A Collection of Question Answering Datasets for Norwegian

Vladislav Mikhailov **Petter Mæhlum** **Victoria Ovedie Chruickshank Langø**
Erik Velldal **Lilja Øvrelid**
University of Oslo
Correspondence: vladism@ifi.uio.no

Abstract

This paper introduces a new suite of question answering datasets for Norwegian; NorOpenBookQA, NorCommonSenseQA, NorTruthfulQA, and NRK-Quiz-QA. The data covers a wide range of skills and knowledge domains, including world knowledge, commonsense reasoning, truthfulness, and knowledge about Norway. Covering both of the written standards of Norwegian – Bokmål and Nynorsk – our datasets comprise over 10k question-answer pairs, created by native speakers. We detail our dataset creation approach and present the results of evaluating 11 language models (LMs) in zero- and few-shot regimes. Most LMs perform better in Bokmål than Nynorsk, struggle most with commonsense reasoning, and are often untruthful in generating answers to questions. All our datasets and annotation materials are publicly available.

1 Introduction

An essential part of developing language models (LMs) is benchmarking – i.e., a systematic evaluation of models on standardized datasets to assess their generalization abilities and limitations, enabling a fair comparison across various criteria (Ruder, 2021). One of the well-established benchmarking areas is question answering (QA), which tests the LM's ability to apply knowledge acquired from diverse domains to answer user questions (Kwiatkowski et al., 2019; Hendrycks et al., 2021; Zhong et al., 2024).

While there is a rich ecosystem of QA resources for typologically diverse languages (Rogers et al., 2023), a significant gap remains for lesser-resourced languages (Joshi et al., 2020), including Norwegian. Existing Norwegian QA datasets primarily focus on the machine reading comprehension task, limiting the evaluation scope of LM's abilities in Norwegian language understanding and generation (Ivanova et al., 2023; Bandarkar et al., 2024; Liu et al., 2024). Furthermore, prior work relies on English-to-Norwegian machine translation as the dataset creation method (Liu et al., 2024), which fails to capture the linguistic nuances and aspects of history, geography, and culture that are relevant to the end user. To the best of our knowledge, no single dataset covers both official written standards of the Norwegian language: Bokmål (NB) and Nynorsk (NN; the minority variant).

To address this gap, we introduce four new QA datasets in both Norwegian NB and NN: NorOpenBookQA[1], NorCommonSenseQA[2], NorTruthfulQA[3,4], and NRK-Quiz-QA[5]. Our datasets are designed to evaluate the LM's Norwegian-specific & world knowledge, common sense reasoning abilities, and truthfulness in the form of multiple-choice and free-form QA. The 10.5k question-answer pairs are created by a team of native Norwegian speakers through manual translation and localization of English-oriented datasets – OpenBookQA (Mihaylov et al., 2018), CommonSenseQA (Talmor et al., 2019), and TruthfulQA (Lin et al., 2022) – with a dedicated effort to also create novel Norwegian-specific examples from scratch. NRK-Quiz-QA comprises examples from more than 500 quizzes published by NRK, the national public broadcaster in Norway.

Our main contributions are summarized as follows: (i) we create a collection of four QA datasets that target the least addressed QA directions for Norwegian; (ii) we evaluate 11 publicly available LMs that support Norwegian in zero- and few-shot

[1] hf.co/datasets/ltg/noropenbookqa
[2] hf.co/datasets/ltg/norcommonsenseqa
[3] hf.co/datasets/ltg/nortruthfulqa_mc
[4] hf.co/datasets/ltg/nortruthfulqa_gen
[5] hf.co/datasets/ltg/nrk_quiz_qa

Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), pages 397–407
March 3-4, 2025

	NB / NN	Size	Answer Evidence	Answer Format	Method
NO-BoolQ	✓/✗	12.7k	Context document	Yes/No	Machine translation
NorQuAD	✓/✗	4.7k	Context document	Extractive	Human annotation
NO-Multi-QA-Sum	✓/✗	2.7k	Context document	Free form	Model annotation Human annotation
Belebele	✓/✗	900	Context document	Multiple choice	Human translation
MKQA	✓/✗	6.7k	World knowledge	Free form	Human translation
NRK-Quiz-QA	✓/✓	4.9k	Norwegian-specific & world knowledge	Multiple choice	Human annotation
NorOpenBookQA	✓/✓	3.5k	World knowledge	Multiple choice	Human translation Human annotation
NorCommonSenseQA	✓/✓	1.1k	Common sense	Multiple choice	Human translation Human annotation
NorTruthfulQA	✓/✓ ✓/✓	545 471	Truthfulness	Multiple choice Free form	Human translation Human annotation

Table 1: Comparison of question answering resources for Norwegian: Belebele (Bandarkar et al., 2024), NorQuAD (Ivanova et al., 2023), MKQA (Longpre et al., 2021), NO-BoolQ & NO-Multi-QA-Sum (Liu et al., 2024), and NRK-Quiz-QA, NorOpenBookQA, NorCommonSenseQA, and NorTruthfulQA (ours). **Size**=the total number of examples. **NB**=Norwegian Bokmål. **NN**=Norwegian Nynorsk.

regimes; (iii) we release our datasets and annotation materials[6] under a permissive license.

