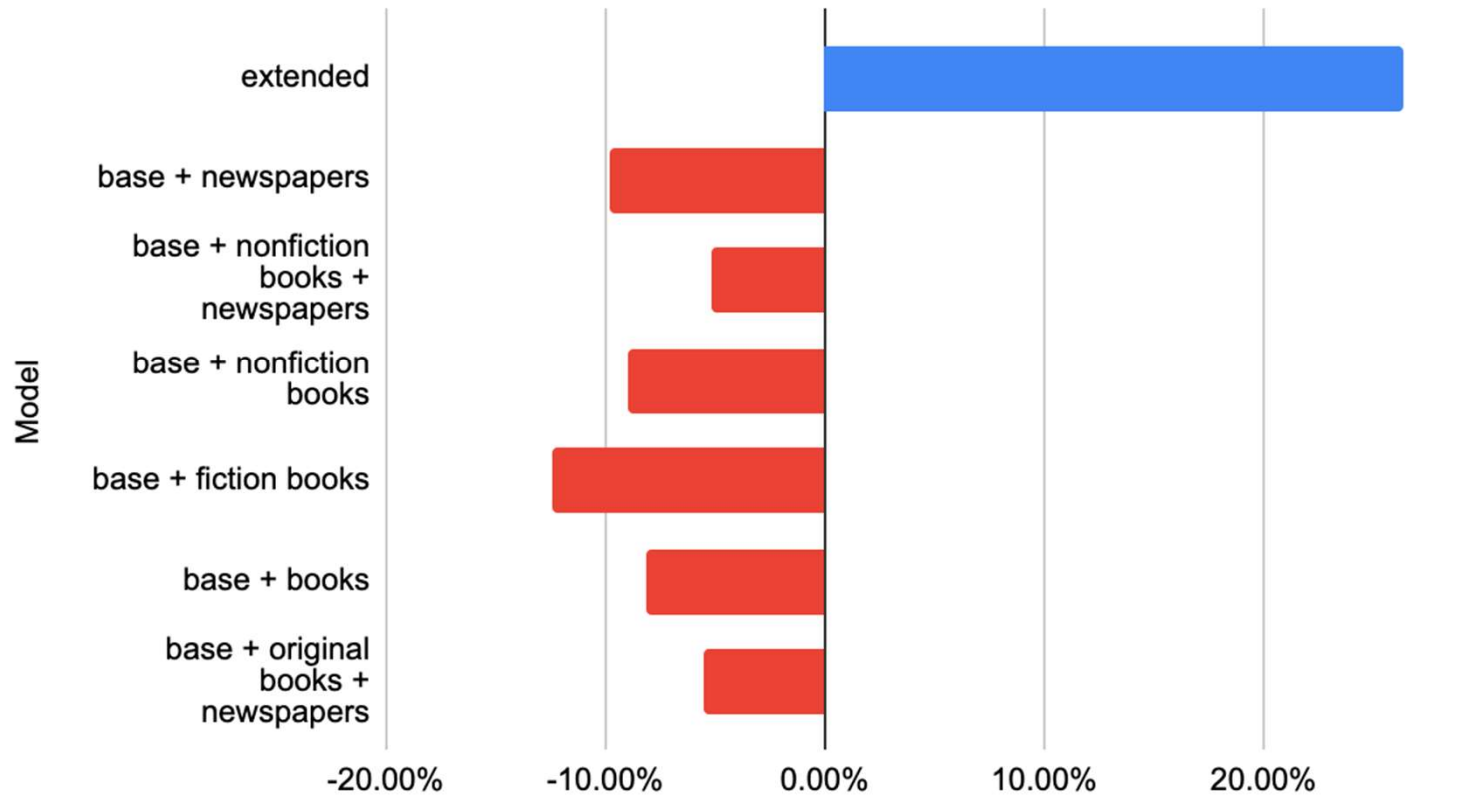


Evaluation: Ablations from Scratch

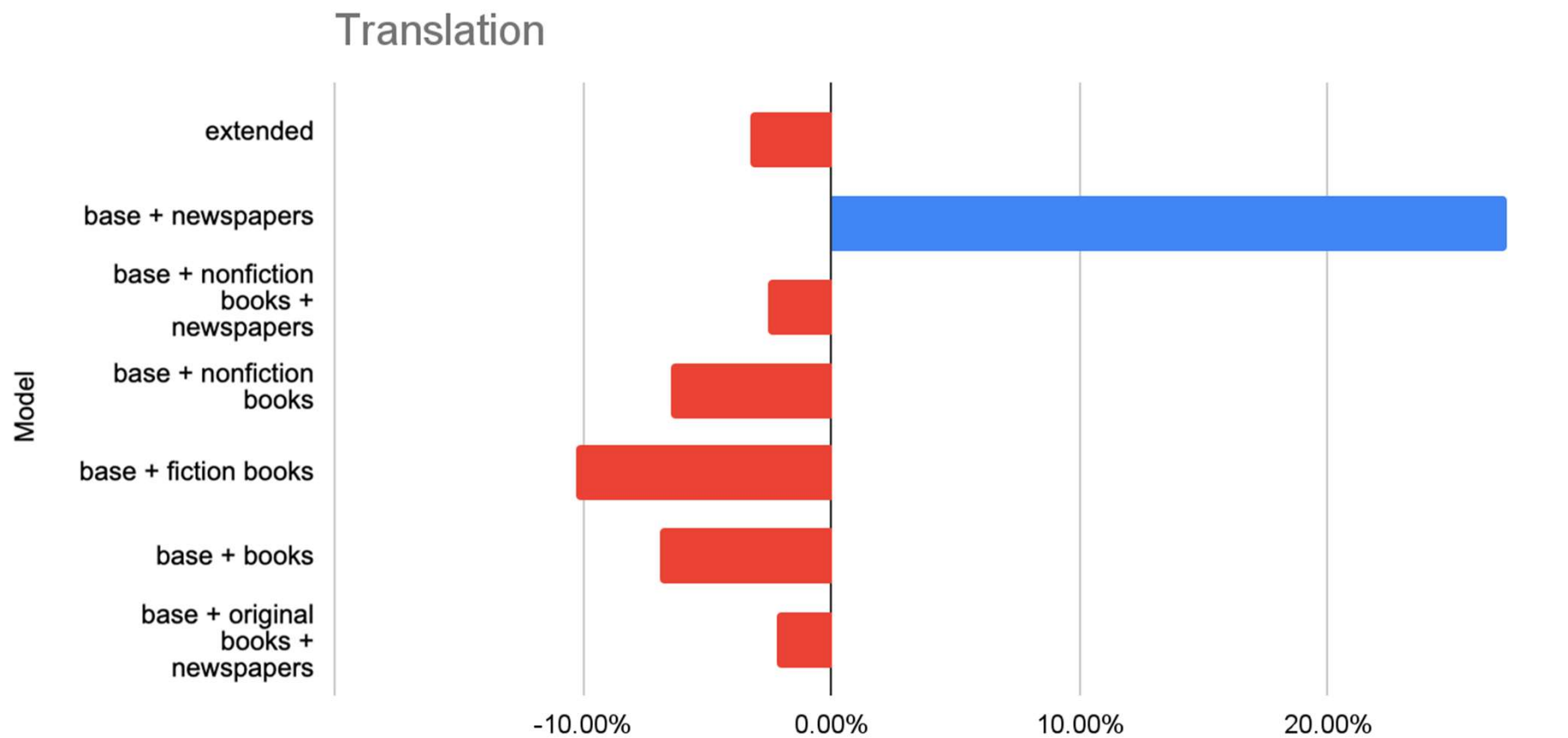


Evaluation: Ablations from Scratch

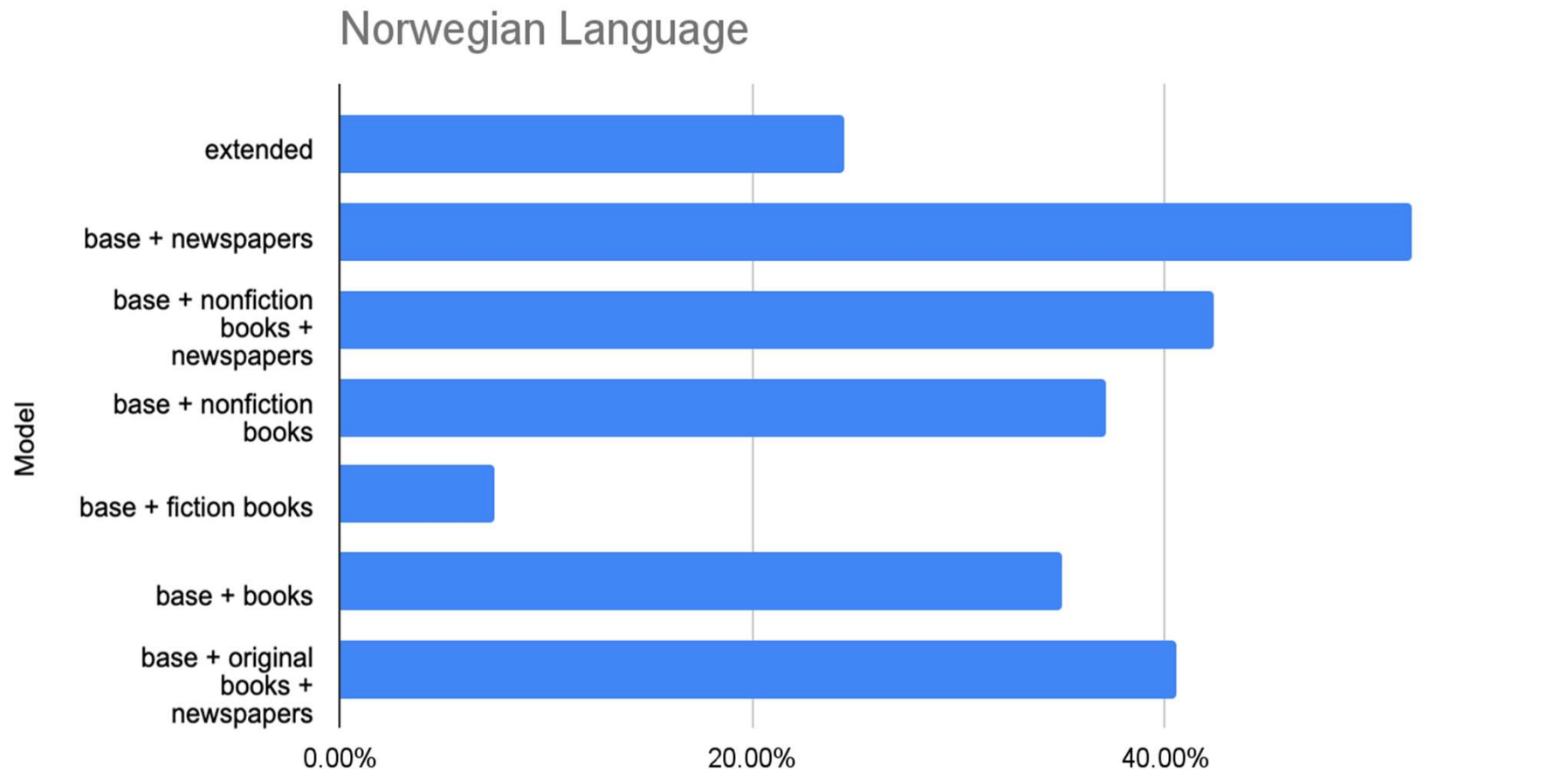
Summarization



