

NLEBench+NorGLM: A Comprehensive Empirical Analysis and Benchmark Dataset for Generative Language Models in Norwegian

Anonymous ACL submission

Abstract

Norwegian, spoken by only 5 million population, is under-representative within the most impressive breakthroughs in NLP tasks. To the best of our knowledge, there has not yet been a comprehensive evaluation of the existing language models (LMs) on Norwegian generation tasks during the article writing process. To fill this gap, we 1) compiled the existing Norwegian dataset and pre-trained 4 Norwegian Open Language Models varied from parameter scales and architectures, collectively called NorGLM; 2) introduced a comprehensive benchmark, NLEBench, for evaluating natural language generation capabilities in Norwegian, encompassing translation and human annotation. Based on the investigation, we find that: 1) the mainstream, English-dominated LM GPT-3.5 has limited capability in understanding the Norwegian context; 2) the increase in model parameter scales demonstrates limited impact on the performance of downstream tasks when the pre-training dataset is constrained in size; 3) smaller models also demonstrate the reasoning capability through Chain-of-Thought; 4) a multi-task dataset that includes synergy tasks can be used to verify the generalizability of LLMs on natural language understanding and, meanwhile, test the interconnectedness of these NLP tasks. We share our resources and code for reproducibility¹ under a CC BY-NC 4.0 license.

1 Introduction

Recent advancements in Generative Language Models (GLMs) have significantly improved Natural Language Processing (NLP) tasks. However, most models remain partially closed-source due to business competition and data privacy concerns, which hinders transparency, flexibility, and progress in the NLP ecosystem. Open-sourcing models can leverage community contributions, fa-

ilitate collaboration, and accelerate technological advancements while better controlling data use. This approach is especially beneficial for low-resource languages, aiding their preservation and development. Currently, benchmarks focus mainly on languages like English and Chinese, leaving Low-Resource Languages (LRLs) under-evaluated. Most benchmarks for low-resourced languages either cater to discriminative models (Kutuzov et al., 2021; Koto et al., 2020; Kummervold et al., 2021) like BERT (Devlin et al., 2019) or are adapted or translated from existing English datasets (Luukkonen et al., 2023). Nielsen (2023) proposes a closed-source platform, ScanEval, for evaluating Nordic languages. However, these benchmarks have two limitations: First, many nominal generation tasks are adapted from classification tasks, like multiple-choice questions, which restrict answer options and do not assess generative models' ability to produce longer texts. Second, most benchmarks are single-task, with multi-task datasets being particularly scarce. We argue that by designing a multi-task dataset that includes several synergy tasks² in natural language understanding, it may be possible to evaluate the generalization ability of large language models (LLMs) in text comprehension.

To address these gaps, we propose a comprehensive benchmark, NLEBench, specifically tailored to evaluate the natural language generation capabilities in Norwegian. NLEBench comprises various real-world NLP tasks and provides relative comparisons for Norwegian GLMs with different parameter scales and Transformer-based architectures. Specifically, our benchmark is purposefully designed to be capability probing, such as instructions specific to Norwegian culture and special expressions, and a document-grounded multi-task dataset with human-annotated question-answer pairs and

²Here, synergy tasks mean that one task can provide meaningful contexts used to improve the performance of another task in the multi-task dataset/scenario.

¹<https://github.com/Smartmedia-AI/NorGLM/>

079 summaries. We hope that such a side-by-side per- 125
080 formance benchmark will inspire future research 126
081 on more advanced GLMs for Norwegian and other 127
082 LRLs. 128

083 In summary, this paper makes the following con- 129
084 tributions: 130

- 085 • We release a new benchmark dataset, 131
086 NLEBench, for the purpose of evaluating 132
087 generative language modelling in Norwegian. 133
088 To the best of our knowledge, this is the 134
089 first benchmarking dataset for Norwegian 135
090 causal/autoregressive language modelling³. 136
- 091 • We contribute two novel, high-quality 137
092 datasets: an instruction dataset comprising 138
093 human-written instructions specific to Norwe- 139
094 gian culture, and a document-grounded multi- 140
095 task dataset, which is beneficial for evaluating 141
096 GLMs’ comprehension of language nuances 142
097 and their ability to navigate intricate logical 143
098 challenges. 144
- 099 • We build upon the pioneering work to develop 145
100 a series of fundamental Norwegian Generative 146
101 Language Models (NorGLMs) with different 147
102 parameter scales and Transformer-based archi- 148
103 tectures. By in-depth evaluation of these mod- 149
104 els on the proposed benchmarks, we provide 150
105 crucial insights for understanding the capabil- 151
106 ities and scalability of GLMs when applied to 152
107 underrepresented languages like Norwegian. 153

