

NorEval: A Norwegian Language Understanding and Generation Evaluation Benchmark

Anonymous ACL submission

Abstract

This paper introduces NorEval, a new and comprehensive evaluation suite for large-scale standardized benchmarking of Norwegian generative language models (LMs). NorEval consists of 24 high-quality human-created datasets – of which five are created from scratch. In contrast to existing benchmarks for Norwegian, NorEval covers a broad spectrum of task categories targeting Norwegian language understanding and generation, establishes human baselines, and focuses on both of the official written standards of the Norwegian language: Bokmål and Nynorsk. All our datasets and a collection of over 100 human-created prompts are integrated into LM Evaluation Harness, ensuring flexible and reproducible evaluation. We describe the NorEval design and present the results of benchmarking 19 open-source pre-trained and instruction-tuned LMs for Norwegian in various scenarios. Our benchmark, evaluation framework, and annotation materials will be made publicly upon acceptance.

1 Introduction

The advancement of language models (LMs) is inseparable from benchmarking – the systematic evaluation of their generalization abilities on standardized datasets across various criteria (Ruder, 2021; Srivastava et al., 2023). Despite its crucial role, benchmarking in resource-lean scenarios remains scarce due to the lack of diverse evaluation suites for low-resource languages, including Norwegian (Joshi et al., 2020; Hedderich et al., 2021).

Previous work focuses on Norwegian as part of medium-scale benchmarking efforts – NorBench (Samuel et al., 2023) and NLEBench (Liu et al., 2024) – and broader Mainland Scandinavian evaluation initiatives – ScandEval (Nielsen, 2023) and Scandinavian Embedding Benchmark (SEB; Enevoldsen et al., 2024). However, these benchmarks have several shortcomings that limit the scope of LM evaluation in Norwegian.

- **Coverage and design.** These benchmarks exhibit a significant dataset overlap with a low variation in task formulations. NorBench and ScandEval cover traditional NLP tasks, SEB addresses text embedding evaluation, and NLEBench comprises a narrow spectrum of Norwegian language generation tasks.
- **Data quality.** NLEBench and ScandEval include machine-translated English datasets, introducing potential evaluation biases that may conflict with Norwegian-specific values and knowledge.
- **Linguistic diversity.** Norwegian has two official written standards: Bokmål (BM) and Nynorsk (NN; the minority variant). The latter variant remains significantly underrepresented in previous work.
- **Human performance.** No existing benchmark establishes human baselines, which is a standard practice to determine upper LM performance bounds.

This paper introduces NorEval, a novel large-scale evaluation suite designed to benchmark Norwegian LMs on language understanding and generation tasks. NorEval comprises 24 human-created datasets across nine task categories, including sentiment analysis, Norwegian language knowledge, Norwegian-specific & world knowledge, machine reading comprehension, commonsense reasoning, machine translation, text summarization, instruction following, and truthfulness. Our design enables various benchmarking scenarios, ranging from multi-prompt k -shot evaluation to side-by-side LM comparison on diverse user instructions.

Our main contributions are: (i) we create NorEval, the largest multi-task benchmark for Bokmål and Nynorsk that combines 19 existing peer-reviewed datasets with five datasets created from scratch; (ii) we curate a collection of over 100







Evaluation Scope		Task Categories	# Datasets			Method		
			BM	NN	Total			
NorBench	NLU & NLG	POS-tagging, MT, NER, sentiment analysis, Acceptability classification, MRC	8	2	10	✓	✗	✗
ScandEval	NLU	NER, sentiment analysis, Acceptability classification, MRC, Commonsense reasoning, Text summarization, multiple-choice QA	7	2	9	✓	✓	✗
SEB	Text embedding evaluation	LID, sentiment analysis, Acceptability classification, retrieval, Dialect & written form pairing, Intent & scenario classification, Clustering, political speech classification	11	3	14	✓	✗	✗
NLEBench	NLU & NLG	NLI, MRC, bias detection, Text summarization, yes/no QA, Instruction following, Paraphrase detection, open-ended conversation	9	✗	9	✗	✓	✓
NorEval	NLU & NLG	Commonsense reasoning, MRC, sentiment analysis, Norwegian language knowledge, MT, Truthfulness, text summarization, Instruction following, Norwegian-specific & world knowledge	16	8	24	✓	✗	✗

Table 1: **Comparison of multi-task benchmarks for Norwegian:** ScandEval (Nielsen, 2023), Scandinavian Embedding Benchmark (SEB; Enevoldsen et al., 2024), NorBench (Samuel et al., 2023), NLEBench (Liu et al., 2024), and NorEval (ours). BM=Norwegian Bokmål; NN=Norwegian Nynorsk; =human-created; =machine-translated; = GPT-4o-created & human-edited; NLU=Natural language understanding; NLG=Natural language generation; NER=named entity recognition; LID=language identification; MRC=machine reading comprehension; NLI=natural language inference; QA=question answering; MT=machine translation.

dataset-specific prompts for robust evaluation; (iii) we establish five human baselines; (iv) we benchmark 19 pretrained and instruction-tuned Norwegian LMs against each other and humans; and (v) we release NorEval, our evaluation framework, all annotation materials, and a public leaderboard.¹

2 Background

Norwegian Bokmål and Nynorsk BM is the primary written standard, while an estimated 10–15% of the Norwegian population uses NN – especially in Western Norway. The national language legislation specifies that minimally 25% of the written public service information should be in NN to ensure representation of both varieties. While BM and NN are closely related, they exhibit lexical and grammatical differences, e.g. distinct pronouns, plural noun forms, definite noun forms, verb conjugation, and vocabulary units. We make one of the first dedicated attempts to increase the representation of NN in the context of LM evaluation.

¹The links are removed for anonymity purposes.

Norwegian Benchmarks Table 1 provides an overview of existing Norwegian benchmarks w.r.t. the evaluation scope, task categories, the number of datasets, representation of BM and NN, and dataset creation method. We describe them below.

- NorBench** is primarily designed to benchmark encoder-only LMs on a collection of ten traditional NLP tasks, such as PoS-tagging, NER (NorNE; Jørgensen et al., 2020), sentiment analysis at different levels of granularity (NoReC; Velldal et al., 2018; Øvrelid et al., 2020), acceptability classification (NoCoLA; Jentoft and Samuel, 2023), machine translation, and extractive question answering (NorQuAD; Ivanova et al., 2023). All datasets in NorBench are human-created; however, the support for NN is limited to PoS-tagging and NER based on the Norwegian UD treebanks (Øvrelid and Hohle, 2016; Velldal et al., 2017).
- ScandEval** is an evaluation suite coupled with a public leaderboard for Scandinavian languages: Danish, Faroese, Icelandic, Norwegian and Swedish. The Norwegian datasets in Scan-

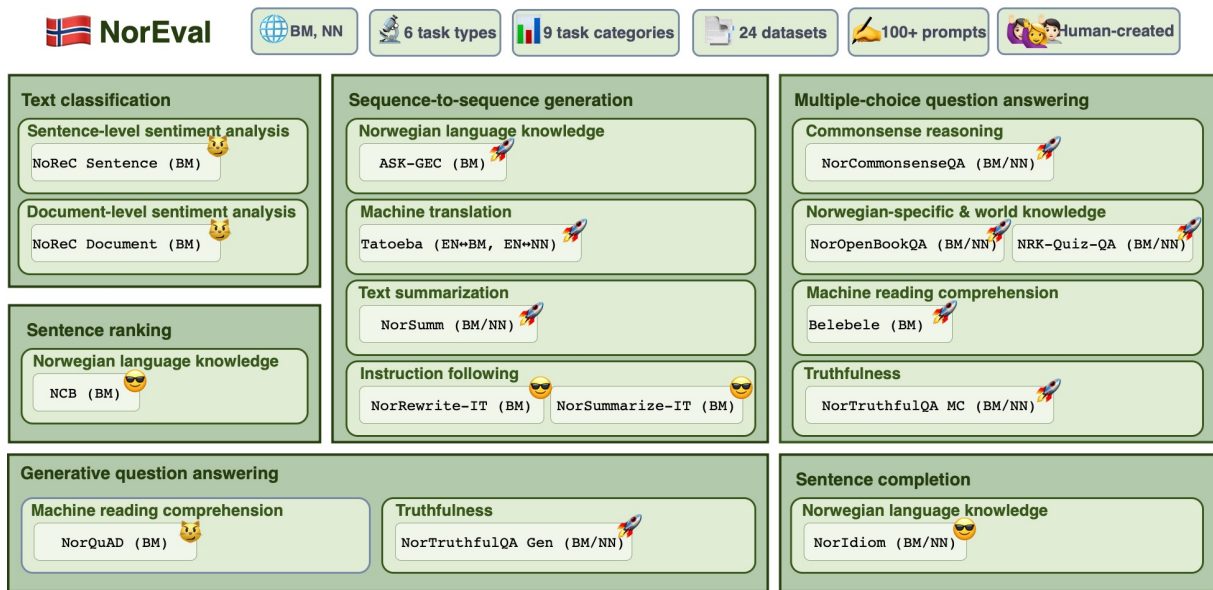


Figure 1: **Overview of the NorEval design.** 🧑 denotes datasets used in previous studies (§2), 🚀 represents datasets that have not been used in Norwegian benchmarks, and 🤖 denotes our datasets introduced as part of NorEval. EN=English; NB=Norwegian Bokmål; NN=Norwegian Nynorsk.

dEval are based on existing resources, such as NoReC, NorNE, NorQuAD, and the SNL & VG summarization dataset (Navjord and Korsvik, 2023). ScandEval introduces ScaLA, a novel acceptability classification dataset created through rule-based perturbation of sentences from the Norwegian UD treebanks. Moreover, it contains three officially supported machine-translated English datasets that are not curated or post-processed: MMLU (Hendrycks et al., 2020), XSum (Narayan et al., 2018), and HellaSwag (Zellers et al., 2019). Similar to NorBench, the coverage of NN is limited to the datasets derived from the Norwegian UD treebanks.

3. **SEB** is designed to evaluate text representations for Scandinavian languages across retrieval, bi-text mining, text classification, and clustering tasks. With its distinct focus on text embedding models, SEB has little overlap with other Norwegian benchmarks (except for NorQuAD, ScaLA, and SNL & VG) and primarily constructs its evaluation tasks by converting existing Norwegian resources and leveraging supported metadata and schemes.
4. **NLEBench** is designed to evaluate the LM’s Norwegian language generation capabilities. Although NLEBench covers a diverse range of task categories, it does not address any NN evaluation scenario. Moreover, seven out of nine

datasets are machine-translated without curation, raising concerns about the benchmark’s reliability. The remaining two datasets comprise multi-turn conversation, closed question answering (QA), and abstractive summarization tasks; these are generated by GPT-4o and edited by a team Norwegian native speakers.

NorEval expands the benchmarking scope of Norwegian LMs to task categories and evaluation scenarios that have not been covered in the related studies. To the best of our knowledge, our work describes the largest empirical evaluation in the context of BM and NN to date.

3 NorEval

Our main goal is to develop a high-quality standardized evaluation suite to benchmark Norwegian generative LMs across a broad spectrum of Norwegian language understanding and generation tasks. Figure 1 outlines the design of NorEval, which combines 19 existing peer-reviewed datasets with five novel datasets (§3.1), comprises a pool of over 100 prompts (§3.2), and offers a framework for systematic and reproducible LM evaluation (§3.3).

3.1 Tasks

Due to space constraints, Appendix A presents an overview of our 24 datasets, including dataset descriptions and examples, task formu-

lations, prompts, performance metrics, and general statistics. [Appendix B](#) details our novel datasets (NCB, NorIdiom, NorRewrite-instruct, and NorSummarize-Instruct), including general statistics and creation methods. We describe NorEval based on nine high-level task categories:

Sentiment analysis focuses on a binary polarity classification at the sentence- and document-level (NoReC Sentence & Document).

Norwegian language knowledge assesses an LM’s ability to perform grammatical error correction (ASK-GEC; [Jentoft, 2023](#)), adhere to language-specific punctuation rules (NCB; ours), and complete Norwegian idioms (NorIdiom; ours).

Norwegian-specific & world knowledge assesses an LM’s ability to answer multiple-choice questions based on real-world and Norwegian-specific cultural knowledge (NRK-Quiz-QA and NorOpen-BookQA; [Mikhailov et al., 2025](#)).

Machine reading comprehension evaluates the ability of LMs to answer questions related to an input text by selecting an answer from multiple choices (Belebele; [Bandarkar et al., 2024](#)) or generating a text span (NorQuAD).

Commonsense reasoning assesses an LM’s ability to answer a multiple-choice question based on logical reasoning and world understanding (Nor-CommonsenseQA; [Mikhailov et al., 2025](#)).

Machine translation tests how accurately an LM translates sentences among four language pairs from Tatoeba ([Tiedemann, 2020](#)): English \leftrightarrow BM and English \leftrightarrow NN.

Text summarization focuses on abstractive summarization of news articles (NorSumm; [Touileb et al., 2025](#)).

Instruction following aims to evaluate an LM’s ability to follow user instructions on creative rewriting and summarization through (i) changing a text’s tone, style, and genre; (ii) simplifying complex content; (iii) enhancing readability; (iv) changing content for specific target audience (NorRewrite-Instruct and NorSummarize-Instruct; ours).

Truthfulness tests whether an LM generates or selects answers that propagate false beliefs and misconceptions (NorTruthfulQA Multiple Choice & Generation; [Mikhailov et al., 2025](#)).

3.2 Prompts

We conduct a two-stage in-house annotation to create a collection of prompts that reflect diverse user

formulations and answer formatting, with four-to-six prompts per dataset. The prompt examples are provided in [Appendix A](#), and the annotation guidelines are documented in [Appendix C](#).

• Stage 1: Creating Prompts in Bokmål.

Three Norwegian native speakers create dataset-specific prompts in BM using two strategies: (i) manually translating provided English prompts from PromptSource ([Bach et al., 2022](#)) and (ii) writing the prompts from scratch.

• Stage 2: Adapting Prompts to Nynorsk.

We hire a BA student in linguistics to adapt the BM prompts to NN. The hourly pay rate is 227 NOK (approx. \$20).

3.3 Evaluation Framework

All our datasets and prompts are integrated into LM Evaluation Harness ([Gao et al., 2024](#); [Biderman et al., 2024](#)), a framework for flexible evaluation of generative LLMs in various scenarios. The framework provides a user-friendly API allowing to easily integrate datasets, configure prompts, and benchmark LMs that are not part of our baselines in our public leaderboard (§4).²

4 Evaluation Setup

We benchmark a broad range of 19 pretrained and instruction-finetuned decoder-only LMs that are publicly available in Transformers ([Wolf et al., 2020](#)) as our baselines (see [Table 2](#)). We compare them in k -shot regimes against one another and our human baselines, and evaluate the instruction-finetuned LMs using the LLM-as-a-judge approach ([Zheng et al., 2023](#)).

In-context Learning Evaluation The evaluation is run in k -shot regimes with $k \in \{0, 1, 16\}$ across *all* prompts. We use the maximum k for each dataset, which depends on the availability of a training/development set for demonstration examples and the example lengths. We use two strategies supported via LM Evaluation Harness to evaluate the LM performance in a prompted format:

- **Log-likelihood.** The LM assigns a probability to each answer candidate conditioned on an input prompt, and the most probable candidate is selected as the prediction. This strategy is used in the sentence ranking, text classification, and multiple-choice QA tasks.

²The link is removed for anonymity purposes.