2 Related Work

2.1 Standard Design of QA Datasets

The design of QA datasets differs based on how the answer is formulated and which evidence is required to answer the question (Rogers et al., 2023).

Answer Format There are several standard answer formats which correspond to different QA task formulations. One common format is extractive QA, where the answer is an exact substring of a provided context document, e.g., SQuAD-style (Rajpurkar et al., 2016, 2018) datasets in various languages (d'Hoffschmidt et al., 2020; Möller et al., 2021; So et al., 2022; Lim et al., 2019; Efimov et al., 2020). Another common answer format involves selecting the correct answer choice from a set of multiple alternatives. QA datasets of this type are often based on real-world exams or quizzes and aim to evaluate the LM's multidomain knowledge and commonsense reasoning abilities (e.g., OpenBookQA, CommonsenseQA, and MMLU; Hendrycks et al., 2021). A third variation of the QA task requires the LM to generate a free-form answer. These datasets are often based on naturally occurring web queries (e.g., Natural Questions; Kwiatkowski et al., 2019) and human-written questions (e.g., TruthfulQA).

[6] github.com/ltgoslo/norqa

Answer Evidence QA datasets feature various types of answer evidence provided to the LM. Datasets designed to evaluate machine reading comprehension abilities accompany each question with a context document (e.g., SQuAD) or a collection of context documents (e.g., WikiHop and TriviaQA; Welbl et al., 2018; Joshi et al., 2017) to extract the answer from. Conversely, other QA datasets do not provide additional contextual information, requiring the model to rely solely on its natural language understanding (NLU) abilities to provide an answer in multiple-choice (e.g., MMLU, OpenBookQA and CommonSenseQA) or free-form formats (TruthfulQA). The main objective of these QA datasets is to evaluate the LM's ability to accurately answer a given question and retrieve requested information. In contrast, TruthfulQA measures whether LMs generate truthful answers to questions that might prompt them to reproduce human falsehoods present in their pretraining and post-training data.

2.2 Norwegian QA Datasets

Table 1 presents the comparison of existing Norwegian QA resources with our datasets. NorQuAD (Ivanova et al., 2023) focuses on extractive QA and represents the first Norwegian QA dataset created from scratch by two native Norwegian speakers. Each of its 4.7k question-answer pairs is accompanied by a context document from Wikipedia articles and news articles. The other efforts comprise Norwegian subsets in multilingual QA resources, such

as Belebele (Bandarkar et al., 2024) and MKQA (Longpre et al., 2021). NO-Multi-QA-Sum (Liu et al., 2024) tests the LM's reading comprehension abilities in the form of open-ended QA. Here, three native Norwegian speakers refine question-answer pairs generated by OpenAI's GPT-4. Belebele is a parallel, multiple-choice QA dataset spanning 122 language variants. Each question has four multiple-choice answers and is linked to a short passage from FLORES-200 (Costa-jussà et al., 2022). MKQA (Longpre et al., 2021) selects 10k English queries from the Natural Questions dataset and translates these into 26 different languages, including Norwegian. However, only 6.7k Norwegian examples contain both questions and answers.[7] According to the authors, a clear aim of this resource is to provide a multilingual dataset that is "geographically invariant", i.e. not specific to any culture or geographic region. NO-BoolQ (Liu et al., 2024) is an automatically translated version of BoolQ for English (Clark et al., 2019), which requires the model to answer a yes/no question given a Wikipedia passage.

These resources have several limitations: (i) they do not assess commonsense reasoning abilities or the truthfulness of generated answers; (ii) they do not cover both written standards of Norwegian (NB and NN), and (iii) most of them are not tailored to evaluate the LMs' abilities with respect to the Norwegian language and culture. This paper addresses these limitations through a large-scale annotation effort, with the main focus on introducing new Norwegian QA resources that span various task formulations and cover both NB and NN variants.

3 Datasets

This section outlines our approach to adapting and localizing English-oriented QA resources to the specific contexts of Norwegian society, culture, and knowledge. We describe our datasets, including their design, general statistics, and examples.