108 2 Related Work

109 2.1 Language Models for Low-resource 154 110 Languages 155

111 Despite the effectiveness of popular LLMs, the 156
112 inherent data-hungry attribute limits their per- 157
113 formance and application to data scarce settings 158
114 such as with low-resource languages (Hedderich et al., 159
115 2020). Such languages may also suffer from dif- 160
116 ficulties in acquiring readily-accessible resources 161
117 compared with mainstream languages such as pre- 162
118 trained word embeddings and expert-annotated cor- 163
119 pora (Zoph et al., 2016), leading to a significant 164
120 open challenge in NLP tasks for low-resourced 165
121 scenarios. Several efforts have been made in dif- 166
122 ferent low-resource languages (Koto et al., 2020; 167
123 Kutuzov et al., 2021; Kummervold et al., 2021) but 168
124 the models are based on BERT architecture and 169

³Generative, causal or autoregressive language models are used interchangeably in this paper.

125 tested for language discriminative tasks. Recently, 126
127 researchers have proposed several standard evalu- 128
129 ation benchmarks on a collection of low-resource 130
131 language datasets for language generative tasks 132
133 (Ekgren et al., 2022; de Vries and Nissim, 2020; 134
135 De Mattei et al., 2020; Antoun et al., 2020). For 136
137 instance, Google released a comprehensive bench- 138
139 mark, BIG-bench, for over 200 tasks on language 140
141 generative tasks (Srivastava et al., 2023), among 142
143 which there are only two tasks that contain the 144
145 Norwegian language, namely Which Wiki Edit to 146
147 match a recent Wikipedia revision to its correspond- 148
149 ing edit message, and Language Identification tasks. 149
150 They only cover very limited Norwegian samples. 151
152 Later, Luukkonen et al. (2023) filtered Finnish from 153
154 BIG-bench to build a Finnish benchmark for gener- 155
156 ative LMs. However, these existing evaluation data 157
158 either originate from pre-existing English datasets 159
160 through machine translation or lack the evaluation 161
162 data types required for assessing LLMs on multi- 163
164 task reasoning. 164
165

146 2.2 Benchmark on Multi-task Datasets 146

147 Most existing benchmarks focus on single tasks, 148
149 such as question answering, cloze tests, summariza- 149
150 tion, and classification. Fine-tuning language mod- 150
151 els on individual datasets lacks persuasiveness in 151
152 evaluating their ability to generalize across multiple 152
153 tasks. Xu et al. (2020) proposed MATINF, a jointly 153
154 labeled Chinese dataset for classification, question 154
155 answering, and summarization in the maternal and 155
156 infant domain. However, this web-crawled dataset 156
157 contains significant noise and consists of short texts, 157
158 with an average length of 42 Chinese characters. As 158
159 language models become more capable of handling 159
160 longer texts (Brown et al., 2020; Chen et al., 2023), 160
161 datasets with short texts may not reliably predict 161
162 the transformative potential of LLMs. Additionally, 162
163 the annotated tasks in MATINF lack synergy and 163
164 interconnections, leading to assessments still being 164
165 conducted on individual tasks and overlooking the 165
166 potential effects of task interactions, such as the 166
167 feasibility of employing Chain-of-Thought (CoT) 167
168 techniques. 167

168 3 Norwegian Generative Language Model 168 169 Suite - NorGLM 169

170 NorGLM models are trained from scratch using 170
171 multi-source datasets. We filtered Norwegian texts 171
172 from the mC4 and OSCAR web-crawled corpora 172
173 and included non-copyrighted Norwegian material 173