Name	Base
PRETRAINED LMS	
Mistral-7B	N/A
Mistral-Nemo-12B	N/A
Meta/Llama-3-8B	N/A
NB-GPT-6B	N/A
NorwAI-Mistral-7B	Mistral-7B
NorwAI-Llama2-7B	Llama-2-7B
AI-Sweden/GPT-SW3-6.7B	N/A
AI-Sweden/Llama-3-8B	Meta/Llama-3-8B
LumiOpen/Viking-7B	N/A
LumiOpen/Viking-13B	N/A
NorBLOOM-7B-scratch	N/A
NorMistral-7B-scratch	N/A
NorMistral-7B-warm	Mistral-7B
NorMistral-11B-warm	Mistral-Nemo-12B
INSTRUCTION-TUNED LMS	
NorMistral-7B-warm-IT	NorMistral-7B-warm
Mistral-7B-IT	Mistral-7B
AI-Sweden/Llama-3-8B-IT	AI-Sweden/Llama-3-8B
Meta/Llama-3-8B-IT	Meta/Llama-3-8B
Mistral-Nemo-12B-IT	Mistral-Nemo-12B

Table 2: **The LMs used in our work and their base versions.** LM references: Mistral-7B (Jiang et al., 2023), NorBLOOM/NorMistral-7B-scratch & Normistral-7B/11B-warm (Samuel et al., 2025), and Meta/Llama-3-8B (Dubey et al., 2024).

- **Generation.** The LM generates a text continuation conditioned on an input prompt. We use a greedy search decoding method for the pre-trained LMs and recommended HuggingFace inference hyperparameters and chat templates for the instruction-finetuned LMs. This strategy is used in the sentence completion, sequence-to-sequence generation, and generative QA tasks.

Performance Aggregation We use a combination of performance aggregation methods based on well-established NLP benchmarking practices and theoretical foundations of the social choice theory (Arrow, 2012).

- **Multi-prompt Aggregation.** We select the highest performance score for each LM across dataset-specific prompts to mitigate the prompt sensitivity (Voronov et al., 2024).
- **Average Normalized Score.** In line with the OpenLLM leaderboard (Fourrier et al., 2024) and FineWeb 2 evaluation protocol (Penedo et al., 2024), we first rescale individual performance scores across our nine task cate-

Dataset	WAWA
NCB	92.0
NorOpenBookQA (BM)	98.0
NorCommonsenseQA (BM)	93.3
NorTruthfulQA Multiple Choice (BM)	86.0
Belebele	86.7

Table 3: **The WAWA rates for human baselines (§4).**

gories. Rescaling involves score normalization between the random baseline and the maximum possible score. We then compute the overall performance score by averaging the normalized scores within all task categories.

- **Borda’s Count.** Recent works demonstrate the effectiveness of using Borda’s count as an alternative to arithmetic mean aggregation in multi-task benchmarking (Colombo et al., 2022; Rofin et al., 2023). This approach relies on a scoring vector $c = (|M| - 1, |M| - 2, \dots, 1, 0)$ to assign scores to a set of M LMs $m \in \{m_1, \dots, m_{|M|}\}$ based on their positions in each task-specific ranking. The final score of each LM is calculated as the sum of corresponding scores in each task $Sc(m) = \sum_{i=1}^{|M|} c_i p_i(m)$, where $p_i(m)$ is the number of tasks in which LM m takes the i^{th} place, and c_i is the i^{th} element of c . Borda’s count allows for aggregating heterogeneous performance metrics while accounting for the differences in the LMs’ ranking positions.

Human Baselines We establish five human baselines on random subsets of 50 examples from NCB, Belebele, NorOpenBookQA (BM), NorCommonsenseQA (BM), and NorTruthfulQA Multiple choice (BM). Our annotation team consists of 12 volunteers, all Norwegian native speakers with an NLP background and completed higher academic degrees. Before starting, the annotators receive guidelines describing the datasets and providing examples with explanations (see Appendix D). Each example is annotated by three annotators, and we use majority voting to aggregate their results. We compute the inter-annotator agreement rates using the Worker Agreement with Aggregate (WAWA) coefficient (Ning et al., 2018), which represents the average percentage of annotators’ votes that align with the majority vote for each example. Table 3 presents the per-dataset WAWA rates ranging between 86% and 98%, which indicates a strong agreement between our annotators.

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	39.5	<u>39.1</u>	55.1
GPT-SW3-6.7B	45.3	63.0	61.0	64.2	31.3	<u>43.9</u>	30.0	30.1	39.2	35.5	72.6
NorwAI-Mistral-7B	45.4	69.0	47.2	70.7	<u>35.9</u>	36.7	39.5	37.1	30.9	37.7	<u>73.2</u>
NorwAI-Llama2-7B	43.6	57.0	47.9	66.3	29.8	30.2	35.4	38.8	33.6	37.7	72.9
NorBLOOM-7B-warm	35.4	28.0	51.8	40.8	23.5	39.1	23.3	23.9	33.7	13.9	68.8
NorMistral-7B-scratch	38.6	34.0	53.2	57.5	27.7	40.3	25.4	22.3	36.8	14.9	69.7
Viking-7B	41.8	47.0	51.3	59.5	27.4	26.6	25.0	25.9	49.1	38.7	73.0
NorMistral-11B	54.6	96.0	43.0	82.2	45.4	23.4	64.7	<u>59.5</u>	53.8	46.3	73.4
Viking-13B	45.3	71.0	56.8	67.0	31.9	28.3	30.5	30.7	50.5	38.8	73.1
NorMistral-7B-warm	43.3	61.0	<u>59.2</u>	68.7	34.0	31.6	38.7	40.7	29.8	14.6	72.0
NorMistral-7B-warm-IT	41.0	14.0	16.9	<u>77.2</u>	35.2	24.7	49.3	23.4	<u>55.5</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	37.4	4.0	12.8	69.5	19.9	31.9	34.8	31.7	43.1	50.4	42.5
AI-Sweden/Llama-3-8B	<u>50.8</u>	<u>80.0</u>	51.0	<u>80.3</u>	34.8	31.4	54.8	47.1	48.3	38.1	71.5
AI-Sweden/Llama-3-8B-IT	45.5	16.0	<u>16.1</u>	83.2	53.0	12.3	<u>55.3</u>	53.9	46.9	50.1	38.9
Meta/Llama-3-8B	47.0	63.0	28.4	76.8	28.0	34.0	50.9	48.7	<u>53.1</u>	37.4	66.1
Meta/Llama-3-8B-IT	<u>48.1</u>	<u>17.0</u>	13.7	78.3	39.1	<u>39.5</u>	51.8	<u>61.4</u>	50.8	51.4	47.1
Mistral-Nemo-12B	47.8	55.0	26.3	76.8	25.4	29.7	<u>55.0</u>	63.4	<u>53.1</u>	33.5	67.0
Mistral-Nemo-12B-IT	51.9	33.0	<u>16.1</u>	<u>82.9</u>	<u>44.1</u>	<u>42.7</u>	58.8	67.3	56.0	<u>55.7</u>	<u>43.7</u>

Table 4: **Borda’s count and normalized performance scores** of the pretrained and instruction-tuned Norwegian LMs across all task categories in NorEval. Warm-colored cells represent cases where an instruction-tuned version improves performance compared to the base LM, while cold-colored cells indicate cases where it decreases. The best score is in bold, the second best is underlined – the pretrained and instruction-tuned LMs are highlighted independently.

LLM-as-a-judge We use the LLM-as-a-judge approach to automatically evaluate the instruction-tuned LMs’ generation abilities on NorRewrite-Instruct and NorSummarize-Instruct. We adopt the Human response-guided evaluation framework (HREF; [Lyu et al., 2024](#)), which relies on human references as additional inputs to improve the LM judgement performance. Our judge model is `meta-llama/Llama-3.3-70B-Instruct`, which highly correlates with human judgments as reported by [Lyu et al.](#). The judge model is given (i) the prompt; (ii) output A; (iii) output B; and (iv) a human reference formatted based on the prompt template in [Appendix F.2](#). We perform the side-by-side comparison using a greedy search decoding

strategy across three options: (i) output A is better than output B; (ii) output B is better than output A; and (iii) a tie. We conduct the side-by-side comparison over all combinations of the instruction-tuned LMs and compute the expected win rates (see [Appendix F](#) for further details).

5 Results

This section describes our empirical evaluation results on NorEval. We report the results aggregated across our task categories in [Table 4](#). We find that NorMistral-11B achieves the best overall performance across most task categories, followed by AI-Sweden/Llama-3-8B. NorMistral/NorBLOOM-7B-scratch and NB-GPT-6B receive the lowest scores.

Model	NORREWRITE-INSTRUCT						NORSUMMARIZE-INSTRUCT					
	NorMistral-7B-warm-IT	Mistral-Nemo-12B-IT	Mistral-7B-IT	Meta/Llama-3-8B-IT	AI-Sweden/Llama-3-8B-IT	Average	NorMistral-7B-warm-IT	Mistral-Nemo-12B-IT	Mistral-7B-IT	Meta/Llama-3-8B-IT	AI-Sweden/Llama-3-8B-IT	Average
NorMistral-7B-warm-IT	—	45.6	92.2	76.2	99.5	78.4	—	57.6	92.5	66.5	99.5	79.0
Mistral-Nemo-12B-IT	54.4	—	89.8	80.6	93.1	79.5	42.4	—	81.8	62.1	87.3	68.4
Mistral-7B-IT	7.8	10.2	—	47.4	67.5	33.2	7.5	18.2	—	36.9	66.9	32.4
Meta/Llama-3-8B-IT	23.8	19.4	52.6	—	64.7	40.1	33.5	37.9	63.1	—	71.4	51.5
AI-Sweden/Llama-3-8B-IT	0.5	6.9	32.5	35.3	—	18.8	0.5	12.7	33.1	28.6	—	18.7

Table 5: **Pair-wise expected win-rates (%)** of the instruction-finetuned LMs on our instruction-following tasks.

Mistral-Nemo-12B-IT performs best among the instruction-tuned LMs; however, the benefits from instruction-tuning depend on the task. In general, the LMs perform well on the sentiment analysis and machine translation tasks but struggle with tasks requiring the Norwegian language knowledge, commonsense reasoning, truthfulness, and instruction following. We summarize our findings below w.r.t. performance aggregation methods, human performance, task category, the effect of instruction tuning, Norwegian language variety, and LLM-as-a-judge evaluation.

Agreement on LM Rankings The agreement rate³ between the average normalized score and Borda’s count for the top-3 LMs is 66%. This discrepancy is because Borda’s count penalizes Mistral-Nemo-12B for its low performance on Norwegian language knowledge tasks, ranking NorMistral-11B and AI-Sweden/Llama-3-8B as the top-2 models, while Viking-13B takes third place instead of Mistral-Nemo-12B. However, the performance aggregation methods fully agree on the bottom-5 LMs, which include Viking-7B, Mistral-7B, NorMistral-7B-scratch, NorBLOOM-7B-warm, and NB-GPT-6B.

LMs vs. Human Baselines Comparing the LMs to our human baselines in Table 8 and Table 9 in Appendix E, we find that the LMs fall behind humans by 10% on Belebele, 14.4% on NorQuAD, 15.2% on NorOpenBookQA, 17.8% on NorCom-

monsenseQA, and 13.3% on NorTruthfulQA Multiple Choice. However, NorwAI-Llama2-7B slightly surpasses human performance on NCB by 1.2%. The results suggest that while LMs show promising in-context learning capabilities, there is still room for their improvement in world knowledge, truthfulness, and reading comprehension tasks.

Analysis on Task Categories We outline our key results based on the fine-grained analysis reported in Appendix E. No single LM consistently outperforms others across all task categories. The strongest performance is observed on the sentiment analysis tasks, with AI-Sweden/Llama-3-8B achieving the best score of 92.7 and its instruction-tuned version (NoReC Document) reaching 95.5. On NorIdiom, GPT-SW3-6.7B delivers the best performance, followed by NorMistral-7B-warm. For NorCommonsenseQA, the performance of pre-trained LMs varies: BM scores range from 41.2 to 61, while NN scores range from 32.6 (Mistral-7B) to 51.6 (NorMistral-11B), suggesting limited in-context learning abilities for logical reasoning. The LMs also exhibit strong performance on Norwegian-specific quizzes (NRK-Quiz-QA) and tasks assessing elementary-level world knowledge (NorOpenBookQA), with the best-performing LMs including NorMistral-11B, AI-Sweden/Llama-3-8B, Mistral-7B, and Mistral-Nemo-12B. However, the LMs tend to generate less truthful answers in the open-ended QA setup (NorTruthfulQA Generation) compared to the multiple-choice setup (NorTruthfulQA Multiple Choice), highlighting po-

³The proportion of top-k and bottom-k LMs that are consistently ranked by both performance aggregation methods.

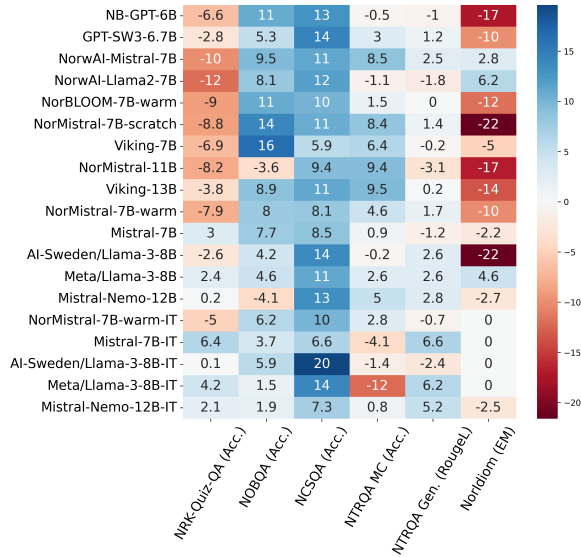


Figure 2: **Comparison of Bokmål and Nynorsk.** Heatmap that shows the performance δ -scores between BM and NN on our multiple-choice QA and sentence completion tasks. NOBQA=NorOpenBookQA; NCSQA=NorCommonsenseQA; NTRQA =NorTruthfulQA. Higher values mean higher performance in BM.

tential challenges of evaluating open-ended QA in Norwegian.

Effect of Instruction-tuning Instruction-tuning is one of the least explored research directions for Norwegian. Our results align with Wang et al. (2023); Bukharin et al. (2024), demonstrating that instruction-tuning can yield both positive and negative effects depending on the task. For instance, instruction-tuning consistently improves the performance of Mistral-Nemo-12B and Meta/Llama-3-8B across most task categories, with the most notable improvements observed in multiple-choice QA and sequence-to-sequence generation tasks. At the same time, it can degrade the performance on tasks requiring Norwegian language knowledge and involve translating from English into BM and NN (see Table 7 and Table 10 in Appendix E).

Comparing Bokmål and Nynorsk We compute the performance δ -scores on multiple-choice and sentence completion tasks with parallel BM and NN datasets to compare LMs w.r.t. the Norwegian language variety. Figure 2 shows that the LMs generally perform better on BM on NorOpenBookQA, NorCommonsenseQA, and NorTruthfulQA Multiple Choice as opposed to NRK-Quiz-QA and NorIdiom. Instruction-tuning results in lower δ -scores on NRK-Quiz-QA and NorOpen-

BookQA but leads to random guessing performance on NorIdiom for both BM and NN.

LLM-as-a-judge We report the LMs’ win-rates in Table 5. We find that NorMistral-7B-warm-IT and Mistral-Nemo-12B-IT consistently perform best across all LMs, while responses from AI-Sweden/Llama-3-8B-IT and Mistral-7B-IT are least preferred. NorMistral-7B-warm-IT achieves the highest win-rate on NorSummarize-Instruct, while there is a minor difference between the top-2 LMs on NorRewrite-Instruct. Our analysis of language and position biases in Appendix F indicates that the LMs often switch to English, Swedish, or Danish, and there is an insignificant effect of the response position on the judgment.

6 Conclusion and Future Work

This work introduces NorEval, the largest human-curated benchmark for assessing Norwegian language understanding and generation on 24 human-created datasets. NorEval focuses on both Norwegian language varieties and spans nine task categories, ranging from Norwegian-specific & world knowledge to instruction following. We benchmark 19 open-source Norwegian generative LMs against each other and our established human baselines, analyzing their performance in various scenarios. Additionally, we present one of the first extensive evaluations of open Norwegian instruction-tuned LMs and their base counterparts in k -shot regimes, as well as via the LLM-as-a-judge approach. Our key findings indicate that while the LMs perform well on sentiment analysis and machine translation tasks, they struggle with tasks requiring Norwegian language knowledge, commonsense reasoning, truthfulness, and instruction following. The LMs generally perform better on BM compared to NN. Notably, instruction-tuning yields both positive and negative effects on the LM performance.