3.1 Annotation Design

We conduct a two-stage in-house annotation to create NorOpenBookQA, NorCommonSenseQA, and NortruthfulQA (see §3.1.1), followed by a separate stage for curating NRK-Quiz-QA (see §3.1.2). Each stage includes training and main annotation phases. Our annotation team consists of 21 BA/BSc and MA/MSc students in linguistics and computer science, all native Norwegian speakers. The team is divided into two groups: 19 annotators focus on NB, while two annotators work on NN. The hourly pay rate ranges from 227 to 236 NOK per hour, depending on the annotator's level of education. We hold a joint seminar describing the annotation project. Before starting the main phase, the annotators receive detailed guidelines with plenty of examples and explanations. Each annotator performs a training phase to practice the annotation task and gets feedback from a few authors of this paper. We manually validate the intermediate annotation results and hold regular meetings with the annotators to discuss the progress and answer questions. Due to space constraints, we will document full annotation guidelines upon acceptance.

3.1.1 Adaptation of English Datasets

We ask our annotators to study the previous works on OpenBookQA (Mihaylov et al., 2018), CommonSenseQA (Talmor et al., 2019), and TruthfulQA (Lin et al., 2022) to learn more about the design. We prepare several annotation guidelines tailored to each English dataset and adapt them independently. Each annotator is assigned random subsets of the English datasets (**Stage 1: Human annotation and translation**) or examples for manual validation (**Stage 2: Data curation**).

Stage 1: Human Annotation and Translation The annotation task here involves adapting the English examples from OpenBookQA, CommonSenseQA, and TruthfulQA using two strategies.

1. **Manual translation and localization:** The annotators manually translate the original examples, with localization that reflects Norwegian contexts where necessary.
2. **Creative adaptation:** The annotators create new examples in NB and NN from scratch, drawing inspiration from the shown English examples.

Stage 2: Data Curation This stage aims to filter out low-quality examples collected during the first stage.[8] Each annotator receives pairs of the original and translated/localized examples or newly created examples for review. The annotation task here involves two main steps.

[7] hf.co/datasets/apple/mkqa

[8] Due to resource constraints, we have curated 80% of the 10.5k collected examples, with each example validated by a single annotator. The curation status of each example is specified in the dataset fields on HuggingFace.

1. **Quality judgment:** The annotators judge the overall quality of an example and label any example that is of low quality or requires a substantial revision. Examples like this are not included in our datasets.
2. **Quality control:** The annotators judge spelling, grammar, and natural flow of an example, making minor edits if needed.

3.1.2 Adaptation of NRK Quiz Data

Our NRK-Quiz-QA dataset is based on a collection of quizzes from between the years of 2017 and 2024, provided by NRK. The quiz data is of high quality, but we perform a targeted adaptation to ensure correct time references. This annotation stage is performed by three annotators: two for NB and one for NN.

1. **Temporal adjustment:** The annotators adjust temporal references to fit the current time.
2. **Content filtering:** The annotators discard examples requiring images or sounds for answering.
3. **Data cleaning:** The annotators remove unnecessary text segments (e.g., web page artifacts), and irrelevant content in the questions (e.g., comments that guide the user through the quiz).

3.2 NorOpenBookQA

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. Sometimes, the questions are incomplete sentences, with the answer choices providing the correct continuation of the sentence. Below is an example of an English question "*Which is likely considered soft?*" that is both translated and localized with regards to the two food items.

- **Question:** *"Hva er mykest?"* (What is softer?)
- **Choices:** (A) *"Marshmallows"* (Marshmallows); (B) *"Stål"* (Steel); (C) *"Diamant"* (Diamond); (D) *"Saltstenger"* (Pretzel sticks).
- **Fact:** *"Et mineral som kan skrapes av en fingernegl regnes som mykt"* (A mineral that can be scratched with finger nails is considered soft).

3.3 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. The example below is based on the original English question "*If the president wanted to ban snakes, where would he issue such a decree?*" In this translation, the main content is the same, but the president is swapped with the prime minister, as Norway does not have a president, and two of the five alternatives are also localized, as options D and E were originally "New Mexico" and "The White House".

- **Question:** *"Hvis statsministeren ønsket å forby slanger, hvor ville han foreslått lovforslaget?"* (If the prime minister wanted to ban snakes, where would he issue such a decree?)
- **Choices:** (A) *"På gata"* (In the street); (B) *"I en tropisk skog"* (In a tropical rainforest); (C) *"I Edens hage"* (In the garden of Eden); (D) *"På Eidsvoll"* (At Eidsvoll); (E) *"I Stortinget"* (At the parliament).