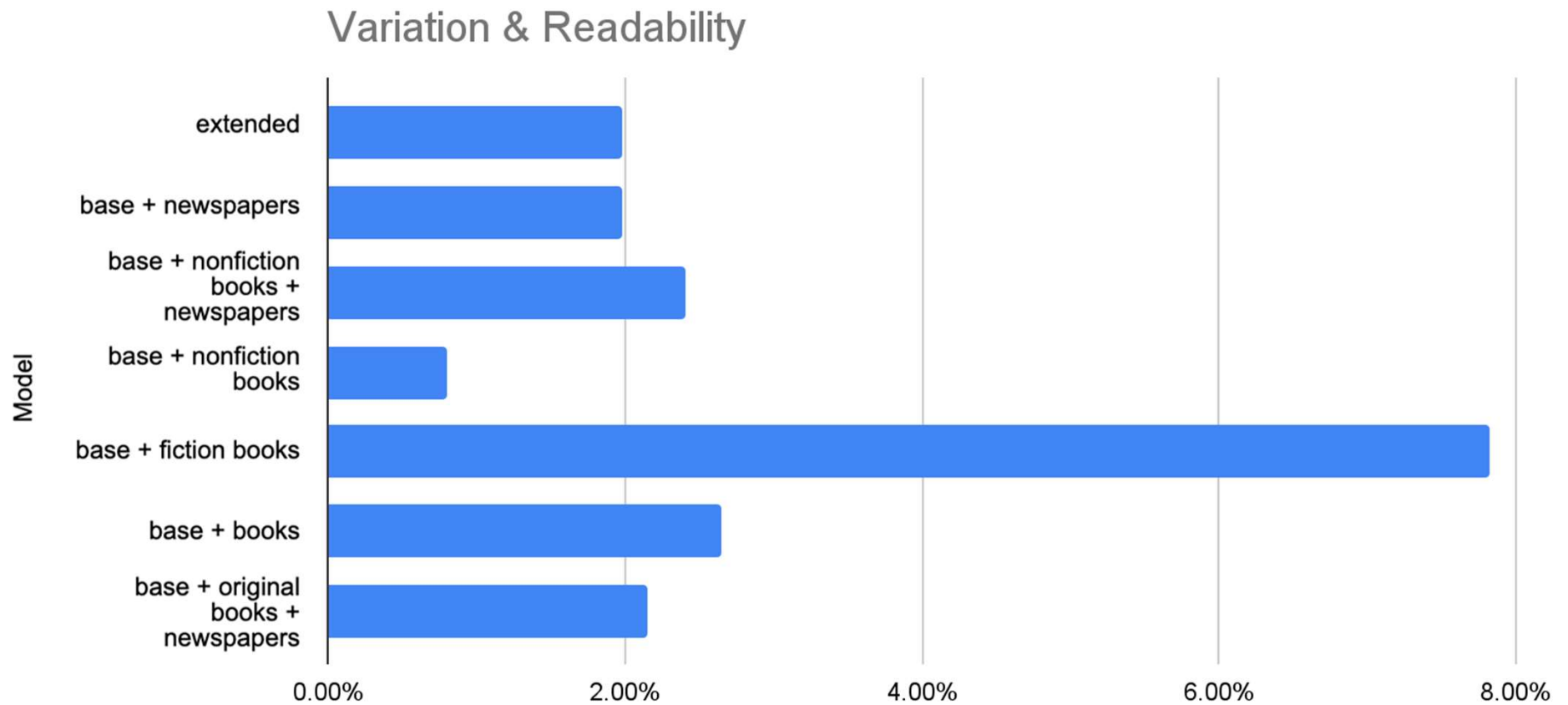
Evaluation: Ablations from Scratch



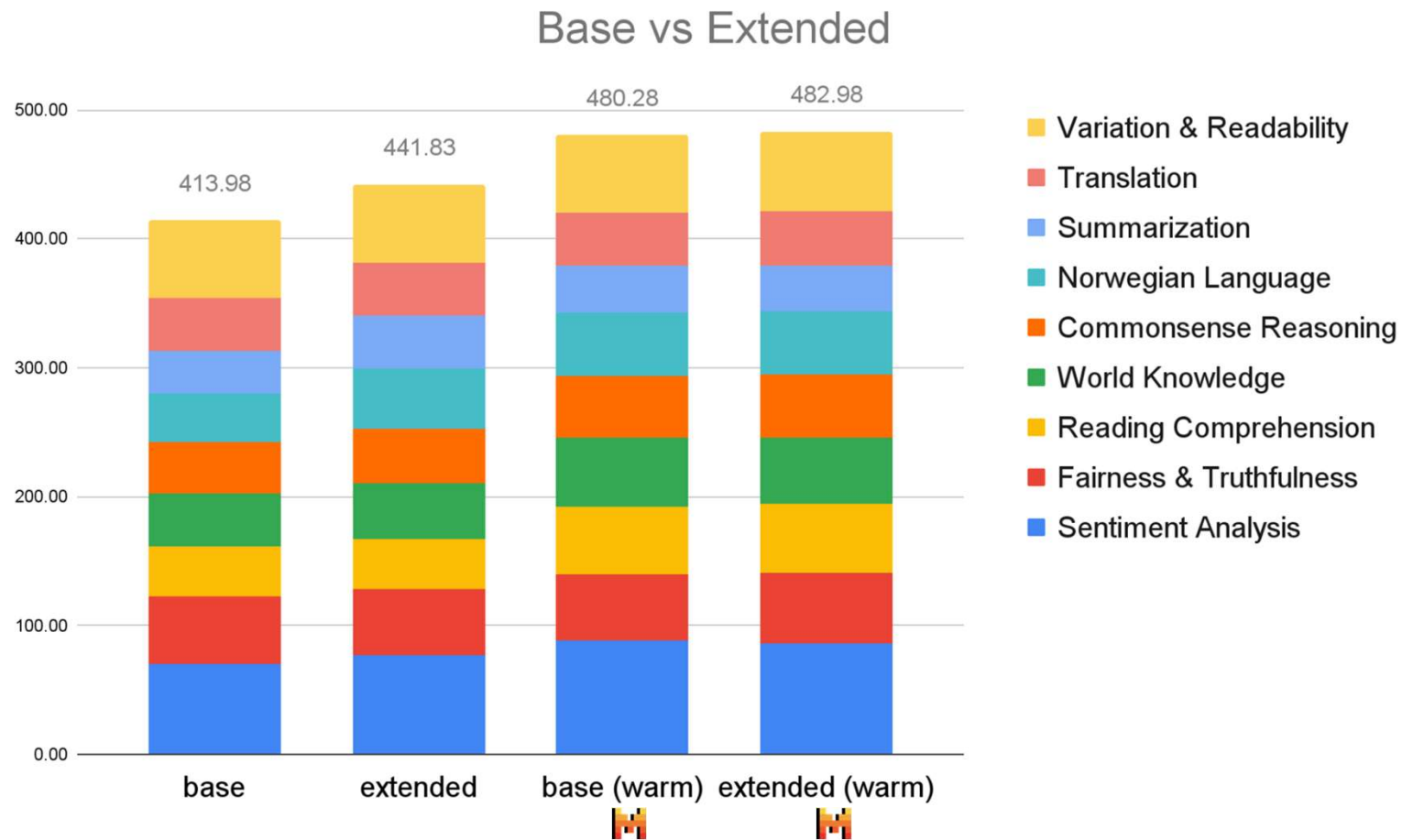
Evaluation: Ablations from Scratch



Evaluation: Ablations from Scratch