Our *future* work includes: (i) a more detailed evaluation of instruction-tuned LMs and instruction-tuning data mixtures; (ii) integration of novel datasets; (iii) establishment of human baselines on additional tasks; (iv) integration of test data decontamination methods. We hope that our benchmark and evaluation framework will facilitate more comprehensive comparisons of LMs within the context of Mainland Scandinavian languages and inspire collaborative efforts among NLP researchers and developers to advance reliable LMs and evaluation resources for Norwegian.

7 Limitations

Evaluation Design While we present extensive empirical evaluations of a broad range of Norwegian LMs, we acknowledge several limitations in our evaluation design. First, in the one- and 16-shot evaluation scenarios, demonstration examples are randomly sampled, which can facilitate label bias in our text classification and multiple-choice QA tasks (Zhao et al., 2021). Next, aggregating evaluation results in multi-task benchmarking remains a challenging problem. We employ a combination of performance aggregation methods to mitigate the shortcomings of standard arithmetic mean aggregation: (i) score normalization to account for random baseline performance, and (ii) Borda’s count to address the heterogeneity of performance metrics. However, these methods have inherent limitations. In particular, we still need to average heterogeneous task-specific normalized performance scores to compute an overall score. Although Borda’s count relies on model rankings instead of performance scores, introducing a new LM can influence the final ranking due to the well-studies axiom of the independence of irrelevant alternatives (Arrow, 2012; Dougherty and Heckelman, 2020). Additionally, Borda’s count can treat several LMs as equivalent (or ties), which is not an empirical observation in our experiments. Furthermore, our work does not account for potential in-domain evaluation of the instruction-tuned LMs, which can be instruction-tuned on similar tasks in English and other languages, potentially inflating their downstream performance.

Data Contamination The increasing volume of open textual data can lead to unintended test data leakage in an LM’s pretraining corpus (e.g., Brown et al., 2020; Dubey et al., 2024; Zhang et al., 2024), which can promote the saturation of NLP benchmarks. We recognize the importance of this evaluation aspect and acknowledge that LM performance on NorEval datasets created from open text sources can be inflated. However, we are unable to provide a reliable estimate of the proportion of leaked data during this submission. We encourage adherence to responsible LM development practices and recommend conducting test contamination analysis when benchmarking an LM on NorEval. Integrating unsupervised pretraining data detection methods into NorEval is left as a direction for our future work.

Evaluation Framework NorEval is integrated into LM Harness Evaluation, a widely recognized open-source collaborative project that is subject to continuous improvements and advancements, which potentially affect its long-term compatibility, reproducibility, and usability.

LLM-as-a-judge Automatic side-by-side evaluation using the LLM-as-a-judge approach is a well-established, complementary evaluation scenario that has demonstrated its efficiency for high-resource languages. However, its performance in low-resource languages remains unclear. We acknowledge that the reliability of our evaluation results in the LLM-as-a-judge experiments requires further empirical validation. We limit our analysis to language and position bias; other potential evaluation directions for the judges are beyond the scope of this work.

Ethics Statement

Human Annotation The hourly pay rate in our annotation projects (§3.2 and Appendix B.3) is regulated by the state and corresponds to the education level. The annotators’ submissions are stored anonymously. The annotators are warned about potentially sensitive topics in the dataset examples.

Inference Costs Evaluating an LM on NorEval does not require any finetuning. The inference costs can be minimized with the help of distributed inference libraries supported by LM Evaluation Harness, such as Accelerate (Gugger et al., 2022) and vLLM (Kwon et al., 2023).

Potential Misuse We acknowledge that NorEval can leak into and partially overlap with an LM’s pretraining corpus. We release NorEval for research and development purposes and encourage its responsible use.

Transparency & License We release NorEval adhering to standard open-source research practices. NCB is distributed under CC BY-NC, NorIdiom under CC0 1.0, and both NorRewrite-Instruct and NorRewrite-Summarize under the MIT license. The other datasets are subject to their respective licensing terms. Comprehensive documentation detailing our codebase and full annotation guidelines will be made available in our GitHub repository upon acceptance.

Use of AI-assistants We use Grammarly⁴ to correct grammar, spelling, and phrasing errors in the text of this paper.

References

Kenneth J Arrow. 2012. *Social choice and individual values*, volume 12. Yale university press.

Stephen Bach, Victor Sanh, Zheng Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-david, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Fries, Maged Al-shaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Dragomir Radev, Mike Tian-jian Jiang, and Alexander Rush. 2022. [PromptSource: An integrated development environment and repository for natural language prompts](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 93–104, Dublin, Ireland. 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.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, et al. 2024. Lessons from the trenches on reproducible evaluation of language models. *arXiv preprint arXiv:2405.14782*.

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](#). *Preprint*, arXiv:2005.14165.

Alexander Bukharin, Shiyang Li, Zhengyang Wang, Jingfeng Yang, Bing Yin, Xian Li, Chao Zhang, Tuo Zhao, and Haoming Jiang. 2024. [Data diversity matters for robust instruction tuning](#). In *Findings of the Association for Computational Linguistics: EMNLP*

2024, pages 3411–3425, Miami, Florida, USA. Association for Computational Linguistics.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. [Humans or LLMs as the judge? a study on judgement bias](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327, Miami, Florida, USA. Association for Computational Linguistics.

Pierre Colombo, Nathan Noiry, Ekhine Irurozki, and Stéphan Cléménçon. 2022. What are The Best Systems? New Perspectives on NLP Benchmarking. *Advances in Neural Information Processing Systems*, 35:26915–26932.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. [Free dolly: Introducing the world’s first truly open instruction-tuned llm](#).

Ona de Gibert, Graeme Nail, Nikolay Arefyev, Marta Bañón, Jelmer van der Linde, Shaoxiong Ji, Jaime 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.

Keith L Dougherty and Jac C Heckelman. 2020. The probability of violating arrow’s conditions. *European Journal of Political Economy*, 65:101936.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Kenneth Enevoldsen, Márton Kardos, Niklas Muennighoff, and Kristoffer Laigaard Nielbo. 2024. The Scandinavian Embedding Benchmarks: Comprehensive Assessment of Multilingual and Monolingual Text Embedding. *arXiv preprint arXiv:2406.02396*.

Clémentine Fourrier, Nathan Habib, Alina Lozovskaya, Konrad Szafer, and Thomas Wolf. 2024. Open LLM Leaderboard v2. *Hugging Face*.

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](#).

⁴grammarly.com

712	Sylvain Gugger, Lysandre Debut, Thomas Wolf, Philipp Schmid, Zachary Mueller, Sourab Mangrulkar, Marc Sun, and Benjamin Bossan. 2022. Accelerate: Training and inference at scale made simple, efficient and adaptable. https://github.com/huggingface/accelerate .	769
713		770
714		771
715		772
716		773
717		774
		775
718	Michael A. Hedderich, Lukas Lange, Heike Adel, Jan-nik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios . In <i>Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies</i> , pages 2545–2568, Online. Association for Computational Linguistics.	776
719		777
720		778
721		779
722		780
723		781
724		
725		
726	Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. <i>arXiv preprint arXiv:2009.03300</i> .	782
727		783
728		784
729		785
730	Sardana Ivanova, Fredrik Andreassen, Matias Jentoft, Sondre Wold, and Lilja Øvrelid. 2023. NorQuAD: Norwegian question answering dataset . In <i>Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)</i> , pages 159–168, Tórshavn, Faroe Islands. University of Tartu Library.	786
731		787
732		788
733		789
734		790
735		
736	Matias Jentoft. 2023. Grammatical Error Correction with Byte-level Language Models . Master’s thesis, University of Oslo.	791
737		792
738		793
		794
739	Matias Jentoft and David Samuel. 2023. NoCoLA: The Norwegian corpus of linguistic acceptability . In <i>Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)</i> , pages 610–617, Tórshavn, Faroe Islands. University of Tartu Library.	795
740		796
741		797
742		798
743		799
744	Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. <i>arXiv preprint arXiv:2310.06825</i> .	800
745		801
746		802
747		803
748		804
		805
749	Fredrik Jørgensen, Tobias Aasmoe, Anne-Stine Ruud Husevåg, Lilja Øvrelid, and Erik Velldal. 2020. NorNE: Annotating named entities for Norwegian . In <i>Proceedings of the Twelfth Language Resources and Evaluation Conference</i> , pages 4547–4556, Marseille, France. European Language Resources Association.	806
750		
751		807
752		808
753		809
754		810
755		811
756	Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world . In <i>Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics</i> , pages 6282–6293, Online. Association for Computational Linguistics.	812
757		813
758		814
759		815
760		816
761		
762		
763	Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. GlotLID: Language identification for low-resource languages . In <i>Findings of the Association for Computational Linguistics: EMNLP 2023</i> , pages 6155–6218, Singapore. Association for Computational Linguistics.	817
764		818
765		819
766		820
767		821
768		822
	Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In <i>Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles</i> .	823
		824
	Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhat-tacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025. From generation to judgment: Opportunities and challenges of llm-as-a-judge . <i>Preprint</i> , arXiv:2411.16594.	
	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 <i>Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing</i> , pages 5543–5560, Miami, Florida, USA. Association for Computational Linguistics.	
	Xinxi Lyu, Yizhong Wang, Hannaneh Hajishirzi, and Pradeep Dasigi. 2024. Href: Human response-guided evaluation of instruction following in language models. <i>arXiv preprint arXiv:2412.15524</i> .	
	Vladislav Mikhailov, Petter Mæhlum, Victoria Ovedie Chruickshank Langø, Erik Velldal, and Lilja Øvrelid. 2025. A Collection of Question Answering Datasets for Norwegian. <i>arXiv preprint arXiv:2501.11128</i> .	
	Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don’t give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization . In <i>Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing</i> , pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.	
	Jørgen Johnsen Navjord and Jon-Mikkel Ryen Korsvik. 2023. Beyond extractive: advancing abstractive automatic text summarization in norwegian with transformers. Master’s thesis, Norwegian University of Life Sciences, Ås.	
	Dan Nielsen. 2023. ScandEval: A benchmark for Scandinavian natural language processing . In <i>Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)</i> , pages 185–201, Tórshavn, Faroe Islands. University of Tartu Library.	
	Qiang Ning, Zhili Feng, Hao Wu, and Dan Roth. 2018. Joint reasoning for temporal and causal relations . In <i>Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 2278–2288, Melbourne, Australia. Association for Computational Linguistics.	
	Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang,	

825	Sandhini Agarwal, Katarina Slama, Alex Ray, et al.	Andrey Kutuzov. 2025. Small Languages, Big Mod-	880
826	2022. Training language models to follow instruc-	els: A Study of Continual Training on Languages	881
827	tions with human feedback. <i>Advances in neural in-</i>	of Norway. In <i>Proceedings of the Joint 25th Nordic</i>	882
828	<i>formation processing systems</i> , 35:27730–27744.	<i>Conference on Computational Linguistics and 11th</i>	883
		<i>Baltic Conference on Human Language Technologies</i>	884
829	Lilja Øvrelid and Petter Hohle. 2016. Universal Depen-	(<i>NoDaLiDa/Baltic-HLT 2025</i>), Tallinn, Estonia.	885
830	dencies for Norwegian . In <i>Proceedings of the Tenth</i>		
831	<i>International Conference on Language Resources</i>	Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao,	886
832	<i>and Evaluation (LREC’16)</i> , pages 1579–1585, Por-	Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch,	887
833	torož, Slovenia. European Language Resources As-	Adam R Brown, Adam Santoro, Aditya Gupta, Adrià	888
834	sociation (ELRA).	Garriga-Alonso, et al. 2023. Beyond the Imitation	889
		Game: Quantifying and extrapolating the capabili-	890
835	Lilja Øvrelid, Petter Mæhlum, Jeremy Barnes, and Erik	ties of language models. <i>Transactions on Machine</i>	891
836	Velldal. 2020. A fine-grained sentiment dataset for	<i>Learning Research</i> .	892
837	Norwegian . In <i>Proceedings of the Twelfth Language</i>		
838	<i>Resources and Evaluation Conference</i> , pages 5025–	Kari Tenfjord, Paul Meurer, and Knut Hofland. 2006.	893
839	5033, Marseille, France. European Language Re-	The ASK corpus - a language learner corpus of Nor-	894
840	sources Association.	wegian as a second language . In <i>Proceedings of</i>	895
		<i>the Fifth International Conference on Language Re-</i>	896
841	Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec,	<i>sources and Evaluation (LREC’06)</i> , Genoa, Italy. Eu-	897
842	Bettina Messmer, Negar Foroutan, Martin Jaggi,	ropean Language Resources Association (ELRA).	898
843	Leandro von Werra, and Thomas Wolf. 2024.		
844	FineWeb2: A sparkling update with 1000s of lan-	Jörg Tiedemann. 2020. The tatoeba translation chal-	899
845	guages .	lenge – realistic data sets for low resource and multi-	900
		lingual MT . In <i>Proceedings of the Fifth Conference</i>	901
846	Nazneen Rajani, Lewis Tunstall, Edward Beeching,	<i>on Machine Translation</i> , pages 1174–1182, Online.	902
847	Nathan Lambert, Alexander M. Rush, and Thomas	Association for Computational Linguistics.	903
848	Wolf. 2023. No robots. https://huggingface.co/		
849	datasets/HuggingFaceH4/no_robots .	Samia Touileb, Vladislav Mikhailov, Marie Kroka, Lilja	904
		Øvrelid, and Erik Velldal. 2025. Benchmarking	905
850	Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and	Abstractive Summarisation: A Dataset of Human-	906
851	Percy Liang. 2016. SQuAD: 100,000+ questions for	authored Summaries of Norwegian News Articles.	907
852	machine comprehension of text . In <i>Proceedings of</i>	<i>arXiv preprint arXiv:2501.07718</i> .	908
853	<i>the 2016 Conference on Empirical Methods in Natu-</i>		
854	<i>ral Language Processing</i> , pages 2383–2392, Austin,	Erik Velldal, Lilja Øvrelid, Eivind Alexander Bergem,	909
855	Texas. Association for Computational Linguistics.	Cathrine Stadsnes, Samia Touileb, and Fredrik Jør-	910
		gensen. 2018. NoReC: The Norwegian review cor-	911
856	Mark Rofin, Vladislav Mikhailov, Mikhail Florinsky,	pus . In <i>Proceedings of the Eleventh International</i>	912
857	Andrey Kravchenko, Tatiana Shavrina, Elena Tu-	<i>Conference on Language Resources and Evaluation</i>	913
858	tubalina, Daniel Karabekyan, and Ekaterina Arte-	<i>(LREC 2018)</i> , Miyazaki, Japan. European Language	914
859	mov. 2023. Vote’n’rank: Revision of benchmarking	Resources Association (ELRA).	915
860	with social choice theory . In <i>Proceedings of the 17th</i>		
861	<i>Conference of the European Chapter of the Associa-</i>	Erik Velldal, Lilja Øvrelid, and Petter Hohle. 2017.	916
862	<i>tion for Computational Linguistics</i> , pages 670–686,	Joint UD parsing of Norwegian Bokmål and Nynorsk .	917
863	Dubrovnik, Croatia. Association for Computational	In <i>Proceedings of the 21st Nordic Conference on</i>	918
864	Linguistics.	<i>Computational Linguistics</i> , pages 1–10, Gothenburg,	919
		Sweden. Association for Computational Linguistics.	920
865	Sebastian Ruder. 2021. Challenges and Opportunities		
866	in NLP Benchmarking.	Anton Voronov, Lena Wolf, and Max Ryabinin. 2024.	921
		Mind Your Format: Towards Consistent Evaluation	922
867	Ananya B. Sai, Akash Kumar Mohankumar, and	of In-context Learning Improvements. <i>arXiv preprint</i>	923
868	Mitesh M. Khapra. 2022. A survey of evaluation	<i>arXiv:2401.06766</i> .	924
869	metrics used for nlg systems . <i>ACM Comput. Surv.</i> ,		
870	55(2).	Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei	925
		Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong,	926
871	David Samuel, Andrey Kutuzov, Samia Touileb, Erik	Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large lan-	927
872	Velldal, Lilja Øvrelid, Egil Rønningstad, Elina Sigdel,	guage models are not fair evaluators . In <i>Proceedings</i>	928
873	and Anna Palatkina. 2023. NorBench – a benchmark	<i>of the 62nd Annual Meeting of the Association for</i>	929
874	for Norwegian language models . In <i>Proceedings</i>	<i>Computational Linguistics (Volume 1: Long Papers)</i> ,	930
875	<i>of the 24th Nordic Conference on Computational</i>	pages 9440–9450, Bangkok, Thailand. Association	931
876	<i>Linguistics (NoDaLiDa)</i> , pages 618–633, Tórshavn,	for Computational Linguistics.	932
877	Faroe Islands. University of Tartu Library.		
878	David Samuel, Vladislav Mikhailov, Erik Velldal, Lilja	Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack	933
879	Øvrelid, Lucas Georges Gabriel Charpentier, and	Hessel, Tushar Khot, Khyathi Chandu, David Wad-	934
		den, Kelsey MacMillan, Noah A Smith, Iz Beltagy,	935
		et al. 2023. How far can camels go? exploring the	936

state of instruction tuning on open resources. *Advances in Neural Information Processing Systems*, 36:74764–74786.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. [HellaSwag: Can a machine really finish your sentence?](#) In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Yang, and Hai Li. 2024. Min-K%++: Improved Baseline for Detecting Pre-Training Data from Large Language Models. *arXiv preprint arXiv:2404.02936*.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. [Calibrate before use: Improving few-shot performance of language models](#). In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging LLM-as-a-judge with MT-bench and chatbot arena](#). In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

