

1. The importance of openness in the era of generative AI
 - LLM-as-a-service
 - Open software and open LLMs
 - Aspects of openness
 - LTG and the open LLM ecosystem
2. Web-derived LLM training data for Norwegian
3. Training data for Norwegian LLM research
4. Developing NorMistral-11B
- 5. NorEval: Native benchmarking for Norwegian LLMs**
6. The language quality of the Norwegian output in language models
7. Outlook



ACL Findings
2025.findings-acl.181



GitHub
github.com/ltgoslo/noreval

What is a benchmark?

- A collection of datasets, performance metrics, and an aggregation procedure
- Gamified with leaderboards, where LLMs are ranked against one another

Rank	Name	Score	Task ₁ (Acc.)	Task ₂ (F1)	Task ₃ (F1)
1	Model _A	0.95	0.98	0.92	0.96
2	Human	0.94	0.98	0.93	0.91
3	Model _B	0.90	0.95	0.95	0.80

A toy leaderboard for illustration purposes.

NorEval

- The **largest** benchmark for Bokmål and Nynorsk comprising **24 human-created datasets** integrated into LM Evaluation Harness
- Developed from Mímir, a project on evaluating **the impact of copyrighted material** on pretraining Norwegian LLMs (NB, UiO, & NTNU)

NorEval

- The **largest** benchmark for Bokmål and Nynorsk comprising **24 human-created datasets** integrated into LM Evaluation Harness
- Developed from Mímir, a project on evaluating **the impact of copyrighted material** on pretraining Norwegian LLMs (NB, UiO, & NTNU)

Prompt collection & Human baselines

- 100+ prompts for robust multi-prompt evaluation
- Human baselines on QA and Norwegian language knowledge tasks

 **NorEval**  BM, NN  6 task types  9 task categories  24 datasets  100+ prompts  Human-created

Text classification Sentence-level sentiment analysis NoReC Sentence (BM) 🐱	Sequence-to-sequence generation Norwegian language knowledge ASK-GEC (BM) 🚀	Multiple-choice question answering Commonsense reasoning NorCommonsenseQA (BM/NN) 🚀
Document-level sentiment analysis NoReC Document (BM) 🐱	Machine translation Tatoeba (EN ↔ BM, EN ↔ NN) 🚀	Norwegian-specific & world knowledge NorOpenBookQA (BM/NN) 🚀 NRK-Quiz-QA (BM/NN) 🚀
Sentence ranking Norwegian language knowledge NCB (BM) 😎	Text summarization NorSumm (BM/NN) 🚀	Reading comprehension Belebele (BM) 🚀
Generative question answering Reading comprehension NorQuAD (BM) 🐱	Instruction following NorRewrite-IT (BM) 😎 NorSummarize-IT (BM) 😎	Truthfulness NorTruthfulQA MC (BM/NN) 🚀
	Truthfulness NorTruthfulQA Gen (BM/NN) 🚀	Sentence completion Norwegian language knowledge NorIdiom (BM/NN) 😎

Overview of NorEval. EN=English; BM=Norwegian Bokmål; NN=Norwegian Nynorsk.

Instruction following 🖐️

- **Instruction:** Skriv dette som et eventyr!
- **Context:** Opplysningstiden var en samfunnskritisk åndsretning i Europa på 1700-tallet (fra 1688 til 1789) som uttrykte en tro på alle menneskers evner og fremmet renessansens humanistiske idealer...
- **Human response:** Det var en gang i Europa for mange hundre år siden et kongerike med all makt...

Norwegian-specific & world knowledge 🌐

- **Question:** "Æ træng læsta: Læsta er kjekt å ha. I alle fall sånn innimellom. Men hva er det for noe?"
- **Option A:** "Venner"
- **Option B:** "Lesestoff"
- **Option C:** "Ro"
- **Option D:** "Ullsokker"

- **Main idea**

- Do not update the LLM weights
- Show k examples in the input

- **k -shot evaluation regimes with $k \in \{0, 1, 16\}$**

- Probability estimation: The LLM scores each answer option conditioned on the prompt
- Generation: The LLM generates a response conditioned on the prompt

tekst: God betjening

sentiment: positiv

tekst: Vi ble ikke imponert over lydstyrken.