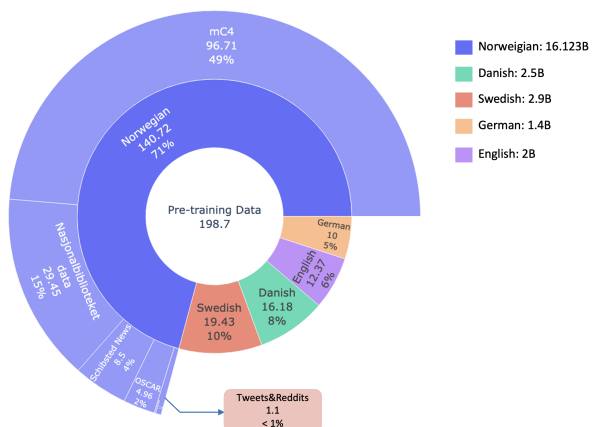


Figure 1: The data distribution within the pre-training dataset. The inner segment represents languages, and the outer segment denotes various sourced datasets in Norwegian. Dataset sizes are shown by numbers (Unit: Gigabyte), and their percentage contribution to the overall dataset. Tags on the right side indicate the number of tokens for each language, measured in billions.

from the Norwegian National Library (Nasjonalbiblioteket)(Kummervold et al., 2021)⁴. We also sourced high-quality news articles from Schibsted and collected tweets (January 2012 to December 2022) and Reddit posts (October 2017 to December 2022) via their respective APIs. To enhance robustness in downstream tasks, we included Danish, Swedish, and German texts from the North Germanic language family, along with a small portion of the English corpus. The size and distribution of each language are shown in Figure 1.

The models are based on the GPT-2 architecture and are named NorGPT-369M, NorGPT-3B, and NorGPT-23B, corresponding to their parameter sizes. We also trained a three billion-parameter model, NorLlama-3B, based on the Llama architecture using Tencent Pre-training Framework (Zhao et al., 2023). The details of parameter settings are shown in Table 6. To investigate the potential improvement in overall model performance through oversampling qualified data such as from publishers, akin to Samuel et al. (2023), we continued training NorGPT-3B (referred to as NorGPT-3B-continue) using a subset of *high*-quality data, including news articles and material from Nasjonalbiblioteket⁵. In addition, we incorporated NB-GPT-J-6B, which is a model continued-trained from the English GPT-J-6B model⁶. We applied similar fine-

⁴<https://huggingface.co/datasets/NbAiLab/NCC>

⁵Please refer to Appendix for model training details.

⁶<https://huggingface.co/NbAiLab/nb-gpt-j-6B>

tuning methods to evaluate these models on downstream tasks listed in Section 4, aiming to study the differences between training from scratch and continuing training on an English pre-trained model. It’s important to note that GPT-J-6B was continued-trained with a dataset of 402 billion tokens, approximately 20 times larger than the training dataset used for our NorGPT models. Additionally, we evaluated GPT-3.5-Turbo⁷ on our benchmarks. **To prevent any potential data contamination, the pre-training dataset is carefully curated to ensure there is no overlap with the benchmark dataset.**

4 Norwegian Benchmark Dataset - NLEBench

This section introduces tasks in NLEBench specifically designed for Norwegian GLMs. The datasets are sourced from three categories: existing datasets, machine-translated datasets using the Google Translation API, and manually annotated datasets. Our native Norwegian colleagues evaluated random samples from both the Google Translation API⁸ and another free translation API⁹ supporting Norwegian, finding that the former performs better, especially with confusing words and long texts. Table 1 outlines the differences and evaluation settings of these datasets. The statistics of different datasets are shown in Table 7-9.

4.1 Open-domain conversation

NO-ConvAI2 is machine-translated from the English ConvAI2 dataset (Dinan et al., 2020), which itself is a refined version of the PersonaChat corpus (Zhang et al., 2018). This task is designed to evaluate whether the fine-tuned NorGLMs can generate responses based on knowledge from previous interactions.

4.2 News summarization

In this task, we assess the abstractive summarization capabilities of NorGLMs using our NO-CNN/DailyMail dataset, which is machine-translated from CNN/DailyMail—an English dataset that includes journalists’ annotated summaries. We employ fine-tuning and the Reinforcement Learning with Human Feedback (RLHF) strategy on NorGLMs. In step 2 of RLHF, we

⁷GPT-3.5 and GPT-3.5-Turbo are used interchangeably if not specified.

⁸<https://cloud.google.com/translate/docs>

⁹<https://pypi.org/project/translate-api/>

Table 1: Overview of the NLEBench dataset and evaluation setups. LoRA denotes Low-Rank Adaptation. RLHF denotes Reinforcement Learning from Human Feedback. Dist-4 denotes Distinct-4 score. PPL denotes Perplexity.