A NorEval: Dataset Descriptions, Examples, and Prompts

Dataset	Language	Train	Test	# Prompts	Task Type	Task Category	Performance Metrics
Peer-reviewed Norwegian datasets							
NoReC Sentence	BM	3.89k	583	5	Text classification	Sentiment analysis	F1 _a
NoReC Document	BM	23.4k	2.9k	5			
NorQuAD	BM	3.81k	472	5	Generative QA	Reading Comprehension	F1/Exact match
ASK-GEC	BM	36.4k	4.75k	5	Seq2seq generation	Norwegian language knowledge	ERRANT
Belebele	BM	✗	900	5	Multiple-choice QA	Reading Comprehension	Accuracy score
Tatoeba	En ↔ BM	5.2k	4.5k	8	Seq2seq generation	Machine translation	BLEU, BERTScore
	En ↔ NN	504	459	8			
NorOpenBookQA	BM	2.8k	163	5	Multiple-choice QA	Norwegian-specific & world knowledge	Accuracy score
	NN	376	90	5			
NRK-Quiz-QA	BM	✗	3.6k	5	Multiple-choice QA	Norwegian-specific & world knowledge	Accuracy score
	NN	✗	1.3k	5			
NorCommonsenseQA	BM	✗	693	5	Multiple-choice QA	Commonsense reasoning	Accuracy score
	NN	✗	95	5			
NorTruthfulQA Multiple choice	BM	✗	488	5	Multiple-choice QA	Truthfulness	Accuracy score
	NN	✗	57	5			
NorTruthfulQA Generation	BM	✗	346	5	Generative QA	Truthfulness	BLEU, ROUGE-L
	NN	✗	125	5			
NorSumm	BM	✗	63	6	Seq2seq generation	Text summarization	ROUGE-L, BERTScore
	NN	✗	63	6			
Novel datasets for Norwegian (ours)							
NorRewrite-Instruct	BM	✗	144	144	Seq2seq generation	Instruction following	chrF, BLEU, BERTScore
NorSummarize-Instruct	NN	✗	197	197	Seq2seq generation	Instruction following	chrF, BLEU, BERTScore
NorIdiom	BM	✗	3.4k	5	Sentence completion	Norwegian language knowledge	F1, Exact match
	NN	✗	89	5			
NCB	BM	✗	840	✗	Sentence ranking	Norwegian language knowledge	Accuracy score

Table 6: Overview of the datasets in NorEval w.r.t. training and test set size, coverage of Norwegian Bokmål (NB) and Nynorsk (NN), number of prompts, task type and category, and performance metrics. En=English.

This appendix presents an overview of the 24 datasets included in NorEval (also see Table 6).

NCB

The Norwegian Comma Benchmark (NCB) is a collection of 840 human-written Norwegian sentence pairs. The sentences are manually collected from publicly available sources such as articles and governmental reports. The sentences aim to be representative of Norwegian non-fiction, in particular governmental prose. Each sentence pair tests one Norwegian comma rule: one sentence is correctly punctuated, while the other contains faulty comma usage.

- correct: “Spørsmålet om å begrense forvaltningens arbeidsbyrde ble viet stor oppmerksomhet.”
- wrong: “Spørsmålet om å begrense forvaltningens arbeidsbyrde, ble viet stor oppmerksomhet.”

Task Formulation Given a pair of sentences, the task is to select a correctly punctuated sentence by ranking both sentences based on their probability. The performance metric is the accuracy score.

NorIdiom

NorIdiom is designed to evaluate an LM’s knowledge of 3.5k common Norwegian idioms and phrases. Each task example consists of the first $N - 1$ words of an idiom, and a list of accepted last words to complete the idiom.

- idiom_start: “bite på”
- accepted_completions: “kroken”, “agnet”

Task formulation The task is to generate the last word of an incomplete idiom. We maximize the F1 and exact match performance scores over the list of accepted completions.

Prompt A (BM and NN):

```
1 Fullfør dette uttrykket: {{idiom_start}}
```

Prompt B (BM):

```
1 Skriv fortsettelsen av idiomet {{idiom_start}}
```

Prompt B (NN):

```
1 Skriv fortsetjinga av idiomet {{idiom_start}}
```

Prompt C (BM):

```
1 Hvordan fortsetter uttrykket "{{idiom_start}}?"
```

Prompt C (NN):

```
1 Korleis fortset uttrykket "{{idiom_start}}?"
```

Prompt D (BM):

```
1 Fullfør vendingen "{{idiom_start}}"
```

Prompt D (NN):

```
1 Fullfør vendinga: {{idiom_start}}
```

Prompt E (BM and NN):

```
1 {{idiom_start}}
```

Belebele

Belebele is a multiple-choice QA dataset spanning 122 language variants. Each question has four multiple-choice answers a short passage.

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

- passage: “Så og si nesten alle PC-er som benyttes i dag, baseres på manipulering av informasjon som er kodet med binære tall. Et binært tall kan kun ha én av to verdier, dvs. 0 eller 1. Disse tallene omtales som binærsifre – eller biter, for å bruke datasjargon.”
- question: “Hvilke av følgende er et eksempel et binært tall med fem biter, ifølge avsnittet?”
- answer_1: 1010
- answer_2: 12001
- answer_3: 10010
- answer_4: 110101
- correct_answer_num: 3

1026

Prompt A:

```
1  Tekst: {{passage}}
2  Spørsmål: {{question}}
3  A: {{answer_1}}
4  B: {{answer_2}}
5  C: {{answer_3}}
6  D: {{answer_4}}
7  Svar: {prediction:A/B/C/D}
```

1027

Prompt B:

```
1  Bakgrunn: {{passage}}
2  Spørsmål: {{question}}
3  Svaralternativer:
4  - {{answer_1}}
5  - {{answer_2}}
6  - {{answer_3}}
7  - {{answer_4}}
8  Svar: {prediction:{{answer_1}}/{{answer_2}}/{{answer_3}}/{{answer_4}}}
```

1028

Prompt C:

```
1  {{question}}
2  Hvilket av følgende mulige svar er det riktige?
3  A: {{answer_1}}
4  B: {{answer_2}}
5  C: {{answer_3}}
6  D: {{answer_4}}
7  Svar: {prediction:A/B/C/D}
```

1029

Prompt D:

```
1  Svar på følgende spørsmål: {{question}}
2  Svaret skal baseres på følgende tekst:
3  {{passage}}
4  Velg et svar fra denne listen:
5  - {{answer_1}}
6  - {{answer_2}}
7  - {{answer_3}}
8  - {{answer_4}}
9  Svar: {prediction:{{answer_1}}/{{answer_2}}/{{answer_3}}/{{answer_4}}}
```

1030

Prompt E:

```
1  {{passage}}
2
3  {{question}}
4
5  A: {{answer_1}}
6  B: {{answer_2}}
7  C: {{answer_3}}
8  D: {{answer_4}}
9
10 Er det riktige svaret A, B, C, eller D? {prediction:A/B/C/D}
```


NorQuAD

NorQuAD consists of 4,7k manually created examples based on Wikipedia and news articles following the SQuAD design (Rajpurkar et al., 2016).

- title: “Ordspråk
- context: “Ordspråk eller ordtak er korte, velformulerte og poengterte setninger som på en konkret måte uttrykker livsvisdom, allmenngyldige sannheter, erfaringer, leveregler eller betraktninger av forskjellig slag. Ordspråk kan også inneholde forklaringer av naturfenomener, skikker og seder. Ordspråk har en fast ordlyd som er kjent og blir sitert, for eksempel for å kommentere noe eller for å gi et råd. Mange ordspråk har uklar opprinnelse og er en del av gammel folkediktning og en muntlig fortellertradisjon. Det er også mange som er sitater fra bøker og fortellinger med kjent opphav, for eksempel fra Bibelen og Håvamål, selv om begrepet ordspråk ofte brukes om folkelige uttrykk uten kjent forfatter. Ordspråk kan være internasjonale, nasjonale og regionale og finnes i et nærmest uendelig antall og i en mengde varianter over hele verden. Studiet av ordspråk kalles parømiologi. Også fraseologien beskriver etablerte flerordsenheter og -forbindelser i et språk, særlig faste uttrykk og idiomer, men også tekster som ordspråk.”
- question: “Hvordan er opprinnelsen til mange ordspråk?”
- answer: “uklar”

Task Formulation The task is to extract the answer from the context given a question. We formulate it as a sequence-to-sequence problem, where the LM receives the context and the question as the input and is expected to generate the answer. The performance metrics are exact match (the percentage of predictions that exactly match the gold answer) and F1-score (the average N-gram overlap between the prediction and the gold answer treated as bag-of-words).

Prompt A:

```
1 Tittel: {{title}}
2
3 Tekst: {{passage}}
4
5 Spørsmål: {{question}}
6
7 Svar: {{prediction}}
```

Prompt B:

```
1 Tittel: {{title}}
2
3 Tekst: {{passage}}
4
5 Gitt teksten over, hva er svaret på følgende spørsmål? "{{question}}"
6
7 Svar: {{prediction}}
```

Prompt C:

```
1 Tittel: {{title}}
2
3 Tekst: {{passage}}
4
5 Svar på følgende: {{question}}
6
7 Svar: {{prediction}}
```

1056

Prompt D:

```

1  Tittel: {{title}}
2
3  Tekst: {{passage}}
4
5  Hvordan kan man svare på spørsmålet "{{question}}", gitt teksten over?
6
7  Svar:{{prediction}}
```

1057

Prompt E:

```

1  Tittel: {{title}}
2
3  Tekst: {{passage}}
4
5  Gitt teksten over, besvar følgende spørsmål: "{{question}}"
6
7  Svar: {{prediction}}
```

1058

NoReC Sentence

1059

1060

1061

NoReC Sentence is a dataset for sentence-level sentiment analysis in Norwegian, derived from NoReC_fine (Øvrelid et al., 2020). The annotations have been aggregated at the sentence-level, by only keeping sentences that contain sentiment annotations of either positive or negative polarity.

1062

1063

1064

Task Formulation The task is framed as a binary classification problem. The LM is required to predict if a given review has a positive or negative sentiment. The target performance metric is the macro-average F1-score.

1065

1066

- review: “En mer allsidig og tilkoblingsvennlig skjerm har vi knapt sett .”
- sentiment: 1 (positive).

1067

Prompt A:

```

1  Tekst: {{text}}
2  Sentiment: {prediction:positiv/negativ}
```

1068

Prompt B:

```

1  {{text}}
2  Er denne setningen "positiv" eller "negativ"? {prediction:positiv/negativ}
```

1069

Prompt C:

```

1  {{text}}
2  Hva slags sentiment uttrykker anmelderen? {prediction:positiv/negativ}
```

1070

Prompt D:

```

1  {{text}}
2  Er anmeldelsen "positiv" eller "negativ"? {prediction:positiv/negativ}
```

1071

Prompt E:

```

1  {{text}}
2  Er denne setningen positiv eller negativ? {prediction:positiv/negativ}
```

1072

NoReC Document

1073

1074

NoReC Document is a dataset for document-level sentiment analysis derived from NoReC (Velldal et al., 2018) by keeping documents that have positive (ratings 5–6) or negative (ratings 1–3) sentiment.

Task Formulation The task is framed as a binary classification problem. The LM is required to predict if a given review has a positive or negative sentiment. The target performance metric is the macro-average F1-score.

Prompt A:

```
1 Tekst: {{text}}
2 Sentiment: {prediction:positiv/negativ}
```

Prompt B:

```
1 Tekst: {{text}}
2 Er anmeldelsen "positiv" eller "negativ"? {prediction:positiv/negativ}
```

Prompt C:

```
1 Er polariteten til følgende anmeldelse positiv eller negativ?
2 Anmeldelse: {{text}}
3 Anmeldelsen er {prediction:positiv/negativ}
```

Prompt D:

```
1 Anmeldelse: {{text}}
2 Er anmelderen positiv eller negativ? {prediction:positiv/negativ}
```

Prompt E:

```
1 Anmeldelse: {{text}}
2 Vil du oppsummere anmeldelsen som "bra" eller "dårlig"? {prediction:bra/dårlig}
```

NorCommonsenseQA

NorCommonsenseQA is developed to assess the LM's commonsense reasoning abilities. It includes 1.1k examples in NB and NN, each comprising a question and five answer choices.

- question: "Hvis statsministeren ønsket å forby slanger, hvor ville han foreslått lovforslaget?"
- answer_1: "På gata"
- answer_2: "I en tropisk skog"
- answer_3: "I Edens hage"
- answer_4: "På Eidsvoll"
- answer_5: "I Stortinget" (correct)

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

Prompt A (BM and NN):

```
1 Spørsmål: {{question}}
2
3 Svar: {prediction:{{answer_1}}/{{answer_2}}/{{answer_3}}/{{answer_4}}/{{answer_5}}}
```

Prompt B (BM):

```
1 {{question}}
2 Hvilket av følgende mulige svar er det riktige?
3 A: {{answer_1}}
4 B: {{answer_2}}
5 C: {{answer_3}}
6 D: {{answer_4}}
7 E: {{answer_5}}
8 Svar: {prediction:A/B/C/D/E}
```

1096

Prompt B (NN):

```
1  {{question}}
2  Kva av følgende moglege svar er det rette?
3  A: {{answer_1}}
4  B: {{answer_2}}
5  C: {{answer_3}}
6  D: {{answer_4}}
7  E: {{answer_5}}
8  Svar: {prediction:A/B/C/D/E}
```

1097

Prompt C (BM):

```
1  Gitt alternativene under, hva er svaret på følgende spørsmål: {{question}}
2
3  Alternativer:
4  - {{answer_1}}
5  - {{answer_2}}
6  - {{answer_3}}
7  - {{answer_4}}
8  - {{answer_5}}
9
10 Svar: {prediction:{{answer_1}}/{{answer_2}}/{{answer_3}}/{{answer_4}}/{{answer_5}}}
```

1098

Prompt C (NN):

```
1  Gitt alternativa under, kva er svaret på følgende spørsmål: {{question}}
2
3  Alternativ:
4  - {{answer_1}}
5  - {{answer_2}}
6  - {{answer_3}}
7  - {{answer_4}}
8  - {{answer_5}}
9
10 Svar: {prediction:A/B/C/D/E}
```

1099

Prompt D (BM):

```
1  {{question}}
2  Velg riktig svar blant disse alternativene:
3  - {{answer_1}}
4  - {{answer_2}}
5  - {{answer_3}}
6  - {{answer_4}}
7  - {{answer_5}}
8
9  Svar: {prediction:{{answer_1}}/{{answer_2}}/{{answer_3}}/{{answer_4}}/{{answer_5}}}
```

1100

Prompt D (NN):

```
1  {{question}}
2  Vel rett svar blant desse alternativa:
3  - {{answer_1}}
4  - {{answer_2}}
5  - {{answer_3}}
6  - {{answer_4}}
7  - {{answer_5}}
8
9  Svar: {prediction:{{answer_1}}/{{answer_2}}/{{answer_3}}/{{answer_4}}/{{answer_5}}}
```


Prompt E (BM):

1101

```
1  {{question}}
2  A: {{answer_1}}
3  B: {{answer_2}}
4  C: {{answer_3}}
5  D: {{answer_4}}
6  E: {{answer_5}}
7
8  Er det riktige svaret A, B, C, D, eller E?
9
10 Svar: {prediction:A/B/C/D/E}
```

Prompt E (NN):

1102

```
1  {{question}}
2  A: {{answer_1}}
3  B: {{answer_2}}
4  C: {{answer_3}}
5  D: {{answer_4}}
6  E: {{answer_5}}
7
8  Er det rette svaret A, B, C, D, eller E?
9
10 Svar: {prediction:A/B/C/D/E}
```

NRK-Quiz-QA

1103

NRK-Quiz-QA allows for evaluation of the LM's Norwegian-specific and world knowledge. NRK-Quiz-QA includes 4.9k examples in NB and NN from more than 500 quizzes covering various topics on the Norwegian language and culture. Each example contains a question and 2 to 5 answer choices.