sentiment: negativ

tekst: Og han gjør det enormt godt.

sentiment: _____

språkmodell

positiv

In this talk

- Overall results
- Effect of instruction-tuning

More in our paper

- Agreement on LLM rankings
- Analysis of the LLMs' outputs
- LLMs vs. Human Baselines
- Comparison of BM vs. NN
- LLM-as-a-judge

Benchmarking LLMs: Key Results



The LLMs perform well on sentiment analysis and MT tasks, **but struggle with:**

Model	Overall	Borda's Count ↑	Norwegian language knowledge	Sentiment analysis	Commonsense reasoning	Truthfulness	Norwegian-specific & world knowledge	Machine reading comprehension	Text summarization	Instruction following	Machine translation
NB-GPT-6B	33.0	42.0	30.6	34.2	27.9	33.0	29.6	7.8	39.3	<u>39.1</u>	55.1
GPT-SW3-6.7B	45.1	63.0	61.0	64.2	31.3	<u>43.9</u>	30.0	30.1	37.7	35.5	72.6
NorwAI-Mistral-7B	45.5	69.0	47.2	70.7	<u>35.9</u>	36.7	39.5	37.1	31.9	37.7	<u>73.2</u>
NorwAI-Llama2-7B	44.1	59.0	47.9	66.3	29.8	30.2	35.4	38.8	37.5	37.7	72.9
NorBLOOM-7B-warm	35.6	28.0	51.8	40.8	23.5	39.1	23.3	23.9	35.6	13.9	68.8
NorMistral-7B-scratch	38.5	32.0	53.2	57.5	27.7	40.3	25.4	22.3	35.9	14.9	69.7
Viking-7B	41.9	47.0	51.3	59.5	27.4	26.6	25.0	25.9	49.4	38.7	73.0
NorMistral-11B	54.4	94.0	43.0	82.2	45.4	23.4	64.7	<u>59.5</u>	51.7	46.3	73.4
Viking-13B	45.2	69.0	56.8	67.0	31.9	28.3	30.5	30.7	49.3	38.8	73.1
NorMistral-7B-warm	43.6	61.0	<u>59.2</u>	68.7	34.0	31.6	38.7	40.7	33.0	14.6	72.0
NorMistral-7B-warm-IT	40.9	13.0	16.9	77.2	35.2	24.7	49.3	23.4	<u>54.8</u>	56.1	30.5
Mistral-7B	39.7	38.0	23.4	77.7	21.1	46.0	43.5	47.1	29.5	11.6	57.5
Mistral-7B-IT	37.7	4.0	12.8	69.5	19.9	31.9	34.8	31.7	46.2	50.4	42.5
AI-Sweden/Llama-3-8B	<u>51.3</u>	<u>84.0</u>	51.0	<u>80.3</u>	34.8	31.4	54.8	47.1	52.9	38.1	71.5
AI-Sweden/Llama-3-8B-IT	45.7	16.0	<u>16.1</u>	83.2	53.0	12.3	<u>55.3</u>	53.9	48.2	50.1	38.9
Meta/Llama-3-8B	47.0	64.0	28.4	76.8	28.0	34.0	50.9	48.7	<u>53.0</u>	37.4	66.1
Meta/Llama-3-8B-IT	<u>48.2</u>	<u>17.0</u>	13.7	78.3	39.1	<u>39.5</u>	51.8	<u>61.4</u>	51.1	51.4	47.1
Mistral-Nemo-12B	47.6	54.0	26.3	76.8	25.4	29.7	<u>55.0</u>	63.4	<u>50.9</u>	33.5	67.0
Mistral-Nemo-12B-IT	52.1	33.0	<u>16.1</u>	<u>82.9</u>	<u>44.1</u>	<u>42.7</u>	58.8	67.3	57.3	<u>55.7</u>	<u>43.7</u>

Benchmarking LLMs: Key Results



The LLMs perform well on sentiment analysis and MT tasks, **but struggle with:**

- Norwegian language knowledge
- commonsense reasoning
- truthfulness
- instruction following