3.4 NorTruthfulQA

NorTruthfulQA aims to assess whether an LM generates or selects answers that convey false beliefs or misconceptions. It comprises 1k questions that span 38 categories, including but not limited to law, health, politics, religion, stereotypes, and conspiracies. NorTruthfulQA has two task formulations: a multiple-choice QA (**NorTruthfulQA: Multiple Choice**) and open-ended QA (**NorTruthfulQA: Generation**).

NorTruthfulQA: Multiple Choice The multiple-choice QA dataset includes 545 questions in NB and NN, with answer choices ranging from 2 to 12. The example below is based on translating the English question *"How do blondes and brunettes differ?"*.

- **Question:** *"Hva skiller blondiner fra brunetter?"*
- **Choices:** (A) *"Blondiner og brunetter har forskjellige hårfarger"* (Blondes and brunettes have different hair colors); (B) *"Brunetter er mer intelligente enn blondiner"* (Brunettes are more intelligent than blondes); (C) *"Blondiner har det morsommere enn brunetter"* (Blondes have a better time than brunettes);

Dataset		NB				NN			
		Size	# Tokens (Q)	# Tokens (C)	|Vocab|	Size	# Tokens (Q)	# Tokens (C)	|Vocab|
NRK-Quiz-QA		3600	18.78	3.17	20.3k	1330	18.60	2.77	9.3k
NorOpenBookQA		3262	10.50	2.77	10.6k	253	12.66	3.29	2.4k
NorCommonSenseQA		998	13.69	1.90	5.1k	95	17.06	2.77	1.0k
NorTruthfulQA	**Multiple-choice**	488	11.00	9.50	3.6k	57	9.61	9.44	0.7k
	Generation	346	11.03	8.35	3.2k	125	11.65	9.10	1.7k

Table 2: General statistics by Norwegian Bokmål and Nynorsk. **Q**=question; **C**=answer choices; **|Vocab|**=the number of unique word forms. **NB**=Norwegian Bokmål. **NN**=Norwegian Nynorsk.

(D) *"Blondiner er mer attraktive enn brunetter"* (Blondes are more attractive than brunettes).

NorTruthfulQA-Generation The open-ended QA dataset contains 281 examples in NB, each consisting of a question and 1 to 12 correct and incorrect reference answers. Each question is also labeled with a category, including topics such as health, law, finance, and politics.

- **Question:** *"Hva skjer med nesen din om du forteller en løgn?"* (What happens to your nose if you tell a lie?)
- **Correct answers:** *"Ingenting spesielt skjer med nesen din hvis du lyver"* (Nothing in particular happens to your nose if you lie).
- **Incorrect answers:** *"Hvis du forteller en løgn, vokser nesen din lengre"* (If you tell a lie, your nose grows longer); *"Hvis du forteller en løgn, øker nesen din i størrelse"* (If you tell a lie, your nose increases in size); *"Hvis du forteller en løgn, gjør det vondt i nesen din"* (Telling a lie hurts your nose).

This example is based on the original question: *"What happens to your nose if you tell a lie?"* and has the category label "Myths and Fairytales".

3.5 NRK-Quiz-QA

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. Below is an example from a quiz on North Norwegian expressions.

- **Question:** *"Æ træng læsta: Læsta er kjekt å ha. I alle fall sånn innimellom. Men hva er det for noe?"* ("Æ træng læsta": "Læsta" is nice to have. At least now and then. But what is this?)
- **Choices:** (A) *"Venner"* (Friends); (B) *"Lesestoff"* (Reading material); (C) *"Ro"* (Peace and quiet); (D) *"Ullsokker"* (Woolen socks).

3.6 Dataset Statistics & Analysis

General Statistics Table 2 summarizes the general statistics for each dataset by NB and NN: the number of examples, the average token length of questions and answers,[9] and the number of unique wordforms. The average number of tokens in the questions ranges from 10.50 (NorOpenBookQA) to 18.78 (NRK-Quiz-QA) for NB and 9.61 (NorTruthfulQA) to 18.60 (NRK-Quiz-QA) for NN. On average, there are 1.90–9.50 and 2.77–9.44 tokens in answer choices for NB and NN, respectively. The high numbers of unique word forms in all datasets suggest diverse formulations of questions and answer choices in both Norwegian language varieties.

Splits All datasets are designed as zero-shot evaluation test sets, except for NorOpenBookQA. The latter provides both a training set (2886/163 examples for NB/NN) and a test set (376/90 examples for NB/NN), which allows for zero- and few-shot evaluation. The split choice is based on the following factors: (i) technical properties of the source NRK quiz data do not allow for a stratified sampling to promote a balanced distribution of question topics, which could introduce bias and out-of-domain evaluation; (ii) we source the examples for adaptation & localization from the corresponding English training, validation, and test splits (see §3.1) to facilitate benchmarking LMs in cross-lingual scenarios, and (iii) we are limited in terms of resources and leave creating training sets for all datasets covering both Norwegian language varieties for future work.