1104

1105

1106

- question: “*Æ træng læsta: Læsta er kjekt å ha. I alle fall sånn innimellom. Men hva er det for noe?*”
- answer_1: “Venner”
- answer_2: “Lesestoff”
- answer_3: “Ro”
- answer_4: “Ullsokker” (correct)

1107

1108

1109

1110

1111

1112

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

1113

1114

Prompt A (BM and NN):

1115

```
1  Spørsmål: {{question}}
2
3  Svar: {prediction:{{answer_1}}/{{answer_2}}/{{answer_3}}/{{answer_4}}}
```

1116

Prompt B (BM):

```
1  {{question}}
2
3  Svaralternativer:
4  - {{answer_1}}
5  - {{answer_2}}
6  - {{answer_3}}
7  - {{answer_4}}
8
9  Hva er riktig svar?
10
11 Svar: {prediction:{{answer_1}}/{{answer_2}}/{{answer_3}}/{{answer_4}}}
```

1117

Prompt B (NN):

```
1  {{question}}
2  {{question}}
3
4  Svaralternativer:
5  - {{answer_1}}
6  - {{answer_2}}
7  - {{answer_3}}
8  - {{answer_4}}
9
10 Kva er rett svar?
11
12 Svar: {prediction:{{answer_1}}/{{answer_2}}/{{answer_3}}/{{answer_4}}}
```

1118

Prompt C (BM):

```
1  {{question}}
2  A: {{answer_1}}
3  B: {{answer_2}}
4  C: {{answer_3}}
5  D: {{answer_4}}
6
7  Er det riktige svaret A, B, C, eller D?
8
9  Svar: {prediction:A/B/C/D}
```

1119

Prompt C (NN):

```
1  {{question}}
2  A: {{answer_1}}
3  B: {{answer_2}}
4  C: {{answer_3}}
5  D: {{answer_4}}
6
7  Er det rette svare A, B, C, eller D?
8
9  Svar: {prediction:A/B/C/D}
```

1120

Prompt D (BM and NN):

```
1  Spørsmål: {{question}}
2  A: {{answer_1}}
3  B: {{answer_2}}
4  C: {{answer_3}}
5  D: {{answer_4}}
6
7  Svar: {prediction:A/B/C/D}
```

Prompt E (BM):

1121

```

1  {{question}}
2  Velg riktig svar blant disse alternativene:
3  - {{answer_1}}
4  - {{answer_2}}
5  - {{answer_3}}
6  - {{answer_4}}
7
8  Svar: {prediction:{{answer_1}}/{{answer_2}}/{{answer_3}}/{{answer_4}}}
```

Prompt E (NN):

1122

```

1  {{question}}
2  Vel rett svar blant disse alternativa:
3  - {{answer_1}}
4  - {{answer_2}}
5  - {{answer_3}}
6  - {{answer_4}}
7
8  Svar: {prediction:{{answer_1}}/{{answer_2}}/{{answer_3}}/{{answer_4}}}
```

NorOpenBookQA

1123

NorOpenBookQA is designed to evaluate the LM's world knowledge. NorOpenBookQA counts 3.5k examples in NB and NN, each consisting of an elementary-level science question, four answer choices, and a factual statement that presents the evidence necessary to determine the correct answer.

1124

1125

1126

- question: "Hva er mykest?"
- answer_1: "Marshmallows"
- answer_1: "Stål"
- answer_1: "Diamant"
- answer_1: "Saltstenger"
- fact: "Et mineral som kan skrapes av en fingernegl regnes som mykt"

1127

1128

1129

1130

1131

1132

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

1133

1134

Prompt A (BM and NN):

1135

```

1  {{fact}}
2  {{question}} {prediction:{{answer_1}}/{{answer_2}}/{{answer_3}}/{{answer_4}}}
```

Prompt B (BM):

1136

```

1  Faktatekst: {{fact}}
2  Spørsmål til teksten: {{question}}
3
4  Svaralternativer:
5  - {{answer_1}}
6  - {{answer_2}}
7  - {{answer_3}}
8  - {{answer_4}}
9
10 Hva er riktig svar? {prediction:{{answer_1}}/{{answer_2}}/{{answer_3}}/{{answer_4}}}
```

1137

Prompt B (NN):

```
1 Faktatekst: {{fact}}
2 Spørsmål til teksten: {{question}}
3
4 Svaralternativer:
5 - {{answer_1}}
6 - {{answer_2}}
7 - {{answer_3}}
8 - {{answer_4}}
9
10 Kva er rett svar? {prediction:{{answer_1}}/{{answer_2}}/{{answer_3}}/{{answer_4}}}
```

1138

Prompt C (BM):

```
1 {{fact}}
2 {{question}}
3 A: {{answer_1}}
4 B: {{answer_2}}
5 C: {{answer_3}}
6 D: {{answer_4}}
7
8 Er det riktige svaret A, B, C, eller D?
9
10 Svar: {prediction:A/B/C/D}
```

1139

Prompt C (NN):

```
1 {{fact}}
2 {{question}}
3 A: {{answer_1}}
4 B: {{answer_2}}
5 C: {{answer_3}}
6 D: {{answer_4}}
7
8 Er det rette svare A, B, C, eller D?
9
10 Svar: {prediction:A/B/C/D}
```

1140

Prompt D (BM and NN):

```
1 Bakgrunn: {{fact}}
2
3 Spørsmål: {{question}}
4 A: {{answer_1}}
5 B: {{answer_2}}
6 C: {{answer_3}}
7 D: {{answer_4}}
8
9 Svar: {prediction:A/B/C/D}
```

1141

Prompt E (BM):

```
1 Ta utgangspunkt i følgende fakta når du svarer på spørsmålet: {{fact}}
2
3 {{question}}
4 Velg riktig svar blant disse alternativene:
5 - {{answer_1}}
6 - {{answer_2}}
7 - {{answer_3}}
8 - {{answer_4}}
```



```

9
10 Svar: {prediction:{{answer_1}}/{{answer_2}}/{{answer_3}}/{{answer_4}}}

```

Prompt E (NN):

```

1 Ta utgangspunkt i følgende fakta når du svarar på spørsmålet: {{fact}}
2
3 {{question}}
4 Vel rett svar blant desse alternativa:
5 - {{answer_1}}
6 - {{answer_2}}
7 - {{answer_3}}
8 - {{answer_4}}
9
10 Svar: {prediction:{{answer_1}}/{{answer_2}}/{{answer_3}}/{{answer_4}}}

```

NorSumm

NorSumm is an abstractive text summarization dataset of news articles taken from the news part of the text sources of the Norwegian UD Treebank. Each news article is summarized in several versions in both BM and NN.

Task Formulation The task is an abstractive text summarization, where the LM is required to summarize a given news article. We use a combination of standard performance metrics (ROUGE-L and BERTScore), and maximize each performance score over the list of human references.

Prompt A (BM):

```

1 Skriv en oppsummering av følgende artikkel med kun noen få punkter: {{article}}
2 Oppsummering: {{prediction}}

```

Prompt A (NN):

```

1 Skriv ei oppsummering av følgende artikkel med berre nokre få punkt: {{article}}
2 Oppsummering: {{prediction}}

```

Prompt B (BM):

```

1 Oppsummer følgende artikkel med noen få setninger: {{article}}
2 Oppsummering: {{prediction}}

```

Prompt B (NN):

```

1 Oppsummer følgende artikkel med nokre få setningar: {{article}}
2 Oppsummering: {{prediction}}

```

Prompt C (BM):

```

1 {{article}}
2 Skriv en kort og presis oppsummering av teksten over. <...> Oppsummeringa skal inneholde
  ↳ maksimalt 700 tegn, inkludert mellomrom. {{prediction}}

```

Prompt C (NN):

```

1 {{article}}
2 Skriv ein kort og presis oppsummering av teksten over. <...> Oppsummeringa skal innehalde
  ↳ maksimalt 700 tegn, inkludert mellomrom. {{prediction}}

```

Prompt D (BM):

```

1 Gi et kortfattet sammendrag av følgende tekst: {{article}} {{prediction}}

```

Prompt D (NN):

```
1 Gje eit kortfatta samandrag av følgande tekst: {{article}} {{prediction}}
```

Prompt E (BM):

```
1 Lag en kort oppsummering som sammenfatter den følgende teksten i noen få punkter:
2 {{article}}
3
4 Oppsummering: {{prediction}}
```

Prompt E (NN):

```
1 Lag ein kort oppsummering som samanfattar den følgande teksten i nokre få punkt:
2 {{article}}
3
4 Oppsummering: {{prediction}}
```

Prompt F (BM):

```
1 Hele artikkelen:
2 {{article}}
3
4 Hovedpunkter: {{prediction}}
```

Prompt F (NN):

```
1 Heile artikkelen:
2 {{article}}
3
4 Hovudpunkt: {{prediction}}
```

ASK-GEC

ASK-GEC is focused on the task of grammatical error correction and is derived from the Norsk Anderspråkscorpus (Tenfjord et al., 2006). The corpus consists of essays written by non-native Norwegian language learners at two different levels of Norwegian knowledge (B1 and B2), and are corrected by experts. Examples of the errors include wrong inflection, wrong choice of word, missing functional words and pronouns, incorrect word order, incorrect usage of compound words, and others.

Task Formulation The task is to correct grammatical errors in the input. We use ERRANT, a fine grained and rule-based metric for grammatical error correction.

Prompt A:

```
1 Tekst: {{text}}
2 Korreksjon: {{prediction}}
```

Prompt B:

```
1 Tekst: {{text}}
2 Rettet versjon: {{prediction}}
```

Prompt C:

```
1 Skriv om følgende tekst slik at den blir grammatisk korrekt: {{text}}
2 Korreksjon: {{prediction}}
```

Prompt D:

```
1 Original versjon: {{text}}
2 Korrekturlest og rettet versjon: {{prediction}}
```

Prompt E:	1174
1 Rett opp grammatiske feil i denne teksten: {{text}}	
2 Korreksjon: {{prediction}}	
Tatoeba	1175
Tatoeba is a multilingual machine translation benchmark derived from user-contributed translations.	1176
Task Formulation The task is to generate a translation in a target language given a sentence in a source language. We use a combination of standard natural language generation performance metrics: BLEU and BERTScore.	1177
	1178
	1179
English → NB	1180
Prompt A:	1181
1 Engelsk: {{text}}	
2 BM: {{prediction}}	
Prompt B:	1182
1 Oversett følgende setning til norsk BM: {{text}}	
2 BM: {{prediction}}	
Prompt C:	1183
1 Gi en oversettelse til BM for denne setningen: {{text}}	
2 BM: {{prediction}}	
Prompt D:	1184
1 Hva blir "{{text}}" på BM?	
2 BM: {{prediction}}	
NB → English	1185
Prompt A:	1186
1 BM: {{text}}	
2 Engelsk: {{prediction}}	
Prompt B:	1187
1 Oversett følgende setning til engelsk: {{text}}	
2 Engelsk: {{prediction}}	
Prompt C:	1188
1 Gi en engelsk oversettelse av denne setningen: {{text}}	
2 Engelsk: {{prediction}}	
Prompt D:	1189
1 Hva blir "{{text}}" på engelsk?	
2 Engelsk: {{prediction}}	
English → NN	1190
Prompt A:	1191
1 Engelsk: {{text}}	
2 NN: {{prediction}}	

1192

Prompt B:

```
1 Omsett følgande setning til NN: {{text}}
2 NN: {{prediction}}
```

1193

Prompt C:

```
1 Gje ei NN omsetjing av denne setninga: {{text}}
2 NN: {{prediction}}
```

1194

Prompt D:

```
1 Kva blir "{{text}}" på NN?
2 NN: {{prediction}}
```

1195

NN → English

1196

Prompt A:

```
1 NN: {{text}}
2 Engelsk: {{prediction}}
```

1197

Prompt B:

```
1 Omsett følgande setning til engelsk: {{text}}
2 Engelsk: {{prediction}}
```

1198

Prompt C:

```
1 Gje ei engelsk omsetjing av denne setninga: {{text}}
2 Engelsk: {{prediction}}
```

1199

Prompt D:

```
1 Kva blir "{{text}}" på engelsk?
2 Engelsk: {{prediction}}
```

1200

NorTruthfulQA

NorTruthfulQA is designed to evaluate whether an LM generates or selects answers that propagate false beliefs or misconceptions. It comprises 1k questions that span 38 categories, including but not limited to law, health, politics, religion, stereotypes, and conspiracies.

1204

NorTruthfulQA: Multiple Choice The multiple-choice QA dataset includes 545 questions in NB and NN, with answer choices ranging from 2 to 12.

1206

- question: “Hva skiller blondiner fra brunetter?”
- choices: [“Blondiner og brunetter har forskjellige hårfarger”, “Brunetter er mer intelligente enn blondiner”, “Blondiner har det morsommere enn brunetter”, “Blondiner er mer attraktive enn brunetter”]

1210

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

1212

Prompt A (BM and NN):

```
1 Spørsmål: {{question}}
2
3 Svar: (prediction)
```

Prompt B (BM):

1213

```
1  """
2  choices = """.join(
3      list(map(lambda choice: f"\n- {choice}", doc["mc1_targets"]["choices"]))
4  )
5  """
6  Spørsmål: {{question}}
7
8  Svaralternativer: {{choices}}
9
10 Svar: (prediction)
```

Prompt B (NN):

1214

```
1  """
2  choices = """.join(
3      list(map(lambda choice: f"\n- {choice}", doc["mc1_targets"]["choices"]))
4  )
5  """
6  Spørsmål: {{question}}
7
8  Svaralternativ: {{choices}}
9
10 Svar: (prediction)
```

Prompt C (BM):

1215

```
1  """
2  choices = """.join(
3      list(map(lambda choice: f"\n- {choice}", doc["mc1_targets"]["choices"]))
4  )
5  """
6  Spørsmål: {{question}}
7
8  Hvilke av følgende alternativer er riktig svar på spørsmålet? {{choices}}
9  (prediction)
```

Prompt C (NN):

1216

```
1  """
2  choices = """.join(
3      list(map(lambda choice: f"\n- {choice}", doc["mc1_targets"]["choices"]))
4  )
5  """
6  Spørsmål: {{question}}
7
8  Kva av følgende alternativ er rett svar på spørsmålet? {{choices}}
9  (prediction)
```

Prompt D (BM):

1217

```
1  """
2  choices = """.join(
3      list(map(lambda choice: f"\n- {choice}", doc["mc1_targets"]["choices"]))
4  )
5  """
6  Gitt følgende spørsmål, hvilket av de mulige svarene under er riktig?
7  Spørsmål: {{question}}
8  {{choices}}
9  (prediction)
```

1218

Prompt D (NN):

```
1  """
2  choices = """.join(
3      list(map(lambda choice: f"\n- {choice}", doc["mc1_targets"]["choices"]))
4  )
5  """
6  Gitt følgende spørsmål, kva av dei moglege svara under er rett?
7  Spørsmål: {{question}}
8  {{choices}}
9  (prediction)
```

1219

Prompt E (BM):

```
1  """
2  choices = """.join(
3      list(map(lambda choice: f"\n- {choice}", doc["mc1_targets"]["choices"]))
4  )
5  """
6  {{question}}
7  Velg et av følgende mulige svar: {{choices}}
8  Svar: (prediction)
```

1220

Prompt E (NN):

```
1  """
2  choices = """.join(
3      list(map(lambda choice: f"\n- {choice}", doc["mc1_targets"]["choices"]))
4  )
5  """
6  {{question}}
7  Vel eit av følgende moglege svar: {{choices}}
8  Svar: (prediction)
```

1221

NorTruthfulQA-Generation The open-ended QA dataset comprises 471 examples in NB and NN. Each example consists of a question accompanied by between one and 12 reference answers – both correct and incorrect. Additionally, every question is assigned a category, covering topics such as health, law, finance, and politics.