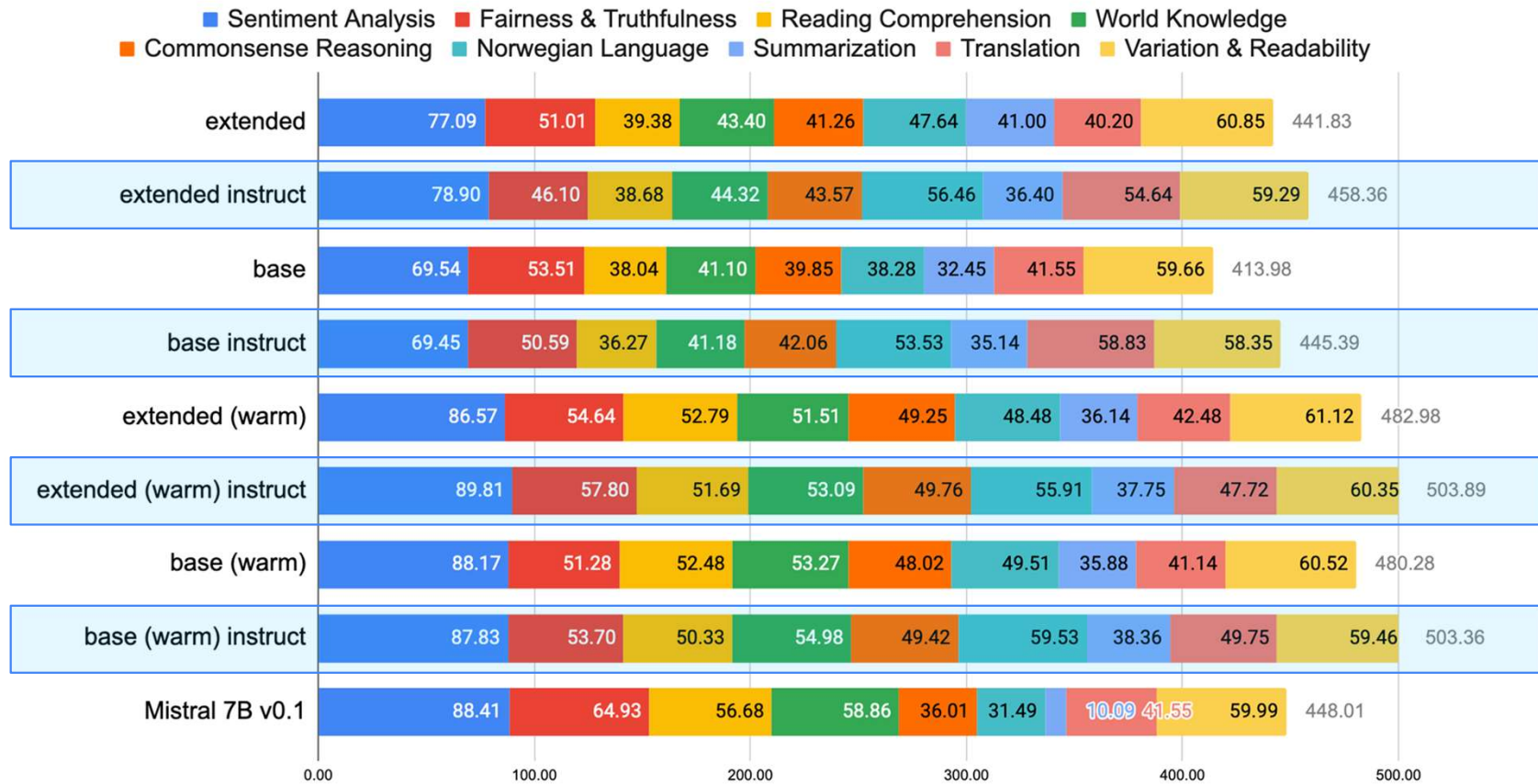


Evaluation: Pre-Training

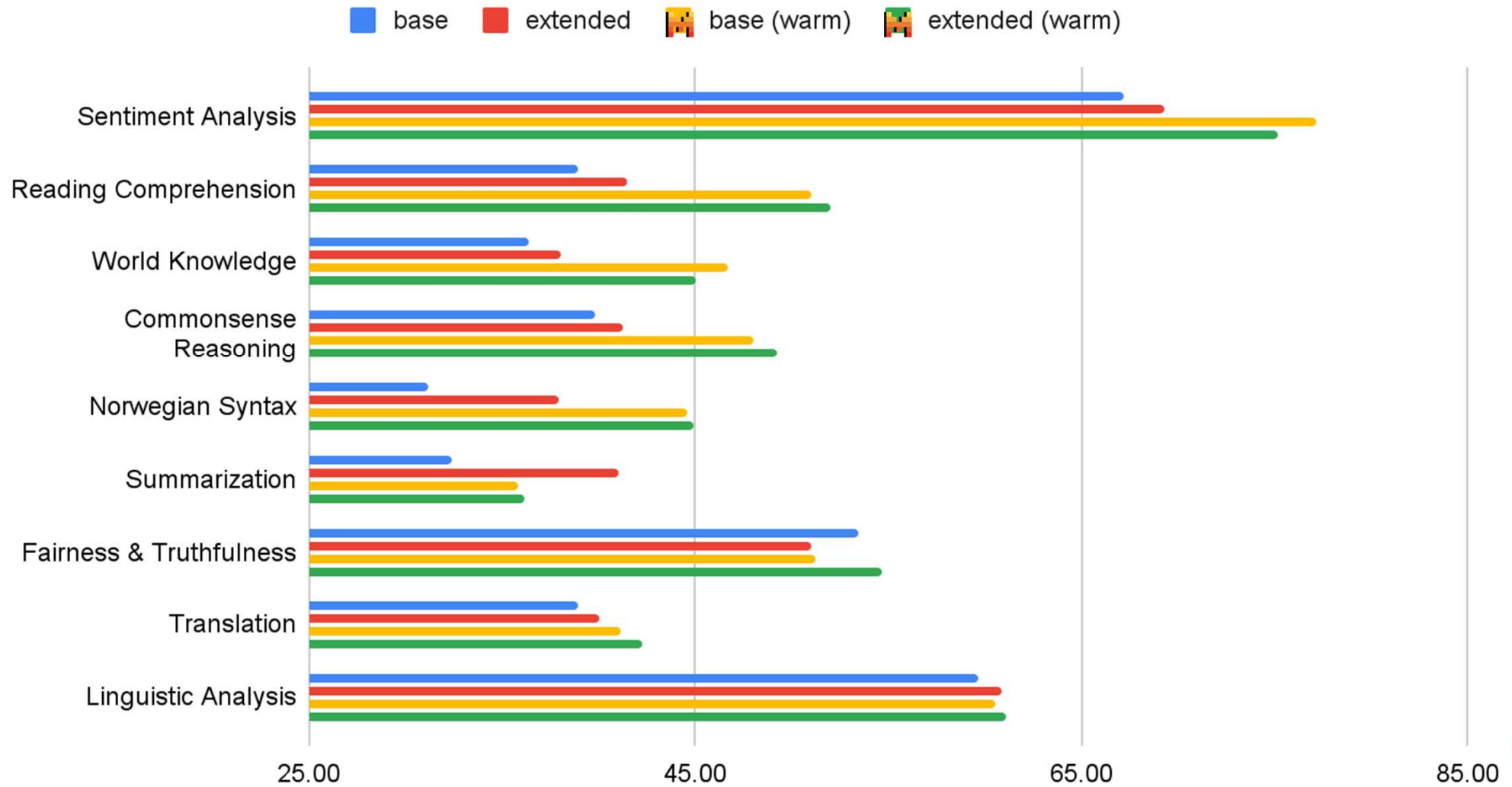


Evaluation: Instruction-following

Core and Instruct models

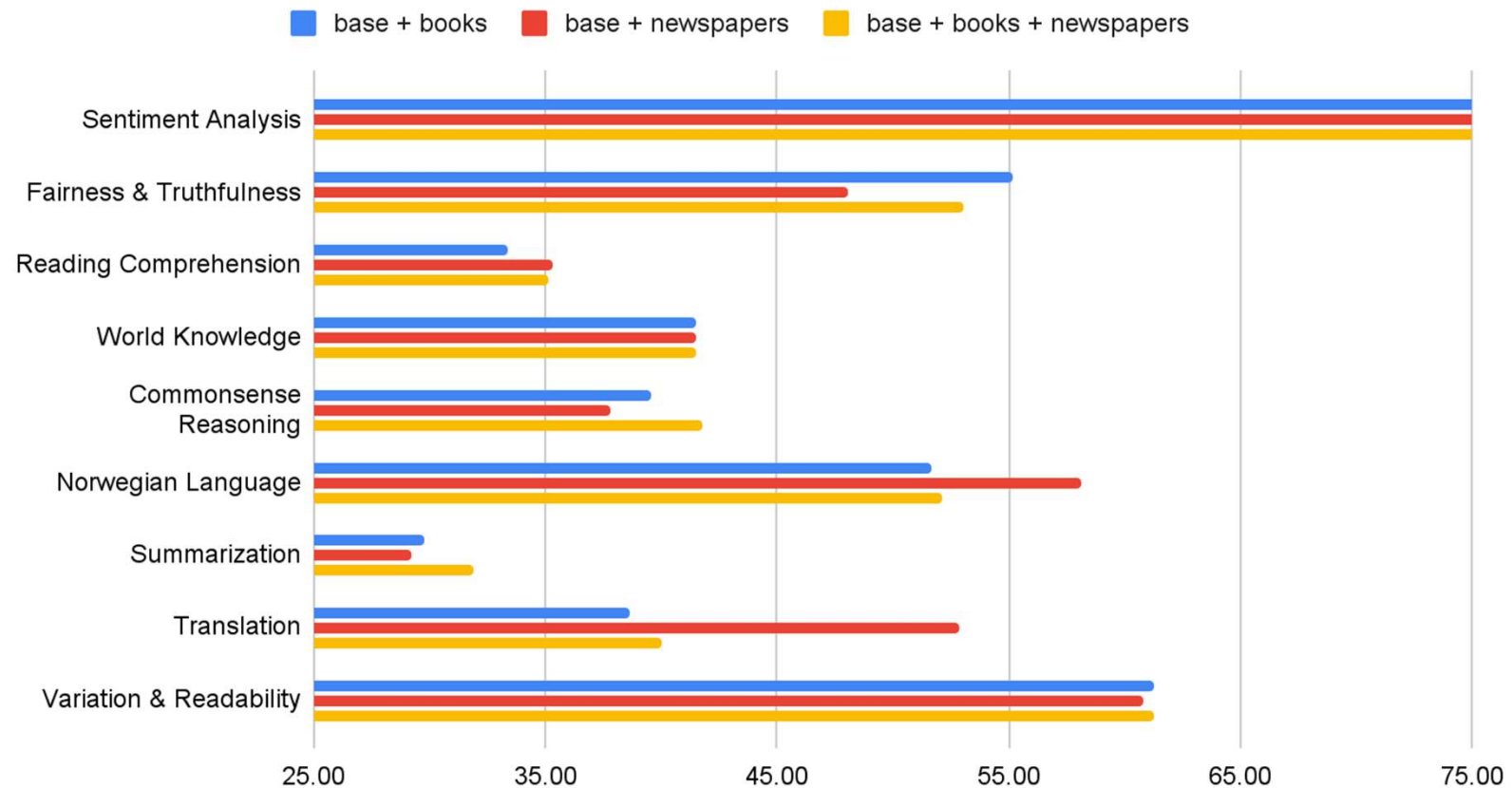