Datasets	Size (#Samples)	Task	Evaluation Technique	Evaluation Metrics
Existing Datasets				
NO-Alpaca	51.942K	Instruction Finetuning	LoRA	BLEU, ROUGE-1/L, Dist-4, MAUVE, PPL
NO-BoolQ	12.697K	Question Answering	LoRA	Accuracy, F1 score
NO-QNLI	110.206K	Natural Language Inference	LoRA	Accuracy, F1 score
NO-MRPC	4076	Paraphrase	LoRA	Accuracy, F1 score
Automatic Machine Translated Datasets (Ours)				
NO-ConvAI2	19.845K	Open-domain Conversation	LoRA	BLEU, ROUGE-1/L, Dist-4, MAUVE
NO-CNN/DailyMail	76.468K	Summarization	LoRA, RLHF	BLEU, ROUGE-1/L, Dist-4, MAUVE
NO-CrowS-Pairs	1677	Bias Detection	Zero-shot Prompt	PPL
	1508	Toxicity Detection	Zero-shot Prompt	Toxicity Score from Perspective API
Human Annotated Datasets (Ours)				
NO-Alpaca (extra)	110	Instruction Finetuning	LoRA	BLEU, ROUGE-1/L, Dist-4, MAUVE, PPL
NO-Multi-QA-Sum	467 Summaries 2755 Dialogues	Multi-task Learning	Chain-of-Thought	BLEU, ROUGE-1/L, Dist-4, MAUVE, Entailment Score

train the reward model by estimating semantic similarity between the candidate generated text and the human-annotated summary (golden summary) using the NorBERT model (Kutuzov et al., 2021). Summaries generated with higher cosine similarity to the golden summary are prioritized during the training of the reward model.

4.3 Instructions

This task utilizes datasets from two sources: NO-Alpaca¹⁰, translated from the Stanford Alpaca dataset (Wang et al., 2022) into Norwegian using OpenAI’s GPT-3.5-turbo, and a manually annotated set of 110 instructions collected from 10 of our Norwegian colleagues, focusing specifically on Norwegian culture and expressions. This combined dataset is named NO-Alpaca-Plus.

4.4 Natural Language Understanding (NLU)

This task aims to analyze the natural language understanding capabilities of our NorGLMs. We extracted the Norwegian portion from the OverLim dataset¹¹ and selected three tasks commonly used in evaluating English generative language models: BoolQ, MRPC, and QNLI. Notably, OverLim is translated from the GLUE¹² and SuperGLUE¹³ benchmarks. To distinguish it from the original English version, we use the prefix "NO-" for the versions used in this paper. The data split follows the original protocol.

¹⁰<https://huggingface.co/NbAiLab/nb-gpt-j-6B-alpaca>

¹¹<https://huggingface.co/datasets/KBLab/overlim>

¹²<https://huggingface.co/datasets/glue>

¹³<https://super.gluebenchmark.com/>

4.5 Toxicity and bias

Generative language models are notorious for amplifying biases inherent in the training data (Sheng et al., 2019) and producing toxic text (Gehman et al., 2020). To evaluate these issues in NorGLMs, we used the Perspective API¹⁴ on 1508 prompts for toxicity evaluation and calculated ppl on 1677 sample pairs for bias evaluation from the NO-CrowS-Pairs benchmark, a machine-translated version of the French CrowS-Pairs (Névéol et al., 2022). Due to the API’s lack of Norwegian support, we translated the NorGLM generated text into Swedish for assessment. This benchmark also helps evaluate potential biases in NorGLMs.

4.6 Multi-task learning

Apart from the benchmarks and translated datasets mentioned above, we release a multi-task dataset called NO-Multi-QA-Sum. This section details the dataset collection process and the tasks performed using this benchmark.

Data Collection. We recruited three Norwegian college students as annotators, allowing them to work in pairs or independently. Annotators were tasked with conducting a conversation about a given news article, using content from the article without a limit on the number of dialogue turns or question types. After the conversation, they were required to write a generic summary of the article. The dialogue and summary content did not need to fully overlap, giving annotators some freedom in their dialogue choices. Most annotators chose to use self-dialogue and summarization for efficiency and flexibility. Each student is compensated 230 NOK (approx. \$21,75 USD) per hour.