Model	Overall	Borda's Count ↑	Norwegian language knowledge	Sentiment analysis	Commonsense reasoning	Truthfulness	Norwegian-specific & world knowledge	Machine reading comprehension	Text summarization	Instruction following	Machine translation
NB-GPT-6B	33.0	42.0	30.6	34.2	27.9	33.0	29.6	7.8	39.3	<u>39.1</u>	55.1
GPT-SW3-6.7B	45.1	63.0	61.0	64.2	31.3	<u>43.9</u>	30.0	30.1	37.7	35.5	72.6
NorwAI-Mistral-7B	45.5	69.0	47.2	70.7	<u>35.9</u>	36.7	39.5	37.1	31.9	37.7	<u>73.2</u>
NorwAI-Llama2-7B	44.1	59.0	47.9	66.3	29.8	30.2	35.4	38.8	37.5	37.7	72.9
NorBLOOM-7B-warm	35.6	28.0	51.8	40.8	23.5	39.1	23.3	23.9	35.6	13.9	68.8
NorMistral-7B-scratch	38.5	32.0	53.2	57.5	27.7	40.3	25.4	22.3	35.9	14.9	69.7
Viking-7B	41.9	47.0	51.3	59.5	27.4	26.6	25.0	25.9	49.4	38.7	73.0
NorMistral-11B	54.4	94.0	43.0	82.2	45.4	23.4	64.7	<u>59.5</u>	51.7	46.3	73.4
Viking-13B	45.2	69.0	56.8	67.0	31.9	28.3	30.5	30.7	49.3	38.8	73.1
NorMistral-7B-warm	43.6	61.0	<u>59.2</u>	68.7	34.0	31.6	38.7	40.7	33.0	14.6	72.0
NorMistral-7B-warm-IT	40.9	13.0	16.9	<u>77.2</u>	35.2	24.7	<u>49.3</u>	<u>23.4</u>	<u>54.8</u>	56.1	30.5
Mistral-7B	39.7	38.0	23.4	77.7	21.1	46.0	43.5	47.1	29.5	11.6	57.5
Mistral-7B-IT	37.7	4.0	12.8	<u>69.5</u>	19.9	31.9	34.8	31.7	46.2	50.4	42.5
AI-Sweden/Llama-3-8B	<u>51.3</u>	<u>84.0</u>	51.0	<u>80.3</u>	34.8	31.4	54.8	47.1	52.9	38.1	71.5
AI-Sweden/Llama-3-8B-IT	<u>45.7</u>	16.0	16.1	83.2	53.0	12.3	<u>55.3</u>	<u>53.9</u>	48.2	50.1	38.9
Meta/Llama-3-8B	47.0	64.0	28.4	76.8	28.0	34.0	50.9	48.7	<u>53.0</u>	37.4	66.1
Meta/Llama-3-8B-IT	<u>48.2</u>	<u>17.0</u>	13.7	<u>78.3</u>	<u>39.1</u>	<u>39.5</u>	51.8	<u>61.4</u>	51.1	51.4	47.1
Mistral-Nemo-12B	47.6	54.0	26.3	76.8	25.4	29.7	<u>55.0</u>	63.4	<u>50.9</u>	33.5	67.0
Mistral-Nemo-12B-IT	52.1	33.0	16.1	<u>82.9</u>	<u>44.1</u>	<u>42.7</u>	58.8	67.3	57.3	<u>55.7</u>	<u>43.7</u>

Benchmarking LLMs: Key Results



Instruction-tuning has both positive and negative effects

The most notable improvements are on:

- multiple-choice QA &
- sequence-to-sequence generation tasks

Model	Overall	Borda's Count ↑	Norwegian language knowledge	Sentiment analysis	Commonsense reasoning	Truthfulness	Norwegian-specific & world knowledge	Machine reading comprehension	Text summarization	Instruction following	Machine translation
NB-GPT-6B	33.0	42.0	30.6	34.2	27.9	33.0	29.6	7.8	39.3	<u>39.1</u>	55.1
GPT-SW3-6.7B	45.1	63.0	61.0	64.2	31.3	<u>43.9</u>	30.0	30.1	37.7	35.5	72.6
NorwAI-Mistral-7B	45.5	69.0	47.2	70.7	<u>35.9</u>	36.7	39.5	37.1	31.9	37.7	<u>73.2</u>
NorwAI-Llama2-7B	44.1	59.0	47.9	66.3	29.8	30.2	35.4	38.8	37.5	37.7	72.9
NorBLOOM-7B-warm	35.6	28.0	51.8	40.8	23.5	39.1	23.3	23.9	35.6	13.9	68.8
NorMistral-7B-scratch	38.5	32.0	53.2	57.5	27.7	40.3	25.4	22.3	35.9	14.9	69.7
Viking-7B	41.9	47.0	51.3	59.5	27.4	26.6	25.0	25.9	49.4	38.7	73.0
NorMistral-11B	54.4	94.0	43.0	82.2	45.4	23.4	64.7	<u>59.5</u>	51.7	46.3	73.4
Viking-13B	45.2	69.0	56.8	67.0	31.9	28.3	30.5	30.7	49.3	38.8	73.1
NorMistral-7B-warm	43.6	61.0	<u>59.2</u>	68.7	34.0	31.6	38.7	40.7	33.0	14.6	72.0
NorMistral-7B-warm-IT	<u>40.9</u>	13.0	16.9	<u>77.2</u>	35.2	24.7	<u>49.3</u>	23.4	<u>54.8</u>	56.1	30.5
Mistral-7B	39.7	38.0	23.4	77.7	21.1	46.0	43.5	47.1	29.5	11.6	57.5
Mistral-7B-IT	<u>37.7</u>	4.0	12.8	69.5	19.9	31.9	34.8	31.7	46.2	50.4	42.5
AI-Sweden/Llama-3-8B	<u>51.3</u>	<u>84.0</u>	51.0	<u>80.3</u>	34.8	31.4	54.8	47.1	52.9	38.1	71.5
AI-Sweden/Llama-3-8B-IT	<u>45.7</u>	16.0	16.1	83.2	53.0	12.3	<u>55.3</u>	53.9	48.2	50.1	38.9
Meta/Llama-3-8B	47.0	64.0	28.4	76.8	28.0	34.0	50.9	48.7	<u>53.0</u>	37.4	66.1
Meta/Llama-3-8B-IT	<u>48.2</u>	<u>17.0</u>	13.7	78.3	39.1	<u>39.5</u>	51.8	<u>61.4</u>	51.1	51.4	47.1
Mistral-Nemo-12B	47.6	54.0	26.3	76.8	25.4	29.7	<u>55.0</u>	63.4	<u>50.9</u>	33.5	67.0
Mistral-Nemo-12B-IT	52.1	33.0	16.1	<u>82.9</u>	<u>44.1</u>	<u>42.7</u>	58.8	67.3	57.3	<u>55.7</u>	<u>43.7</u>

Benchmarking LLMs: Key Results



Instruction-tuning has both positive and negative effects

Decreased performance on Norwegian language knowledge and MT tasks due to:

- Language switching
- Misinterpretation of instructions
- Generation issues

Model	Overall	Borda's Count ↑	Norwegian language knowledge	Sentiment analysis	Commonsense reasoning	Truthfulness	Norwegian-specific & world knowledge	Machine reading comprehension	Text summarization	Instruction following	Machine translation
NB-GPT-6B	33.0	42.0	30.6	34.2	27.9	33.0	29.6	7.8	39.3	<u>39.1</u>	55.1
GPT-SW3-6.7B	45.1	63.0	61.0	64.2	31.3	<u>43.9</u>	30.0	30.1	37.7	35.5	72.6
NorwAI-Mistral-7B	45.5	69.0	47.2	70.7	<u>35.9</u>	36.7	39.5	37.1	31.9	37.7	<u>73.2</u>
NorwAI-Llama2-7B	44.1	59.0	47.9	66.3	29.8	30.2	35.4	38.8	37.5	37.7	72.9
NorBLOOM-7B-warm	35.6	28.0	51.8	40.8	23.5	39.1	23.3	23.9	35.6	13.9	68.8
NorMistral-7B-scratch	38.5	32.0	53.2	57.5	27.7	40.3	25.4	22.3	35.9	14.9	69.7
Viking-7B	41.9	47.0	51.3	59.5	27.4	26.6	25.0	25.9	49.4	38.7	73.0
NorMistral-11B	54.4	94.0	43.0	82.2	45.4	23.4	64.7	<u>59.5</u>	51.7	46.3	73.4
Viking-13B	45.2	69.0	56.8	67.0	31.9	28.3	30.5	30.7	49.3	38.8	73.1
NorMistral-7B-warm	43.6	61.0	<u>59.2</u>	68.7	34.0	31.6	38.7	40.7	33.0	14.6	72.0
NorMistral-7B-warm-IT	40.9	13.0	16.9	77.2	35.2	24.7	49.3	23.4	<u>54.8</u>	56.1	30.5
Mistral-7B	39.7	38.0	23.4	77.7	21.1	46.0	43.5	47.1	29.5	11.6	57.5
Mistral-7B-IT	37.7	4.0	12.8	69.5	19.9	31.9	34.8	31.7	46.2	50.4	42.5
AI-Sweden/Llama-3-8B	<u>51.3</u>	<u>84.0</u>	51.0	<u>80.3</u>	34.8	31.4	54.8	47.1	52.9	38.1	71.5
AI-Sweden/Llama-3-8B-IT	45.7	16.0	16.1	83.2	53.0	12.3	<u>55.3</u>	53.9	48.2	50.1	38.9
Meta/Llama-3-8B	47.0	64.0	28.4	76.8	28.0	34.0	50.9	48.7	<u>53.0</u>	37.4	66.1
Meta/Llama-3-8B-IT	<u>48.2</u>	<u>17.0</u>	13.7	78.3	39.1	<u>39.5</u>	51.8	<u>61.4</u>	51.1	51.4	47.1
Mistral-Nemo-12B	47.6	54.0	26.3	76.8	25.4	29.7	<u>55.0</u>	63.4	<u>50.9</u>	33.5	67.0
Mistral-Nemo-12B-IT	52.1	33.0	16.1	<u>82.9</u>	<u>44.1</u>	<u>42.7</u>	58.8	67.3	57.3	<u>55.7</u>	43.7

Language Generation Evaluation

- **Performance metrics:**
 - Do not capture the diversity of possible outputs
 - Can poorly correlate with humans w.r.t. criteria
- **Human-based evaluation:**
 - Super resource-consuming
 - May not be reproducible
- **LLM-based evaluation:**
 - Suffers from biases
 - Lack of strong judge LLMs for non-English languages



Summarize this news article in three sentences: {{article}}



LLM A



LLM B



Human response



Which LLM is better?

Data contamination

- Evaluation data appears in:
 - Pretraining data
 - Post-training data
- Potential solutions:
 - Responsible data curation
 - Responsible development
 - Responsible evaluation

Dataset	Task	Release date	Train split	Dev split	Test split
CoNLL03	IE	2003	Contaminated	Contaminated	Contaminated
ACE05	IE	2005	Suspicious	Suspicious	Suspicious
OntoNotes	IE	2013	Clean	Clean	Clean
SQuAD	QA	2018	Contaminated	Contaminated	N/A
MNLI	NLI	2018	Contaminated	Contaminated	N/A
QuAC	QA	2019	Suspicious	Suspicious	N/A
Natural Questions	QA	2019	Suspicious	Suspicious	N/A
BoolQ	QA/TC	2019	Suspicious	Suspicious	N/A
GSM8K	Reasoning	2021	Clean	N/A	Clean

Contaminated data in ChatGPT. [LM Contamination Index](#).

Benchmark creation

- **Evaluation data is generated:**
 - What do we want to evaluate?
 - Questionable data quality
 - Language, culture, and society values reflection
- Potential solutions:
 - Thorough human curation of synthetic data
 - Automatic data quality criteria & filtering

English:

Which of the following is not included in the U.S. GDP?

- (A) The U.S. military opens a new base in a foreign country with 1000 U.S. personnel.
- (B) Japanese consumers buy thousands of CDs produced in the United States.
- (C) An American pop singer performs a sold-out concert in Paris.**
- (D) A French theatrical production tours dozens of American cities.

Norwegian Bokmål:

Hvilken av følgende er ikke inkludert i USAs BNP?

- (A) Det amerikanske militæret åpner en ny base i et fremmed land med 1000 amerikanske soldater.
- (B) Japanske forbrukere kjøper tusenvis av CD-er produsert i USA.
- (C) En amerikansk popsanger holder en utsolgt konsert i Paris.**
- (D) En fransk teaterproduksjon turnerer i dusinvis av amerikanske byer.



An example from MMLU and its translation by ChatGPT.