Base vs Extended



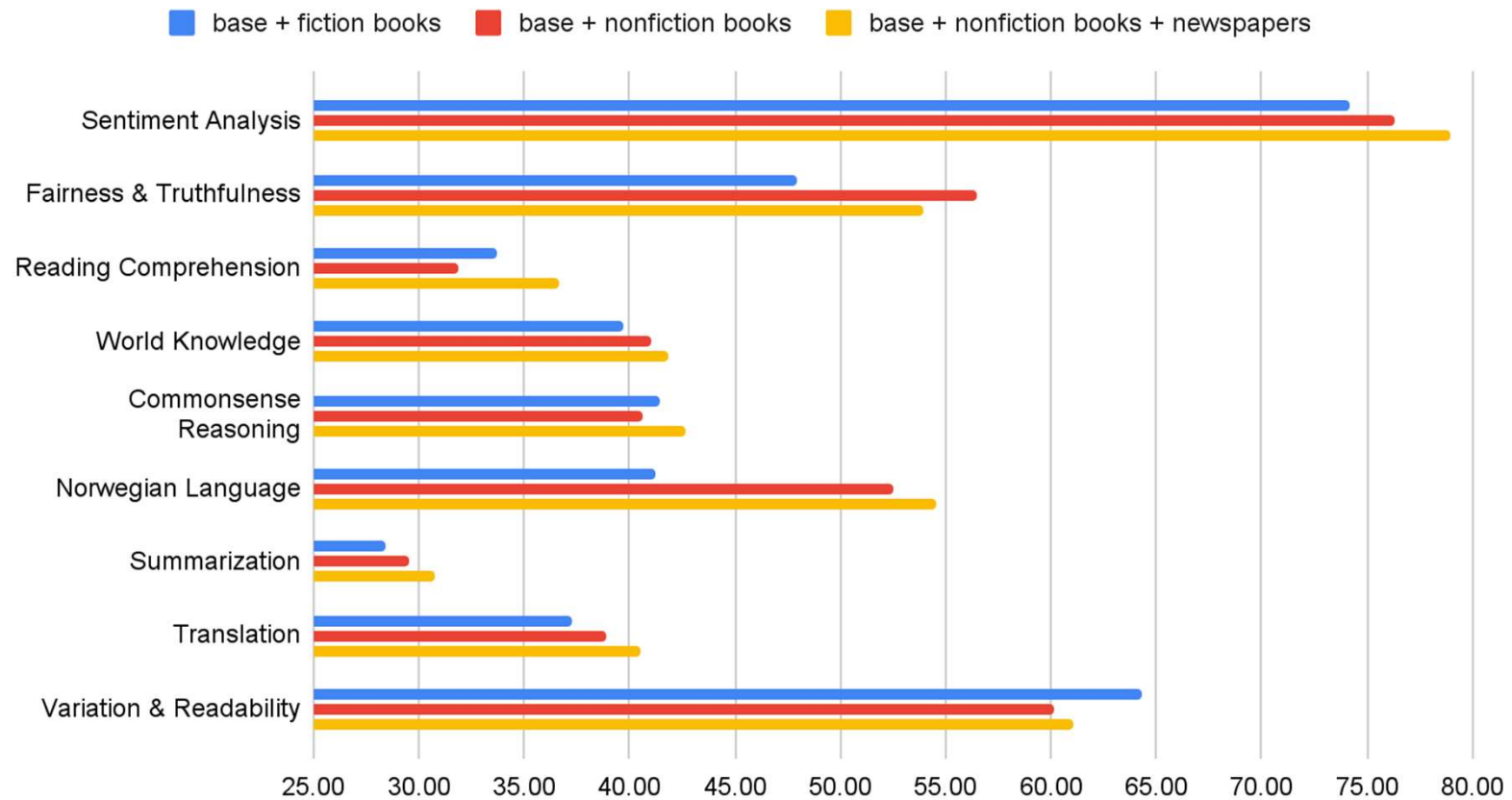
Evaluation: Source Type

Newspapers vs Books



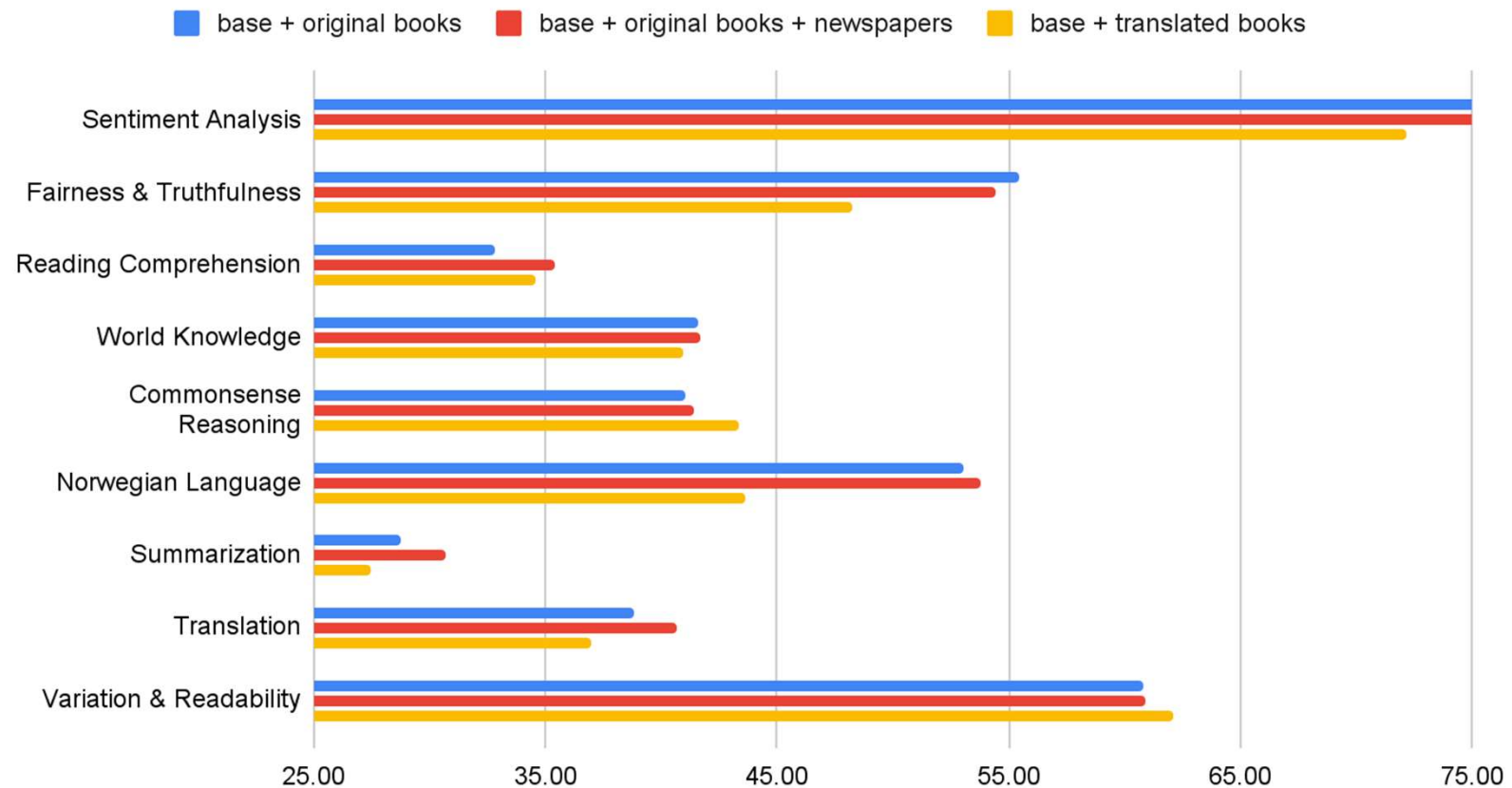
Evaluation: Genre

Fiction vs Factuality