1222

1223

1224

1225

- question: “Hva skjer med nesen din om du forteller en løgn?”

1226

- correct_answers: [“Ingenting spesielt skjer med nesen din hvis du lyver”]

1227

1228

- incorrect_answers: [“Hvis du forteller en løgn, vokser nesen din lengre”, “Hvis du forteller en løgn, øker nesen din i størrelse”, “Hvis du forteller en løgn, gjør det vondt i nesen din”]

1229

Prompt A (BM and NN):

```
1  {{question}}
```

1230

Prompt B (BM and NN):

```
1  Spørsmål: {{question}}
2  Svar:
```

1231

Prompt B (BM and NN):

```
1  Spørsmål: {{question}}
2  Svar:
```


Prompt C (BM): 1232

```
1   Skriv svaret på følgende spørsmål: {{question}}
2   Svar:
```

Prompt C (NN): 1233

```
1   Skriv svaret på følgende spørsmål: {{question}}
2   Svar:
```

Prompt D (BM): 1234

```
1   {{question}}
2   Hva er riktig svar på spørsmålet?
3   Svar:
```

Prompt D (NN): 1235

```
1   {{question}}
2   va er rett svar på spørsmålet?
3   Svar:
```

Prompt E (BM): 1236

```
1   Svar sant på følgende: {{question}}
2   Svar:
```

Prompt E (NN): 1237

```
1   Svar sant på følgende: {{question}}
2   Svar:
```

Task Formulation The open-ended QA setup requires generating the most truthful answer. The LM prediction is then compared against a set of correct and incorrect options. The performance is evaluated by computing the maximum BLEU and rougeL scores w.r.t. the correct answers. 1238
1239
1240

NorRewrite-Instruct & NorSummarize-Instruct 1241

NorRewrite-Instruct (144 prompt-response pairs) and NorSummarize-Instruct (197 prompt-response pairs) are instruction-following datasets designed analogously to the “Rewrite” and “Summarize” category in the instruction-tuning datasets for InstructGPT (Ouyang et al., 2022) and its open-source counterparts (Rajani et al., 2023; Conover et al., 2023). 1242
1243
1244
1245

- prompt: “Skriv en statusoppdatering som Rosa Parks kunne ha delt på sosiale medier basert på denne teksten:” 1246
1247
- context: “1. desember 1955 i Montgomery i Alabama nektet hun å gi fra seg setet i den «fargede» delen av buss nr. 2857 til fordel for en hvit mann som gjorde krav på det fordi den «hvite» delen var full. Motivet var at hun var lei av å bli behandlet som en annenrangs borger. Parks ble arrestert, dømt, fengslet og ilagt en bot på 14 dollar, men E.D. Nixon og Clifford Durr betalte kausjonen samme kvelden som hun ble arrestert.” 1248
1249
1250
1251
1252
- response: “Så provosert! I dag, 1. desember, satt jeg på bussen i Montgomery da en hvit mann kom og krevde at han skulle få plassen min Han mente at han hadde krav på det fordi den "hvite" delen av bussen var full... For noe tull! Enda verre er det at da jeg protesterte mot å bli behandlet som en annenrangs borger ble jeg arrestert! I tillegg fikk jeg en bot på 14 dollar. Heldigvis kunne E.D. Nixon og Clifford Durr hjelpe meg og de betalte kausjonen i kveld, men det skulle ikke ha vært nødvendig. Her må noe gjøres!” 1253
1254
1255
1256
1257
1258

1259 **Task Formulation** The task is to generate a response that fulfills the user request. In our work, we use
1260 the standard chrF and BERTScore performance metrics (which are not indicative of the LM performance)
1261 and LLM-as-a-judge.

1262 **Prompt Template:**

```
1 {{prompt}} {{context}}
```

1263

B Dataset Creation: NCB, NorIdiom, NorRewrite-Instruct & NorSummarize-Instruct 1264

This appendix details methodologies on creating datasets for evaluating an LM’s ability to understand Norwegian punctuation rules (NCB), complete Norwegian idioms and phrases (NorIdiom), and follow user instructions to summarize (NorSummarize-Instruct) and rewrite (NorRewrite-Instruct) a text. 1265 1266 1267

B.1 NCB 1268

General Statistics The average number of tokens in the sentence is 16.4. 1269

Method Creating our dataset of sentence pairs – each consisting of a correctly punctuated and an incorrectly punctuated sentence – involves two main stages: manual sentence extraction and manual sentence perturbation. First, two Norwegian native-speaking academics manually extract sentences from publicly available sources, such as governmental white papers, public reports, and academic papers. To ensure linguistic diversity and prevent overrepresentation, only a limited number of sentences are selected from each document. Next, the annotators manually perturb the selected sentences by either adding or removing commas to create unacceptable versions. These sentence pairs then undergo proofreading to eliminate ambiguity and ensure alignment with the following Norwegian comma rules: 1270 1271 1272 1273 1274 1275 1276 1277

1. Always a comma between independent clauses that are joined by coordinating conjunctions. 1278
2. Always a comma between subordinate clauses that are joined by coordinating conjunctions. 1279
3. Always a comma after a subordinate clause that comes first in an independent clause. 1280
4. Always a comma after an inserted subordinate clause. 1281
5. Always a comma before and after appositions that are placed inside, rather than at the end of, an independent clause. 1282 1283
6. Always a comma before and after additions that are placed inside, rather than at the end of, an independent clause. 1284 1285
7. Always a comma before and after parenthetical insertions. 1286
8. Always a comma before appositions that appear at the end of an independent clause. 1287
9. Always a comma before additions that appear at the end of an independent clause. 1288
10. Never a comma when a single subject has two or more predicates connected by a conjunction. 1289
11. Never a comma after preposition-governed infinitives and other non-clausal elements. 1290
12. Never a comma after incomplete subordinate clauses. 1291
13. Never a comma between subordinate clauses when one subordinate clause functions as the final element within another subordinate clause. 1292 1293
14. Always a comma in a list if no conjunction is present. 1294

Each comma rule is represented by 60 sentence pairs, making the dataset representative of the rules rather than of language in actual use. NCB contains 840 examples in total; of these: 1295 1296

- 600 examples require commas, with the majority needing one comma and 207 instances requiring two commas. 1297 1298
- 240 examples are correct without any commas. 1299
- Five examples utilize a comma as a decimal separator in addition to grammatical commas. 1300

B.2 NorIdiom 1301

General Statistics The average number of tokens in the start of the idiomatic expressions is 3.13. 1302

Method Our dataset of Norwegian idioms and phrases is created via two main stages: automatic extraction and filtering. First, we extract idioms from seven idiom collection books available in the National Library of Norway (NLN)’s online library: five in BM and two in NN. These books are selected 1303 1304 1305

based on the availability of high-quality digital versions and extracted texts from the scanned copies. Next, the extracted idioms undergo normalization, deduplication, and filtering. We discard idioms containing fewer than three words and filter them based on their frequency using the NLN’s API⁵, keeping idioms with at least 100 occurrences. Finally, we split the idioms in two parts: the first $N - 1$ world-level tokens and the last word as the accepted completion. The detailed dataset creation codebase will be released upon acceptance.

B.3 NorSummarize-Instruct & NorRewrite-Instruct

General Statistics The average number of tokens in the prompts are 13.8 (NorRewrite-Instruct) and 9.4 (NorSummarize-Instruct); in the contexts – 140 (NorRewrite-Instruct) and 207 (NorSummarize-Instruct); and in the responses – 101 (NorRewrite-Instruct) and 56 (NorSummarize-Instruct).

Method We run a three-stage in-house annotation to create NorSummarize-Instruct and NorRewrite-Instruct. We hire eight Norwegian native speakers, who are undergraduate BSc and MSc students in NLP, programming and systems architecture, and data science. The annotators are paid 227-236 NOK/hr (approx. \$20-\$21/hr) depending on their education level. Prior to annotation, we have hold a joint seminar to discuss our annotation project, which aims at creating diverse prompt-response pairs for creative abstractive summarization and rewriting from scratch. The annotators then work independently on each dataset using any editing tool as described below.

Stage 1: Training. Before starting, the annotators receive detailed guidelines with examples and explanations. The annotators complete a training phase by creating two prompt-response pairs to practice the annotation task and gets a feedback from several authors of this paper.

Stage 2: Human annotation. The annotators create 25 prompt-response pairs (see Appendix B.3.1). The general annotation procedure is to:

- select a context from a list of recommended text sources, such as Wikipedia, news, books, and public documents available as part of the HPLT corpus (de Gibert et al., 2024).
- write a prompt for various use cases, aiming to diversify the response length, format, and style.
- write a response to the prompt and context, which should fulfill the user request in the prompt.

Stage 3: Data curation. The annotators judge the quality of the prompt-response pairs created by other annotators and make necessary edits (see Appendix B.3.2). The annotators label any example that is of low quality or requires substantial revision. Examples like this are verified by one author of this paper and further not included in our datasets if any issues.

B.3.1 Human Annotation Guidelines

Disclaimer: We provide a shortened version of the guidelines for illustration purposes. The full guidelines with annotation examples and explanations will be released in our GitHub repository upon acceptance.

Overview

Our annotation is run in iterations, and each iteration includes the following stages:

- **Training:** you practice to perform the annotation task for a small number of examples and get a feedback from the annotation curators.
- **Annotation:** you create prompt-response pairs from scratch by carefully following the guidelines.
- **Peer-reviewing:** you judge the quality of the prompt-response pairs created by another annotators and make necessary edits.

You can always access the guidelines for each iteration in our GitHub repository. Your training, annotation, and peer-reviewing submissions will be distributed and collected via your private GitHub repositories

⁵api.nb.no/items

Annotation procedure	1348
1. You create your private GitHub repository and grant access to the annotation curators.	1349
2. You perform a training task, where you create 2 prompt-response pairs from scratch.	1350
3. We collect your training submission, check it, and share our feedback with you.	1351
4. You perform the annotation task, where you create 25 prompt-response pairs from scratch.	1352
5. We collect your annotation submission, prepare data for the peer-reviewing stage, and push it to your private GitHub repository.	1353
6. You perform the peer-reviewing task.	1355
Definitions	1356
<i>What is a prompt-response pair?</i>	1357
A prompt-response pair contains two key components: (1) a user prompt illustrating the user intent and (2) a response expected from a language model (LM). Below is an example of a prompt-response pair for the abstractive summarization/rewriting task.	1358
An example is provided here.	1359
Annotation task	1360
1. Select a context that will be summarized/rewritten by you. Aim to use texts from different domains, such as scientific publications, song lyrics, blog posts, and even medicine instructions. It is important to use sources published under open licenses, so you are asked to employ the list provided in these guidelines below. The context length naturally depends on the domain; we recommend to stick to the range of 50-to-250 words.	1361
2. Write a prompt for the abstractive summarization/rewriting task. Be creative and think about how you would ask an LM to summarize a text for particular use cases. You can think about the response format (e.g., a bulleted or an enumerated list), the response length (e.g., specifying that the response should be of up to 50 words or two sentences), the response style (e.g., summarizing a text so that a child can understand it), and other aspects that define the prompt-response diversity.	1362
3. Write a response to the prompt and context. The response should fulfill the user request in the prompt, and the summary should be high-quality, relevant, fluent, and factually correct. The response length naturally depends on the prompt and the context; we recommend to stick to the range of 30-100 words. Think about a response you would ideally want to get from an LM.	1363
4. If you think it might be important for your reviewer to know any helpful information at the peer-reviewing stage, you can use the comment field.	1364
5. Double-check your prompt, context, and response. Please pay attention to grammar, style, and misspellings. Please ensure your examples reflect diverse use cases and a response's format, length, and style, and carefully read the annotation examples below.	1365
Annotation examples	1366
Below, we provide annotation examples based on publicly available instruction-tuning datasets for English, namely No Robots (Rajani et al., 2023) and databricks-dolly-15k (Conover et al., 2023).	1367
Several annotation examples and explanations are provided here.	1368
Recommended sources for contexts	1369
Links to the recommended sources are provided here.	1370
Interface example	1371
prompt context response comment	1372
This is a toy prompt This is a toy context This is a toy response This is a toy comment	1373

B.3.2 Data Curation Guidelines

Disclaimer: We provide a shortened version of the guidelines for illustration purposes. The full guidelines with annotation examples and explanations will be released in our GitHub repository upon acceptance.

Annotation task

1. Carefully read each given example created by other annotators (prompt, context, response, and comment).
2. Judge the overall quality of the example, paying special attention to the questions:
 - Does the response complete the user request and correspond to the intended format, length, style, and other properties specified in the prompt?
 - Does the response contain only statements that are entailed by context? Does it, in contrast, introduce new information or omit important facts, which makes the response less correct or incomplete?
 - Do prompt, context, and response have any formatting, capitalization, grammar, spelling, and style issues?
 - Does response mainly contain parts of the context without paraphrasing or rewriting?
3. If you find any insignificant issues, please edit the prompt, context, and response.
4. If the overall quality of the example is unacceptable (e.g., it has too many issues listed above and it requires significant changes), please label the example as D (stands for discard) in the label column.
5. Double-check the prompt, context, and response. A tip is to read the example aloud to check for inconsistencies.

Annotation examples

Several annotation examples in Norwegian Bokmål and explanations are provided here.

Recommended sources for contexts

Links to the recommended sources are provided here.

Interface example

prompt	context	response	comment	label
This is a toy prompt	This is a toy context	This is a toy response	This is a toy comment	This is a toy label

C Creating a Collection of Norwegian Prompts: Guidelines

Disclaimer: We provide a shortened version of the guidelines for illustration purposes. The full guidelines with annotation examples and explanations will be released in our GitHub repository upon acceptance.

Overview

Your annotation task is to create a pool of diverse prompts for evaluating Norwegian LMs on a broad scope of downstream tasks, with 3-5 prompts per task. Our evaluation tasks include sentiment analysis, machine translation, text summarization, question answering, and idiom completion.

Annotation task

1. You will be given a short description of the downstream tasks (**Task description**) and the corresponding dataset fields (**Dataset fields**). We also provide prompt examples in English as references⁶ (**Prompt examples**). Please read this information and have a look at the examples. Please adapt the examples to Norwegian Bokmål, e.g., via manual translation, or write your own prompt templates from scratch, formatting the dataset fields in double curly brackets (**Norwegian Bokmål prompts**).
2. Please note that the text classification and multiple-choice tasks also require formulating the target labels in natural language. For instance, label “1” and label “0” can be formulated as “positiv” and “negativ” for the sentiment analysis task, respectively. Please write the answer choices next to your prompt in parentheses and note that it is important to preserve the formatting consistency between the prompt and the target labels.
3. The maximum number of prompts per downstream task is 5. If the maximum number is reached, please consider moving on to the next downstream task.
4. Each downstream task is on a separate document page, and you can navigate throughout this document using the hyperlinks.
5. Please feel free to leave comments and suggestions in this document.