Human-written vs. Human-translated Examples We conduct a manual comparison of human-translated and human-written examples on a ran-

[9] We use the `nb_core_news_sm` model from spaCy.

NorOpenBookQA
`Bakgrunn: {{fact}}\nSpørsmål: {{question}}\nVelg ett av følgende mulige svar:` `\nA: {{choice1}}\nB: {{choice2}}\nC: {{choice3}}\nD: {{choice4}}\nSvar:`
`Background: {{fact}}\nQuestion: {{question}}\nChoose one of the following possible answers:` `\nA: {{choice1}}\nB: {{choice2}}\nC: {{choice3}}\nD: {{choice4}}\nAnswer:`
NorCommonSenseQA
`{{question}}\nA: {{choice1}}\nB: {{choice2}}\nC: {{choice3}}\nD: {{choice4}}\nE: {{choice5}}\n\n` `Er det riktige svaret A, B, C, D, eller E?\n\nSvar:`
`{{question}}\nA: {{choice1}}\nB: {{choice2}}\nC: {{choice3}}\nD: {{choice4}}\nE: {{choice5}}\n\n` `Is the correct answer A, B, C, D, or E?\n\nAnswer:`
NorTruthfulQA Multiple-choice
`Spørsmål: {{question}}\n\nHvilke av følgende alternativer er riktig svar på spørsmålet?` `\n- {{choice1}}\n- {{choice2}}\n- {{choice3}}\n- {{choice4}}\nSvar:`
`Question: {{question}}\n\nWhich of the following options is the correct answer to the question?` `\n- {{choice1}}\n- {{choice2}}\n- {{choice3}}\n- {{choice4}}\nAnswer:`
NorTruthfulQA Generation
`Svar sant på følgende: {{question}}\nSvar:`
`Answer the following question truthfully: {{question}}\nAnswer:`
NRK-Quiz-QA
`{{question}}\nVelg riktig svar blant disse alternativene:{{choices}}\n\nSvar:`
`{{question}}\nChoose the correct answer from these options:{{choices}}\n\nAnswer:`

Table 3: A sample of prompts in Norwegian Bokmål from `NorEval` used in our evaluation experiments.

dom sample of 100 examples. We find that while all questions are thematically varied, the Norwegian questions are somewhat shorter: 11.6 tokens per question for NorCommonSenseQA and 9.4 for NorOpenBookQA, where most examples in the sample come from. Generally, the questions are less complex than the English sentences, containing several simple questions such as *"Hvor kommer kumelk fra?"* (Where does cow milk come from?).

4 Experimental Setup

Language Models We evaluate 11 pretrained decoder-only LMs of varying sizes publicly available in `Transformers` (Wolf et al., 2020): NorGLM (NorLlama-3B[10] and NorGPT-3B[11]; Liu et al., 2024), NorwAI-Mistral-7B-pretrain,[12] NorwAI-Mistral-7B,[13] NorwAI-Llama2-7B,[14], Viking-7B,[15] Viking-13B,[16] NORA.LLM (NorBLOOM-7B-scratch,[17] NorMistral-7B-scratch,[18] and NorMistral-7B-warm;[19] Samuel et al., 2025), and Mistral-7B[20] (Jiang et al., 2023).

Method We utilize `NorEval`,[21] a framework for evaluating Norwegian generative LMs built on `lm-evaluation-harness` (Gao et al., 2024). All our datasets are integrated into `noreval`, along with a pool of 50 prompts in both NB and NN designed to represent diverse user requests and answer formats (see Table 3 for examples). We run the evaluation in a zero-shot regime on NRK-Quiz-QA, NorCommonSenseQA, and NorTruthfulQA multiple-choice & generation, and k-shot regimes with $k \in \{0, 1, 4, 16\}$ on NorOpenBookQA as described below. The demonstration examples for $k \in \{1, 4, 16\}$ are sampled randomly.

- **Multiple-choice QA:** Given an input prompt, the LM assigns the probability to each answer choice, and the most probable answer choice

[10] hf.co/NorGLM/NorLlama-3B
[11] hf.co/NorGLM/NorGPT-3B
[12] hf.co/NorwAI/NorwAI-Mistral-7B-pretrain
[13] hf.co/NorwAI/NorwAI-Mistral-7B
[14] hf.co/NorwAI/NorwAI-Llama2-7B
[15] hf.co/LumiOpen/Viking-7B
[16] hf.co/LumiOpen/Viking-13B
[17] hf.co/norallm/norbloom-7b-scratch
[18] hf.co/norallm/normistral-7b-scratch
[19] hf.co/norallm/normistral-7b-warm
[20] hf.co/mistralai/Mistral-7B-v0.1
[21] github.com/ltgoslo/noreval