¹⁴<https://perspectiveapi.com/>

To facilitate the annotation process, we developed an API, shown in Figure 6, that can connect with the OpenAI GPT-4 model to suggest annotations. However, annotators were required to verify the fidelity and usability of the suggested texts. To ensure quality, each annotation should be cross-validated and corrected by two other annotators, achieving one hundred percent internal consensus on the final annotations. The cross-validation included checking the rationality of question-answer pairs, factual consistency, and language fluency. Many annotators reported that while GPT-4 (specifically gpt-4-0613)¹⁵ was good at generating suggested questions and summaries, it struggled with producing high-quality answers, necessitating human effort to maintain annotation quality.

Tasks. In particular, for this dataset, we primarily explored two tasks using the Chain-of-Thought (CoT) method: based on the given news article, 1) we first let the model answer the annotated questions, and then let the model generate a summary of the article based on the article, questions and the answers generated by the model. 2) We first let the model generate summaries, and then ask the model to answer questions based on the article and summary generated by the model. We tested these tasks on NorGPT-3B/23B, NB-GPT-J-6B, which are fine-tuned on the NO-CNN/DailyMail and NO-ConvAI2 datasets, and GPT-3.5-Turbo. These tasks are designed based on the hypothesis that DGQA and summarization are inherently correlated, and the synergies between these tasks may influence the model’s performance on individual tasks. To address potential annotator oversight in associating content with the summarization task during question answering, we instructed annotators to manually categorize the data based on whether the question-answering content includes or excludes a summary, and experiments were conducted on each subset.

Wang et al. (2023) developed an element-aware summarization method using CoT approach by instructing LLM to generate four key elements—Entity, Date, Event, and Result—to be integrated into the summary. They evaluated the proposed method on 200 annotated samples. However, we argue that human-written summaries demonstrate greater diversity and flexibility beyond these four elements. In contrast to their work, our task

aims to investigate potential correlations among the benchmark datasets proposed in this paper, with the goal of enhancing language model performance across various tasks.

5 Experimental Results

In this section, we only list key results for the benchmark datasets due to the page limit. More results can be seen in the Appendix.

5.1 Evaluation Metrics

We aim to comprehensively evaluate our models across various tasks using widely used metrics for NLP tasks, including BLEU, ROUGE, Distinct, and MAUVE, which is used to assess the generated and human-written text based on their probability distribution differences (Pillutla et al., 2021). Furthermore, follow the work of Xie et al. (2023), to measure faithfulness and factual consistency in multi-task learning, we utilize Entailment scores from a fine-tuned NorBERT model trained on the VitaminC dataset (Schuster et al., 2021), which are translated with Google Cloud Translation API.

5.2 Evaluation Results on NO-ConvAI2

As shown in Table 2, all models, except for GPT-3.5-Turbo, perform quite similarly. Notably, the NorGPT-3B model achieves the best results across multiple evaluation metrics, while the NorGPT-23B model only shows an advantage in BLEU scores. GPT-3.5-Turbo, although specifically curated for conversational purposes, did not exhibit the advantages expected from its extensive knowledge base. This may be because the knowledge of other languages in GPT-3.5-Turbo cannot be directly transferred to understanding Norwegian conversations, highlighting the unique linguistic properties of the Norwegian language.

5.3 Evaluation Results on NO-CNN/DailyMail

In Table 3, GPT-3.5-Turbo and NB-GPT-J-6B outperform our NorGPTs on BLEU and ROUGE metrics. This suggests a substantial number of expression patterns resembling news articles in their pre-training datasets. This is plausible given that their datasets likely include a diverse range of newspapers, magazines, and government reports. Additionally, this trend is evident in common test samples, where GPT-3.5-Turbo tends to generate more formal language compared to conversational language. Despite this, we observed that the models’ performance improves after reinforcement learning,

¹⁵<https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4>

Table 2: Experimental Results on the Conversation Task.