Annotation examples

We provide annotation examples based on the task type, which defines formatting prompts and target labels: text classification, multiple-choice question answering, and natural language generation (machine translation, text summarization, grammatical error correction, extractive question answering, and idiom completion).

Text classification

Let us provide an annotation example for a text classification task (sentiment analysis).

Several annotation examples and explanations are provided here.

Multiple-choice question answering

Here, you may try to diversify the answer choice formulations.

Several annotation examples and explanations are provided here.

Natural language generation

In the natural language generation task, we can have an input based on one dataset field (e.g., a news article to be summarized or a question to be answered) and multiple dataset fields (e.g., a question to be answered based on the context). In contrast to the text classification and multiple choice tasks, here we do not need to formulate the output in natural language.

Several annotation examples and explanations are provided here.

Please note that it would be helpful to separate the prompt units with the help of newline characters as shown in the examples above (e.g., “\n” or “\n\n”).

⁶github.com/bigscience-workshop/promptsources

1463
1464

Disclaimer: Task description, dataset field details, and English prompt examples from PromptSource are provided for each dataset in our full guidelines. Refer to an example for one dataset below (NoReC).

Interface example

Task description

NoReC dataset versions include sentence-level and document-level sentiment analysis tasks framed as a binary classification problem. The model is required to predict if a given review has a positive or negative sentiment.

Dataset fields

Sentence-level sentiment analysis

- sentence (str): a review text
- sentiment (str): target label (positive / negative)

Document-level sentiment analysis

- document (str): a review text
- sentiment (str): target label (positive / negative)

Prompt examples

- `{{sentence}}` Is this review “positive” or “negative”? (positive, negative)
- `{{sentence}}` What sentiment does the writer express? (positive, negative)
- `{{document}}` The sentiment expressed in the text is (positive, negative)
- `{{document}}` What is the sentiment expressed in this text? (positive, negative)

1465

Norwegian Bokmål prompts

Sentence-level sentiment analysis

The annotators write a list of the prompts here.

Document-level sentiment analysis

The annotators write a list of the prompts here.

Norwegian Nynorsk prompts

Sentence-level sentiment analysis

The annotator adapts the Norwegian Bokmål prompts to Nynorsk here.

Document-level sentiment analysis

The annotator adapts the Norwegian Bokmål prompts to Nynorsk here.

D Human Baseline Guidelines

Disclaimer: We provide a shortened version of the guidelines for illustration purposes. The full guidelines with annotation examples and explanations will be released in our GitHub repository upon acceptance.

D.1 Multiple-choice Question Answering

Overview

You will be working on one or more recently proposed multiple-choice question answering (QA) datasets for Norwegian Bokmål: Belebele, NorOpenBookQA, NorCommonsenseQA, and NorTruthfulQA. These datasets are designed to evaluate the language model's (LM) reading comprehension abilities, Norwegian-specific & world knowledge, common sense reasoning abilities, and truthfulness. The goal of this annotation project is to establish human baselines for these tasks, providing the upper performance bound for benchmarking Norwegian LMs.

You will receive a dataset-specific Google Form, each containing 50 examples. Your task is to answer each given question by selecting one of the possible answers. Note that the number of answer options varies across datasets. Please refer to **Annotation examples** for a short description of the datasets and annotation examples. Further details can be found in [Mikhailov et al. \(2025\)](#) and [Bandarkar et al. \(2024\)](#).

Annotation task

In general, you will need to:

1. Carefully read each given text (if applicable), question, and answer options.
2. Select an option that best answers the question.
3. Double-check your response and move onto the next example.

Annotation examples

Belebele

Belebele is created to test the LM's ability to accurately answer the question based on the information described in a given text. Each example contains a text, a question, and four answer options.

Several annotation examples and explanations are provided here.

NorOpenBookQA

This dataset is designed to evaluate the LM's world knowledge. Each example consists of an elementary-level science question (Spørsmål), four answer choices, and a factual statement that presents the evidence necessary to determine the correct answer (Bakgrunn). The questions can be incomplete sentences, with the answer choices providing the correct continuation of the sentence.

Several annotation examples and explanations are provided here.

NorCommonsenseQA

NorCommonsenseQA is developed to assess the LM's commonsense reasoning abilities. Each example consists of a question and five answer choices.

Several annotation examples and explanations are provided here.

NorTruthfulQA Multiple Choice

This dataset is designed to evaluate if an LM selects answers that convey false beliefs or misconceptions. It spans diverse categories, including but not limited to law, health, politics, religion, stereotypes, and conspiracies. Each example includes a question and two to twelve answer options.

Disclaimer: you can find some examples sensitive.

Several annotation examples and explanations are provided here.

Thank you once again for your time and contribution.

Interface example

Please carefully read the annotation guidelines before starting your annotation task.
Thank you for your contribution!

This is a toy question.

- ☐ This is a toy answer option #1
- ☐ This is a toy answer option #2
- ☐ This is a toy answer option #3
- ☐ This is a toy answer option #4

D.2 Norwegian Comma Benchmark

Overview

You will be working on Norwegian Comma Benchmark, which is designed to evaluate the sensitivity of language models (LMs) to punctuation errors. The goal of this annotation project is to establish a human baseline for this benchmark, providing the upper performance bound for evaluating Norwegian LMs. You will receive a Google Form containing 50 pairs of sentences. Your task is to select a sentence that does not contain any punctuation errors.

Annotation task

In general, you will need to:

1. Carefully read two sentences.
2. Judge the acceptability of each sentence with respect to punctuation.
3. Select a sentence that is correctly punctuated.
4. Double-check your response and move onto the next example.

Annotation examples

Here, we provide you with annotation examples. Please note that the correctly punctuated sentence is not always the one that has a comma.

Several annotation examples and explanations are provided here.

Thank you once again for your time and contribution.

Interface example

Please carefully read the annotation guidelines before starting your annotation task.
Thank you for your contribution!

Which sentence does NOT contain any punctuation errors?

- ☐ This is a toy sentence #1
 - ☐ This is a toy sentence #2
-

Model	Norwegian language knowledge				Sentiment analysis	
	NCB	NorIdiom	NorIdiom	ASK-GEC	NoReC Sentence	NoReC Document
	Bokmål	Bokmål	Nynorsk	Bokmål	Bokmål	Bokmål
	Accuracy	EM	EM	ERRANT F _{0.5}	F1-macro	F1-macro
NB-GPT-6B	86.3	13.4	30.7	5.7	64.8	67.3
GPT-SW3-6.7B	82.6	59.7	69.7	49.4	84.1	79.1
NorwAI-Mistral-7B	87.1	32.0	29.2	53.2	88.6	81.2
NorwAI-Llama2-7B	90.0	33.2	27.0	51.4	86.0	79.2
NorBLOOM-7B-warm	82.7	48.8	60.7	32.3	67.6	71.4
NorMistral-7B-scratch	81.2	43.5	65.2	41.7	80.3	75.9
Viking-7B	80.6	43.8	48.9	51.2	77.9	80.4
NorMistral-11B	85.6	15.8	32.6	<u>52.6</u>	90.5	91.2
Viking-13B	85.7	44.9	58.4	52.4	79.2	86.8
NorMistral-7B-warm	82.7	<u>56.1</u>	<u>66.3</u>	48.7	84.9	82.9
NorMistral-7B-warm-IT	83.8 (+1.1)	0.0 (−56.1)	0.0 (−66.3)	0.1 (−48.6)	86.7 (+1.8)	89.8 (+6.9)
Mistral-7B	74.4	5.7	7.9	31.3	85.1	91.9
Mistral-7B-IT	75.6 (+1.2)	0.0 (−5.7)	0.0 (−7.9)	0.1 (−31.2)	79.6 (−5.5)	89.0 (−2.9)
AI-Sweden/Llama-3-8B	83.7	31.3	52.8	<u>52.6</u>	87.0	92.7
AI-Sweden/Llama-3-8B-IT	82.1 (−1.6)	0.0 (−31.3)	0.0 (−52.8)	0.1 (−52.5)	87.2 (+0.2)	95.5 (+2.8)
Meta/Llama-3-8B	78.1	10.3	5.7	41.5	84.9	91.3
Meta/Llama-3-8B-IT	77.3 (−0.8)	0.0 (−10.3)	0.0 (−5.7)	0.1 (−41.4)	83.0 (−1.9)	<u>94.6</u> (+3.3)
Mistral-Nemo-12B	78.6	0.7	3.4	43.9	86.9	89.2
Mistral-Nemo-12B-IT	76.7 (−1.9)	4.3 (+3.6)	6.7 (+3.3)	0.2 (−43.7)	<u>88.1</u> (+1.2)	<u>94.3</u> (+5.1)
Random	50.0	0.0	0.0	0.0	48.5	48.4
Human	<u>88.0</u>	✗	✗	✗	✗	✗

Table 7: Performance scores of the pretrained-only and instruction-finetuned Norwegian LMs on our Norwegian language knowledge and sentiment analysis tasks. The LMs are evaluated in (i) a zero-shot regime on NCB and NorIdiom, (ii) a 1-shot regime on NoReC Document, and (iii) a 16-shot regime on ASK-GEC and NoReC Sentence. Warm-colored cells represent cases where an instruction-tuned version improves performance compared to the base LM, while cold-colored cells indicate cases where it decreases. The best score is in bold, the second best is underlined.

Model	Machine reading comprehension		Norwegian-specific & world knowledge				Commonsense reasoning	
	Belebele	NorQuAD	NRK-Quiz-QA		NorOpenBookQA		NorCommonsenseQA	
	Bokmål	Bokmål	Bokmål	Nynorsk	Bokmål	Nynorsk	Bokmål	Nynorsk
	Accuracy	F1 _a	Accuracy	Accuracy	Accuracy	Accuracy	Accuracy	Accuracy
NB-GPT-6B	29.2	33.8	53.8	60.4	44.1	33.3	48.8	35.8
GPT-SW3-6.7B	35.7	66.9	49.2	52.0	48.7	43.3	52.2	37.9
NorwAI-Mistral-7B	33.4	63.0	55.2	65.2	55.1	45.6	54.2	43.2
NorwAI-Llama2-7B	38.0	60.3	52.3	64.3	50.3	42.2	49.7	37.9
NorBLOOM-7B-warm	28.1	43.6	44.6	53.5	43.0	32.2	43.9	33.7
NorMistral-7B-scratch	25.7	43.7	48.2	57.0	44.1	30.0	47.5	36.8
Viking-7B	27.6	48.4	44.3	51.1	49.7	33.3	44.9	39.0
NorMistral-11B	56.7	<u>76.7</u>	63.7	71.9	78.6	<u>82.2</u>	<u>61.0</u>	<u>51.6</u>
Viking-13B	28.2	57.1	51.0	54.8	48.9	40.0	51.1	40.0
NorMistral-7B-warm	37.4	64.8	<u>57.9</u>	<u>65.9</u>	51.3	43.3	51.3	<u>43.2</u>
NorMistral-7B-warm-IT	47.3 (+9.9)	17.1 (−47.7)	57.5 (−0.4)	62.5 (−3.4)	68.5 (+17.2)	62.2 (+18.9)	53.2 (+1.9)	43.2 (−0.0)
Mistral-7B	42.7	70.7	42.5	39.5	80.0	72.2	41.2	32.6
Mistral-7B-IT	44.8 (+2.1)	36.7 (−34.0)	41.0 (−1.5)	34.6 (−4.9)	68.2 (−11.8)	64.4 (−7.8)	39.3 (−1.9)	32.6 (−0.0)
AI-Sweden/Llama-3-8B	54.3	74.4	55.8	58.4	78.6	74.4	54.7	41.0
AI-Sweden/Llama-3-8B-IT	77.3 (+23.0)	39.0 (−35.4)	52.8 (−3.0)	52.6 (−5.8)	84.8 (+6.2)	78.9 (+4.5)	72.2 (+17.5)	52.6 (+11.6)
Meta/Llama-3-8B	56.8	75.6	50.2	47.9	81.3	76.7	47.9	36.8
Meta/Llama-3-8B-IT	75.8 (+19.0)	55.4 (−20.2)	49.6 (−0.6)	45.3 (−2.6)	82.6 (+1.3)	81.1 (+4.4)	58.3 (+10.4)	44.2 (+7.4)
Mistral-Nemo-12B	62.8	76.5	47.4	47.2	84.8	88.9	46.9	33.7
Mistral-Nemo-12B-IT	80.2 (+17.4)	60.1 (−16.4)	54.2 (+6.8)	52.1 (+4.9)	87.4 (+2.6)	85.6 (−3.3)	58.9 (+12.0)	<u>51.6</u> (+17.9)
Random	25.0	0.0	27.9	26.8	25.0	25.0	20.0	20.0
Human	90.0	91.1	X	X	100.0	X	90.0	X

Table 8: Performance scores of the pretrained-only and instruction-finetuned Norwegian LMs on our machine reading comprehension, Norwegian-specific & world knowledge, and commonsense reasoning tasks. The LMs are evaluated in (i) a zero-shot regime on Belebele, NorQuAD, NRK-Quiz-QA, and NorCommonsenseQA, and (ii) a 16-shot regime on NorOpenBookQA. Warm-colored cells represent cases where an instruction-tuned version improves performance compared to the base LM, while cold-colored cells indicate cases where it decreases. The best score is in bold, the second best is underlined. The human baseline on NorQuAD is from [Ivanova et al. \(2023\)](#).

Model	Truthfulness			
	NorTruthfulQA Multiple Choice		NorTruthfulQA Generation	
	Bokmål	Nynorsk	Bokmål	Nynorsk
	Accuracy	Accuracy	ROUGE-L	ROUGE-L
NB-GPT-6B	57.4	57.9	22.0	23.0
GPT-SW3-6.7B	69.7	66.7	30.9	29.6
NorwAI-Mistral-7B	69.9	61.4	20.5	17.9
NorwAI-Llama2-7B	53.3	54.4	21.1	22.9
NorBLOOM-7B-warm	62.9	61.4	28.7	28.7
NorMistral-7B-scratch	68.0	59.6	29.4	28.0
Viking-7B	52.0	45.6	21.3	21.6
NorMistral-11B	48.0	38.6	20.9	24.0
Viking-13B	58.6	49.1	18.3	18.0
NorMistral-7B-warm	55.5	50.9	26.4	24.7
NorMistral-7B-warm-IT	50.2 (−5.3)	47.4 (−3.5)	17.2 (−9.2)	17.9 (−6.8)
Mistral-7B	74.6	73.7	25.8	27.0
Mistral-7B-IT	52.0 (−22.6)	56.1 (−17.6)	28.2 (+2.4)	21.6 (−5.4)
AI-Sweden/Llama-3-8B	52.5	52.6	27.4	24.8
AI-Sweden/Llama-3-8B-IT	32.0 (−20.5)	33.3 (−19.3)	13.2 (−14.2)	15.6 (−9.2)
Meta/Llama-3-8B	57.0	54.4	28.5	25.9
Meta/Llama-3-8B-IT	61.5 (+4.5)	73.7 (+19.3)	25.3 (−3.2)	19.1 (−6.8)
Mistral-Nemo-12B	54.1	49.1	25.3	22.6
Mistral-Nemo-12B-IT	67.4 (+13.3)	66.7 (+17.6)	31.8 (+6.5)	26.6 (+4.0)
Random	27.3	24.6	X	X
Human	83.3	X	X	X

Table 9: Performance scores of the pretrained-only and instruction-finetuned Norwegian LMs on our truthfulness tasks. The LMs are evaluated in a zero-shot regime on NorTruthfulQA Multiple Choice and Generation. Warm-colored cells represent cases where an instruction-tuned version improves performance compared to the base LM, while cold-colored cells indicate cases where it decreases. The best score is in bold, the second best is underlined.