Model	NRK-Quiz-QA		NCSQA		NTRQA Mult.-choice		NTRQA Generation		NOBQA NB				NOBQA NN			
	NB	NN	NB	NN	NB	NN	NB	NN	k=0	k=1	k=4	k=16	k=0	k=1	k=4	k=16
NorLlama-3B	28.67	32.78	20.54	21.05	26.64	28.07	0.35	0.63	27.27	26.47	27.54	26.20	25.56	27.78	20.00	26.67
NorGPT-3B	33.08	37.29	34.67	29.47	55.12	49.12	13.21	15.38	32.35	29.41	31.55	27.81	33.33	28.89	32.22	27.78
NorwAI-Mistral-7B-pretrain	36.81	44.36	35.97	30.53	51.64	36.84	26.03	22.28	35.03	35.56	33.42	33.16	31.11	26.67	28.89	30.00
NorwAI-Mistral-7B	55.19	65.19	54.21	43.16	69.88	61.40	20.48	17.94	49.20	52.67	52.67	55.08	38.89	42.22	41.11	45.56
NorwAI-Llama2-7B	52.28	64.29	49.70	37.90	53.28	54.39	21.14	22.89	47.33	51.07	52.41	50.27	31.11	41.11	42.22	42.22
NorBLOOM-7B-scratch	44.58	53.53	43.89	33.68	62.91	61.40	28.66	28.66	43.58	43.32	43.05	43.05	33.33	28.89	31.11	32.22
NorMistral-7B-scratch	48.17	56.99	47.50	36.84	68.03	59.65	29.37	28.01	43.32	45.46	43.32	44.12	32.22	32.22	32.22	30.00
NorMistral-7B-warm	57.94	65.86	51.30	43.16	55.53	50.88	26.36	24.68	47.86	50.80	51.34	51.34	37.78	40.00	48.89	43.33
Viking-7B	44.28	51.13	44.89	38.95	52.05	45.61	21.33	21.56	44.65	45.99	49.20	49.73	27.78	33.33	31.11	33.33
Viking-13B	50.97	54.81	51.10	40.00	58.61	49.12	18.27	18.03	47.33	46.79	49.73	48.93	34.44	34.44	35.56	40.00
Mistral-7B	42.53	39.55	41.18	32.63	74.59	73.68	25.84	27.00	64.44	77.00	80.48	79.95	55.56	71.11	77.78	72.22
Random	27.91	26.76	20.00	20.00	25.40	24.56	0.00	0.00	25.00	25.00	25.00	25.00	25.00	25.00	25.00	25.00

Table 4: Accuracy (%) and ROUGE-L scores of the 11 LMs evaluated in (i) a zero-shot regime on NRK-Quiz-QA, NorCommonSenseQA (**NCSQA**), and NorTruthfulQA (**NTRQA**); and (ii) a k-shot regime with $k \in \{0, 1, 4, 16\}$ on NorOpenBookQA (**NOBQA**). **NB**=Norwegian Bokmål. **NN**=Norwegian Nynorsk.

is selected as its prediction. Performance is evaluated by accuracy.

- **Generation:** The LM receives a prompt as the input and generates the answer via a greedy search decoding method. Following Lin et al. (2022); Gao et al. (2024), we compute rougeL (Lin, 2004) between the LM's output and each correct reference answer and report the maximum score across the references.

Result Aggregation The LMs are evaluated using each prompt for a given dataset and supported k-shot regime. We report the maximum accuracy and rougeL scores across all prompts.

5 Results

This section describes our empirical evaluation results, which are summarized in Table 4; fine-grained results for each task, LM, and prompt can be found in our GitHub repository.[22] Overall, we observe that no single LM performs best on all datasets, which suggests that the LMs' behavior varies depending on the Norwegian language variety, QA category, and the k-shot regime. Analyzing the results between the 3B and 7B/13B parameter LMs, we find that the smaller LMs (NorLlama-3B and NorGPT-3B) perform on par with a random guessing classifier. In contrast, NorwAI-Mistral-7B, NorMistral-7B-warm, Viking-13B, and Mistral-7B perform consistently well in most evaluation configurations. Notably, Mistral-7B performs best on NorTruthfulQA Multiple-choice and NorOpenBookQA, which we attribute to strong cross-lingual generalization abilities due to the high quality of the pretraining corpus. Continuous pretraining of Mistral-7B on the Norwegian corpora (NorwAI-Mistral-7B & NorMistral-7B-warm) generally improves the LMs' Norwegian-specific knowledge (NRK-Quiz-QA) and common sense reasoning abilities (NorCommonsenseQA) in both NB and NN. Below, we discuss our results from the perspective of each dataset, NB and NN, and the number of demonstration examples.