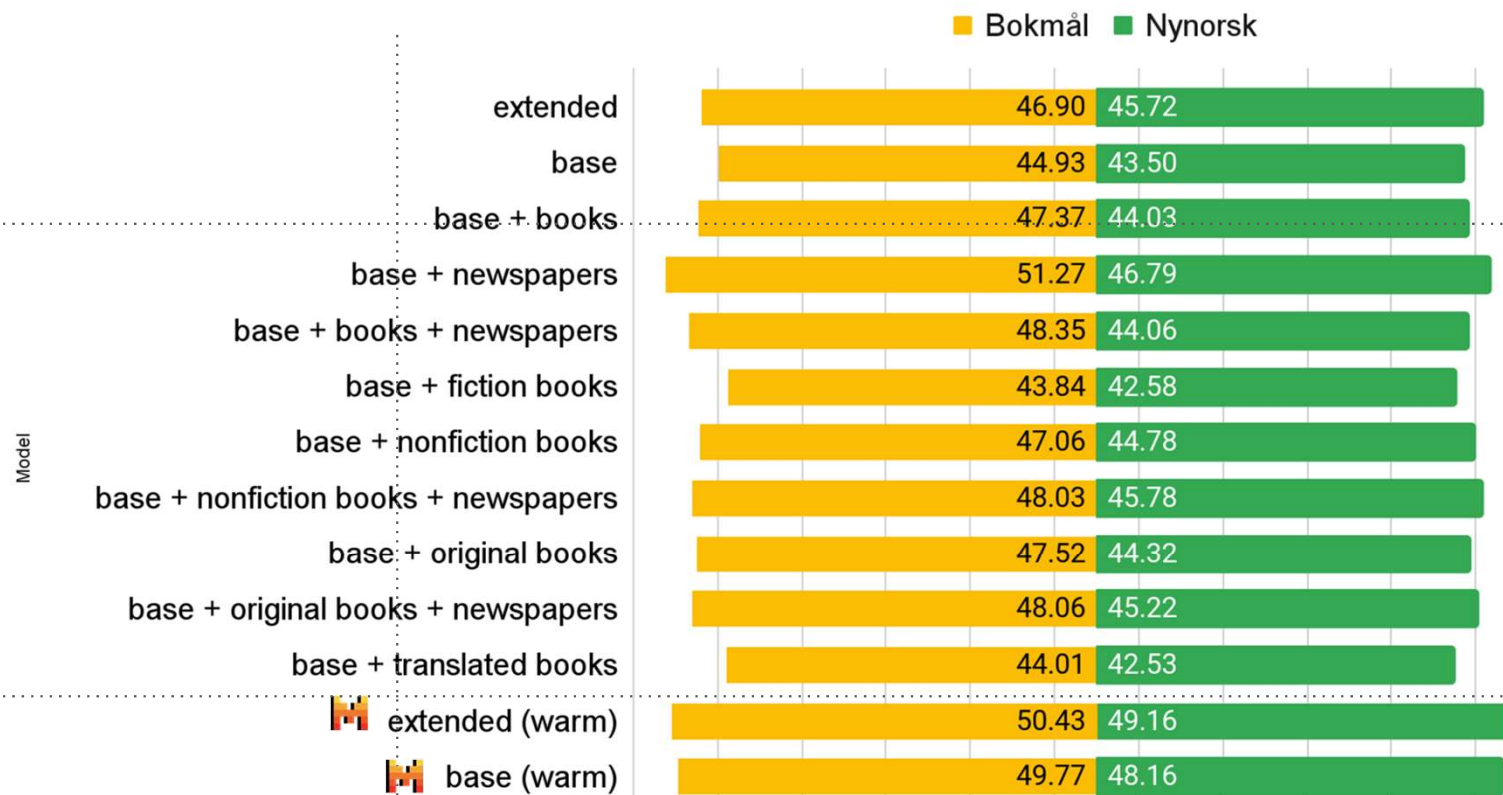
Evaluation: Original Literature

Original vs Translated



Evaluation: Language

Bokmål and Nynorsk



Conclusions

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.
1. **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.
1. **Instruction fine-tuning consistently increases the performance** of all core models, regardless of their pre-training data.

Consequences

- Informing the negotiations with the rights holders
- Acknowledgment at the Government level
- Funding for new unit for language modeling
- Staffing 20 new people dedicated to AI
- Specialized in-house hardware and quota in HPCs
- Central role in the AI ecosystem in Norway
- Digital strategy and budget line for 2025-2030

Takk!

Questions?

Javier de la Rosa
versae@nb.no



AI-lab

National Library of Norway



www.nb.no

ai.nb.no