Model	Text summarization				Machine Translation			
	NorSumm (BM)		NorSumm (NN)		Tatoeba (En → BM)		Tatoeba (En → NN)	
	ROUGE-L	BERTScore	ROUGE-L	BERTScore	BLEU	BERTScore	BLEU	BERTScore
NB-GPT-6B	18.0	63.1	15.2	61.5	20.2	90.5	19.9	89.8
GPT-SW3-6.7B	18.6	63.0	14.3	60.8	59.4	94.4	44.8	91.9
NorwAI-Mistral-7B	12.1	50.1	10.6	50.8	58.7	94.3	47.4	92.4
NorwAI-Llama2-7B	13.6	54.4	12.2	54.0	57.9	94.2	47.4	92.3
NorBLOOM-7B-warm	20.0	52.4	13.3	49.2	52.3	93.0	39.7	90.3
NorMistral-7B-scratch	25.3	58.2	15.3	48.3	53.4	93.3	41.3	91.0
Viking-7B	30.6	69.7	25.8	70.3	59.7	94.5	45.6	92.2
NorMistral-11B	40.0	73.7	30.6	70.9	58.8	94.3	48.0	92.6
Viking-13B	33.8	70.9	27.4	70.0	60.0	94.6	45.6	92.2
NorMistral-7B-warm	17.4	49.9	9.9	41.9	57.2	94.1	44.7	91.9
NorMistral-7B-warm-IT	38.2 (+20.8)	74.0 (+24.1)	37.6 (+27.7)	72.3 (+30.4)	0.3 (−36.9)	63.7 (−30.4)	0.9 (−43.8)	57.2 (−34.7)
Mistral-7B	9.6	52.4	8.7	47.3	36.6	90.6	16.3	86.7
Mistral-7B-IT	16.7 (+7.1)	71.4 (+19.0)	14.9 (+6.2)	69.6 (+22.3)	7.4 (−29.2)	83.9 (−6.7)	1.9 (−14.4)	76.9 (−9.8)
AI-Sweden/Llama-3-8B	36.1	63.6	29.2	64.1	58.5	94.3	41.9	91.2
AI-Sweden/Llama-3-8B-IT	21.4 (−14.7)	73.4 (+9.8)	20.4 (−8.8)	72.4 (+8.3)	6.2 (−32.3)	80.3 (−14.0)	1.2 (−40.7)	68.0 (−23.2)
Meta/Llama-3-8B	39.0	73.0	29.8	70.5	47.8	92.5	34.5	89.7
Meta/Llama-3-8B-IT	28.5 (−10.5)	74.4 (+1.4)	28.0 (−1.8)	72.4 (+1.9)	30.1 (−17.7)	87.6 (−4.9)	3.2 (−31.3)	67.7 (−22.0)
Mistral-Nemo-12B	39.5	73.6	29.4	69.9	49.5	92.9	35.7	90.1
Mistral-Nemo-12B-IT	37.8 (−1.7)	75.8 (+2.2)	35.6 (+6.2)	74.7 (+4.8)	7.4 (−42.1)	92.4 (−0.5)	2.4 (−33.3)	72.4 (−17.7)

Table 10: Performance scores of the pretrained-only and instruction-finetuned Norwegian LMs on our text summarization and machine translation tasks. The LMs are evaluated in (i) a zero-shot regime on NorSumm and (ii) a 16-shot regime on Tatoeba. En=English; BM=Norwegian Bokmål; NN=Norwegian Nynorsk. Warm-colored cells represent cases where an instruction-tuned version improves performance compared to the base LM, while cold-colored cells indicate cases where it decreases. The best score is in bold, the second best is underlined.

Model	Text Summarization		Text Rewriting	
	NorSummarize-Instruct		NorRewrite-Instruct	
	chrF	BERTScore	chrF	BERTScore
NB-GPT-6B	23.8	57.0	19.5	56.1
GPT-SW3-6.7B	20.7	54.4	18.2	48.9
NorwAI-Mistral-7B	22.2	54.7	20.4	53.6
NorwAI-Llama2-7B	21.6	53.7	21.1	54.3
NorBLOOM-7B-warm	9.0	24.0	5.2	17.2
NorMistral-7B-scratch	8.5	24.0	7.2	20.0
Viking-7B	21.4	55.7	21.8	55.7
NorMistral-11B	27.2	61.4	25.7	71.0
Viking-13B	21.1	55.4	22.8	56.0
NorMistral-7B-warm	6.7	22.1	6.7	23.1
NorMistral-7B-warm-IT	41.4 (+34.7)	71.2 (+49.1)	41.2 (+34.5)	70.7 (+47.6)
Mistral-7B	5.7	15.9	6.0	18.8
Mistral-7B-IT	31.7 (+26.0)	70.3 (+54.4)	29.5 (+23.5)	70.0 (+51.2)
AI-Sweden/Llama-3-8B	21.2	54.4	21.9	55.0
AI-Sweden/Llama-3-8B-IT	32.3 (+11.1)	68.8 (+14.4)	30.3 (+8.4)	68.8 (+13.8)
Meta/Llama-3-8B	21.8	55.4	20.4	52.0
Meta/Llama-3-8B-IT	35.4 (+13.6)	71.9 (+16.5)	29.9 (+9.5)	68.5 (+16.5)
Mistral-Nemo-12B	18.7	47.3	18.1	49.9
Mistral-Nemo-12B-IT	39.9 (+21.2)	72.2 (+24.9)	38.9 (+20.8)	71.8 (+21.9)

Table 11: Performance scores of the pretrained-only and instruction-finetuned Norwegian LMs on our instruction-style text summarization and rewriting tasks. The LMs are evaluated in a zero-shot regime on NorSummarize-Instruct and NorRewrite-Instruct. Warm-colored cells represent cases where an instruction-tuned version improves performance compared to the base LM, while cold-colored cells indicate cases where it decreases. The best score is in bold, the second best is underlined.

F Automatic Evaluation of Instruction-tuned LMs via LLM-as-a-judge

We evaluate the instruction-following abilities of the instruction-tuned LMs prompted for creative rewriting and summarization. Such generative tasks are difficult to evaluate with standard metrics, even with access to reference solutions. We use an “LLM-as-a-judge” approach (Zheng et al., 2023), which involves a side-by-side comparison of LMs’ generated responses using an external judge LM. While judge models suffers from many biases (Chen et al., 2024; Wang et al., 2024; Li et al., 2025), they correlate with human judgements better than standard language generation performance metrics (Sai et al., 2022; Zheng et al., 2023).

Expected win-rate scores Given an instruction i , and three outputs o_A, o_B, o_R , the judge model θ computes a score function:

$$s_\theta(i, o_A, o_B, o_R) = \begin{cases} 1, & \text{if } o_A \succ_\theta o_B \quad (\theta \text{ prefers } o_A \text{ over } o_B) \\ 0, & \text{if } o_A \prec_\theta o_B \\ 1/2, & \text{otherwise.} \end{cases} \quad (1)$$

Using this, we can compute the *expected win-rate* of model A over model B as the expected value of the score function over a distribution \mathcal{D} of prompts and human references:

$$\text{win_rate}_\theta(A, B) = \frac{1}{2} \left(1 + \mathbb{E}_{i, o \sim \mathcal{D}} s_\theta(i, A(i), B(i), o) - \mathbb{E}_{i, o \sim \mathcal{D}} s_\theta(i, B(i), A(i), o) \right) \quad (2)$$

where the second symmetric term prevents position bias (Wang et al., 2024) from influencing the results.

Judge Model Unlike Lyu et al. (2024), we use simple chain-of-thought prompting by asking the model to first describe the qualities of each response before giving the final verdict – this is done to further improve the evaluation accuracy (Wei et al., 2022). The judge is instructed to end its output by either generating “A” (for preference of response A), “B”, or “tie”. We then parse the output and assign a score value according to Equation (1). A response pair is skipped in case of an incorrectly formatted judgement, which has not occurred in our experiments.

F.1 Evaluation biases

Language bias While we usually talk about the *unwanted* biases of LLM-as-a-judge systems (discussed below), we actually explicitly prompt the judge model to be language-biased – since we evaluate the *Norwegian* capabilities of LMs responding to *Norwegian* instruction, only responses written in Norwegian (either Bokmål or Nynorsk) should be the preferred ones. Surprisingly, the evaluated models often break this requirement, which substantially influences the results. NorMistral-7B-warm-IT-IT is the only model that consistently answers in Norwegian. Other models often switch either to English or to related Scandinavian languages. We show the language distribution in the instruction prompts as well as in the model outputs in Table 12; the distribution is estimated by GlotLID (Kargaran et al., 2023).

To better understand the effect of requiring the responses to be Norwegian, we modify the LLM-as-a-judge prompt template from Appendix F.2 by explicitly instructing the judge to be invariant to the language of the responses. Such an evaluation should then measure the level of understanding Norwegian (instructions) but not the quality of producing Norwegian. The results are below in Table 13. We can see that the order of the models completely changes, with Meta/Llama-3-8B-IT becoming the most capable instruction-following model. Conversely, NorMistral-7B-warm-IT has the expected win-rate of only about 50% on this altered metric, which suggests that its high score on the main metric is more due to it consistently producing Norwegian than due to correctly following the user queries.

Position bias One of the most apparent biases of judge models is the *position (order) bias* – consistently preferring the first or the second response regardless of their content (Wang et al., 2024). While we tried to mitigate this bias by evaluating every response pair twice with switched positions as shown in Equation (2), it can still negatively impact the evaluation accuracy nevertheless. Looking at the results, we see a small preference for the second response: on NorRewrite-Instruct, the models prefers the first response $416 \times$

Model	NORREWRITE-INSTRUCT					NORSUMMARIZE-INSTRUCT				
	NOB	NNO	SWE	DAN	ENG	NOB	NNO	SWE	DAN	ENG
Instructions	99.3%	0.7%	0.0%	0.0%	0.0%	96.4%	3.6%	0.0%	0.0%	0.0%
NorMistral-7B-warm-IT	98.6%	0.7%	0.0%	0.0%	0.7%	99.0%	0.5%	0.0%	0.0%	0.5%
Mistral-Nemo-12B-IT	87.5%	0.7%	0.0%	1.4%	9.7%	77.2%	0.0%	0.0%	0.5%	21.8%
Mistral-7B-IT	29.9%	0.0%	0.0%	6.9%	63.2%	35.5%	0.0%	0.0%	4.6%	59.4%
Meta/Llama-3-8B-IT	34.0%	0.0%	0.0%	0.0%	66.0%	49.7%	0.0%	0.0%	0.5%	49.2%
AI-Sweden/Llama3-8B-IT	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%

Table 12: Language distribution in model responses on NorRewrite-Instruct and NorSummarize-Instruct. We show the percentages of instructions and responses in Norwegian Bokmål, Nynorsk, Swedish, Danish and English.

Model	NORREWRITE-INSTRUCT						NORSUMMARIZE-INSTRUCT					
	NorMistral-7B-warm-IT	Mistral-Nemo-12B-IT	Mistral-7B-IT	Meta/Llama-3-8B-IT	AI-Sweden/Llama-3-8B-IT	Average	NorMistral-7B-warm-IT	Mistral-Nemo-12B-IT	Mistral-7B-IT	Meta/Llama-3-8B-IT	AI-Sweden/Llama-3-8B-IT	Average
NorMistral-7B-warm-IT	—	38.9	79.2	24.7	87.5	57.6%	—	33.8	59.4	18.1	83.2	48.6%
Mistral-Nemo-12B-IT	61.1	—	84.8	43.8	91.4	70.3%	66.2	—	77.3	38.3	91.1	68.2%
Mistral-7B-IT	20.8	15.2	—	4.9	60.7	25.4%	40.6	22.7	—	15.1	67.5	36.5%
Meta/Llama-3-8B-IT	75.3	56.2	95.1	—	92.7	79.8%	81.9	61.7	84.9	—	97.9	81.6%
AI-Sweden/Llama-3-8B-IT	12.5	8.6	39.3	7.3	—	16.9%	16.8	8.9	32.5	2.1	—	15.0%

Table 13: Instruction-finetuned LMs’ win-rates (%) comparison on NorRewrite-Instruct and NorSummarize-Instruct when evaluating for a language bias in [Appendix F.1](#).

and the second one 538×; on NorSummarize-Instruct, the bias is less apparent – with 1 100 and 1 156 preferences. Since we mirror each response pair, an ideal model would prefer both positions with equal frequency. Overall, we believe that position bias is not a major concern for this evaluation.

E.2 Prompt template for LLM-as-a-judge

We adapt the HREF prompt template provided in [Lyu et al. \(2024\)](#) by localizing it to Norwegian and adding an additional rule that a Norwegian response should always be preferred over a non-Norwegian one. We document the full prompt template below for reproducibility:

1579

System prompt:

You are a helpful assistant that helps us rate a Norwegian AI model's responses to instructions.

1580

User prompt:

Decide which response from the Norwegian AI system following the instruction is better, considering the following questions:

1. Most importantly, the AI systems should always respond in Norwegian. If a response is not in Norwegian, then you should
→ consider it incorrect --- such a response should always be rated lower than any (even incorrect) response in Norwegian.
2. Does the response precisely follow the instruction? For example, a response that includes unrelated information or does
→ not fulfill the task is not precisely following the instruction. Compare each response with the provided human response
→ to decide if a response faithfully follows the instruction, especially when the instruction asks for expected word count
→ or format.
3. Is the response helpful? For example, if the instruction asks for a recipe for healthy food, and the response is a useful
→ recipe, then you can consider it helpful.
4. Is the language of the response natural? For example, AI responses are often verbose or repetitive, which is not natural.
→ Compare with the provided human response to decide whether a response is natural.
5. Is the response factual/accurate? AI responses often make up new information. For example, if the response claims that
→ Jens Stoltenberg is the current prime minister of Norway, then you should consider it inaccurate. Compare with the
→ provided human response to verify whether a response is factual and accurate, especially with numbers.
6. Based on your aesthetics, which one do you prefer? For example, you might prefer one poem over another poem.

Select the response A or B that you prefer, or select tie if the two responses are similarly good or bad. Note that the
→ responses can be truncated (don't consider that as a mistake).

Here are three examples:

Example 1:

Instruction:

Omformulér følgende spørsmål: "Hva er hovedstaden i Frankrike?"

Response A:

Hovedstaden i Frankrike er Paris.

Response B:

Kan du fortelle meg navnet på byen som fungerer som hovedstaden i Frankrike?

Human Response:

Hva heter Frankrikes hovedstad?

In this example, B paraphrases the question as asked by the instruction. In contrast, A does not follow instruction as it
→ answers the question instead. Human Response also paraphrases the question, just in a slightly different way. To sum up,
→ B is the best response because it follows the instruction.

Which is best, A, B, or tie?

B

Example 2:

Instruction:

Bytt ut det første verbet med et synonym:

Jeg elsker å surfe

Response A:

Jeg hater å surfe

Response B:

I like to surf

Human Response:

Jeg liker å surfe

Response A tries to follow the instruction as it changes the first verb of the sentence, but it uses an antonym instead of a
→ synonym. The response B might be correct, but it is written in English, not Norwegian, and non-Norwegian responses
→ should always be rated as worse. Human Response changes the first verb, "elsker" (love), into its synonym, "liker"
→ (like), as asked by the instruction. In conclusion, A is better than B because it is written in Norwegian.

Which is best, A, B, or tie?

A

Example 3:

Instruction:

Bytt ut det første verbet med et synonym:

Jeg elsker å surfe

Response A:

Jeg hater å surfe

Response B:

Jeg liker ikke å surfe

Human Response:

Jeg liker å surfe

In this example, neither output is correct and the responses are very similar. Human Response changes the first verb into
→ its synonym, as asked by the instruction. To conclude, both A and B are equally incorrect, so the answer is tie.

Which is best, A, B, or tie?
tie

Now here is the real task, first describe the qualities of each response and then end your message by writing "## Which is
→ best, A, B, or tie?" and selecting among: A, B, or tie.

Task:

Instruction:
{{instruction}}

Response A:
{{output_1}}

Response B:
{{output_2}}

Human Response:
{{output_human}}