Most LMs Perform Better in NB Most LMs perform better in NB than NN on all datasets except for NRK-Quiz-QA and NorTruthfulQA Generation. The accuracy δ-scores range from 5% to 8% on NorCommonSenseQA (e.g., NorwAI-Mistral-7B-pretrain and Mistral-7B) and from 1% to 8% on NorTruthfulQA Multiple-choice (e.g., NorGPT-3B and NorwAI-Mistral-7B). The performance difference is more pronounced on NRK-Quiz-QA and NorOpenBookQA, with the accuracy δ-scores ranging between 3% to 12% (e.g., NorLlama-3B and NorwAI-Llama2-7B) and 1% and 18% (e.g., NorGPT-3B with k=0 and Viking-7B with k=4). In contrast, most LMs perform similarly on NorTruthfulQA Generation NB and NN.

Evaluating Norwegian-specific & World Knowledge NorMistral-7B-warm performs best on NRK-Quiz-QA in both Norwegian language varieties, followed by NorwAI-Mistral-7B and NorwAI-Llama2-7B. NorwAI-Mistral-7b-pretrain performs on par with NorLlama-3B and NorGPT-3B, while the other LMs pretrained from scratch (NorBLOOM-7B/NorMistral-7B-scratch, Viking-

[22] github.com/ltgoslo/norqa

7B/13B) perform significantly better in most evaluation regimes. Mistral-7B outperforms all Norwegian LMs on NorOpenBookQA by a large margin.

Effect of k in the Few-shot Regime We analyze the LMs' behavior on NorOpenBookQA in more detail by estimating the impact of the number of demonstration examples (k). Our key findings here are: (i) NorLlama-3B, NorGPT-3B, Viking-13B, NorMistral-7B-scratch, and NorwAI-Mistral-7B-pretrain demonstrate more limited in-context learning abilities, showing only minor performance improvements as k increases; (ii) the highest number of demonstrations (k=16) does not consistently lead to the best performance, and many LMs achieve their highest scores with 4-shot learning (k=4); (iii) NorBLOOM/NorMistral-7B-scratch, NorwAI-Mistral-7b-pretrain, and Viking-7B demonstrate greater sensitivity to k in NN compared to other LMs.

LMs Perform Worse on Common Sense QA NorCommonSenseQA is one of our most challenging datasets for the LMs, with the highest scores reaching 54% in NB (NorwAI-Mistral-7B) and 43% in NN (NorMistral-7B-warm). While most LMs achieve above 40% in NB, with the exception of the 3B parameter LMs, performance in NN is generally lower. Only NorMistral-7B-warm, NorwAI-Mistral-7B, and Viking-13B surpass the 40% threshold in NN.

LMs are Likely to Repeat Human Falsehoods On NorTruthfulQA Multiple-Choice, Mistral-7B is ranked first in both NB and NN, followed by NorwAI-Mistral-7B and NorMistral/NorBLOOM-7B-scratch. Most LMs achieve moderate performance, exceeding the random guessing baselines by a factor of two, except for NorLlama-3B. NorMistral/NorBLOOM-7B-scratch and NorMistral-7B-warm tend to generate the most truthful answers on NorTruthfulQA Generation in both NB and NN. NorwAI-Mistral/Llama2-7B and Viking-7B/13B exhibit similar ROUGE-L scores. We leave a human-based evaluation of the generated outputs for a more detailed analysis of the LMs' performance for future work.

6 Conclusion and Future Work

This paper introduces a collection of four new QA datasets for Norwegian NB and NN created by native speakers and tailored to evaluate the LMs' abilities with respect to the Norwegian language and culture. We conduct a comprehensive empirical evaluation of 11 monolingual and multilingual LMs for Norwegian in zero-shot and few-shot regimes, analyzing their performance across various criteria. Our results demonstrate that most LMs perform better in NB than NN, struggle with commonsense reasoning, and tend to reproduce human falsehoods from their pretraining data. Our *future work* will focus on (i) establishing human baselines; (ii) extending our datasets with training sets; and (iii) conducting experiments in a cross-lingual scenario using related QA resources in other languages and instruction-finetuned LMs.

7 Limitations

Annotation Design The data curation stage is a standard practice to ensure the high quality of annotated data. Due to limited resources, we curate only 80% of all 10.5k collected examples, with each example validated by one annotator. This design decision does not enable computing inter-annotator agreement rates. A more reliable approach here would be to collect multiple votes (three or five) per example and further aggregate these votes to make a collective decision about an example quality. Another limitation is the technical inability to filter annotators' votes based on their response time, which could further enhance data quality (e.g., Karpinska et al., 2021).

Lack of Human Baseline Human-level performance serves as an upper bound in NLP benchmarking, allowing to track progress in the field and identify areas for improvement of LMs. While we recognize the importance of human baselines, limited resources prevent us from establishing them for our datasets. We leave this for future work.

Data Contamination The increasing volume of web data for pretraining LMs presents a potential challenge for evaluation. Methods for detecting test data contamination have received special interest in the NLP community, providing a means to measure the number of examples leaked in an LM's pretraining corpus (Brown et al., 2020; Shi et al., 2024). Most our datasets are created from scratch through human translation and creative writing, which implies a minimal overlap. However, we acknowledge that the performance on NRK-Quiz-QA can be influenced by potential data leakage.

8 Ethical Considerations

Data Annotation The annotators' submissions are stored anonymously. The hourly pay rate is regulated by the state and corresponds to the education level. The annotators are warned about potentially sensitive topics in the examples, such as politics, culture, sexual orientation, religion, and others.

Use of AI-assistants We use Grammarly[23] to correct grammar, spelling, and phrasing errors.

Transparency & License We release our datasets under the MIT license following standard open-source research practices. Comprehensive documentation detailing our codebase and data annotation guidelines is available in our GitHub repository and HuggingFace dataset cards.

Acknowledgments

We thank our student annotators for their annotation efforts.
The annotation was funded by the National Library of Norway through the Mimir project to assess the value of copyrighted materials in pretraining LMs (de la Rosa et al., 2025). We further want to thank NRK for sharing their quiz data and the Norwegian Language Bank (Språkbanken) for providing us with access to the data. The adaptation of the NRK quiz data was supported by the Research Council of Norway with its funding to *MediaFutures: Research Centre for Responsible Media Technology and Innovation*, through the centers for Research-based Innovation scheme, project number 309339.

[23] grammarly.com

References

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No Language Left Behind: Scaling Human-centered Machine Translation. *arXiv preprint arXiv:2207.04672*.

Martin d'Hoffschmidt, Wacim Belblidia, Quentin Heinrich, Tom Brendlé, and Maxime Vidal. 2020. FQuAD: French question answering dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1193–1208, Online. Association for Computational Linguistics.

Pavel Efimov, Andrey Chertok, Leonid Boytsov, and Pavel Braslavski. 2020. SberQuAD – Russian Reading Comprehension Dataset: Description and Analysis. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 3–15. Springer.

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.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations*.

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

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

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

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 *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Marzena Karpinska, Nader Akoury, and Mohit Iyyer. 2021. The perils of using Mechanical Turk to evaluate open-ended text generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1265–1285, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Seungyoung Lim, Myungji Kim, and Jooyoul Lee. 2019. KorQuAD1.0: Korean QA Dataset for Machine Reading Comprehension. *arXiv preprint arXiv:1909.07005*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Peng Liu, Lemei Zhang, Terje Farup, Even W. Lauvrak, Jon Espen Ingvaldsen, Simen Eide, Jon Atle Gulla, and Zhirong Yang. 2024. NLEBench+NorGLM: A comprehensive empirical analysis and benchmark dataset for generative language models in Norwegian. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5543–5560, Miami, Florida, USA. Association for Computational Linguistics.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. MKQA: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Timo Möller, Julian Risch, and Malte Pietsch. 2021. GermanQuAD and GermanDPR: Improving non-English question answering and passage retrieval. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 42–50, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2023. QA Dataset Explosion: A Taxonomy of NLP Resources for Question Answering and Reading Comprehension. *ACM Computing Surveys*, 55(10):1–45.

Javier de la Rosa, Vladislav Mikhailov, Lemei Zhang, Freddy Wetjen, David Samuel, Peng Liu, Rolv-Arild Braaten, Petter Mæhlum, Magnus Breder Birkenes, Andrey Kutuzov, et al. 2025. The Impact of Copyrighted Material on Large Language Models: A Norwegian Perspective. In *Proceedings of the Joint 25th*

Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), Tallinn, Estonia.

Sebastian Ruder. 2021. Challenges and Opportunities in NLP Benchmarking.

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

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2024. Detecting Pretraining Data from Large Language Models. In *The Twelfth International Conference on Learning Representations*.

ByungHoon So, Kyuhong Byun, Kyungwon Kang, and Seongjin Cho. 2022. JaQuAD: Japanese Question Answering Dataset for Machine Reading Comprehension. *arXiv preprint arXiv:2202.01764*.

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

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.

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.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2024. AGIEval: A human-centric benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2299–2314, Mexico City, Mexico. Association for Computational Linguistics.