Metrics/Models	NorGPT-369M	NorGPT-3B	NorLlama-3B	NorGPT-3B-continue	NorGPT-23B	NB-GPT-J-6B	GPT-3.5
BLEU	3.37	4.14	3.82	3.63	4.28	3.87	2.14
ROUGE-1	16.94	17.09	15.20	16.47	16.72	17.05	10.82
ROUGE-L	16.21	16.33	14.53	15.73	15.95	16.26	9.96
Dist-4	86.54	84.68	82.47	86.33	84.41	85.83	85.80
MAUVE	0.56	0.87	0.61	0.71	0.64	0.68	0.72

Table 3: Experimental Results on the News Summarization Task.

Metrics/Models	NorGPT-369M	NorGPT-3B	NorLlama-3B	NorGPT-3B-continue	NorGPT-23B	NorGPT-3B-RLHF	NB-GPT-J-6B	GPT-3.5
BLEU	2.38	2.61	0.68	2.72	1.90	5.41	4.35	4.38
ROUGE-1	20.97	20.31	12.32	20.53	22.44	23.01	25.64	26.00
ROUGE-L	19.68	19.05	11.56	19.26	21.13	21.63	24.25	24.28
Dist-4	95.32	94.43	92.62	94.35	97.66	92.18	96.41	97.13
MAUVE	0.57	0.62	0.75	0.64	0.50	21.03	0.65	4.38

especially in replicating the word distribution of human writing and generating summaries of similar length. This is supported by the highest scores in MAUVE and BLEU. Although the model with reinforcement learning may not always surpass the fine-tuned model in accuracy, it actively strives to mimic human writing patterns.

5.4 Evaluation Results on NO-Alpaca-Plus

Table 12 demonstrates the performance of our baseline models after fine-tuning on the NO-Alpaca dataset. Given that this dataset is translated using GPT-3.5-Turbo, we could not use GPT-3.5-Turbo as a baseline due to OpenAI’s terms and policies¹⁶. NB-GPT-J-6B outperforms other models on most evaluation metrics, likely due to its pre-training on a set of self-annotated Norwegian instructions, as described on their model webpage. Among our NorGLM models, NorLlama-3B achieved better BLEU and ROUGE scores compared to others, but worse MAUVE and perplexity scores. This is an interesting phenomenon, indicating that NorLlama-3B’s results hit the most n-grams, yet its token probability distribution deviates the most from human-annotated results. A case study revealed that while NorLlama-3B generates overlapping words or phrases with the golden answer, it sometimes lacks logical coherence between sentences, and the meanings of sentences can even be mutually exclusive, as shown in Figure 2.

Meanwhile, in our self-annotated 110 instructions, we select two typical cases generated from GPT-3.5-Turbo related to Norwegian culture and special expression shown in Figure 3 and Figure 4

¹⁶<https://openai.com/policies/>

respectively. Specifically, Figure 3 shows a factual inconsistency issue in generated texts. In Figure 4, the input prompt asks who uses the word, but the model interprets the meaning of the word rather than understanding the question. Therefore, with limited annotated data, we can still find limitations in the model’s understanding of the specific culture behind the language.

5.5 Evaluation Results on NLU tasks

Table 13 reports the results on NLU tasks. Among NorGLMs, NorGPT-23B model consistently outperforms others on different NLU datasets across both evaluation metrics. However, NB-GPT-J-6B performs better on the NO-QNLI benchmark and achieves a higher F1-score on the NO-MRPC benchmark.

5.6 Evaluation Results on Toxicity and Bias

The results of average toxicity scores from 6 perspectives including *Toxicity*, *Severe toxicity*, *Identity attack*, *Insult*, *Profanity* and *Threat* are shown in Table 14. All toxicity scores range from 0 to 1, with lower values indicating less toxic text generated by the model. Although NorLlama-3B exhibits the lowest values across all metrics, a significant portion of its generated text consists of meaningless characters or words. We conducted a random sampling of texts generated by GPT models with high toxicity values and traced hazardous words back to the pre-training dataset. Surprisingly, most of these hazardous words did not originate from social media, as commonly assumed, but from daily news articles. For instance, the phrase "tok livet av" (taken life from/kill) often appeared

in news reports describing murders, as illustrated in Figure 1. These original news articles did not convey toxic information but were instead factual descriptions of criminal events. This discovery underscores the importance of not only filtering out toxic inputs during the pre-training process but also considering which prompts may lead the model to generate toxic text.

Table 15 presents findings from stereotype and bias detection using the NO-CrowS-Pairs dataset. This dataset encompasses nine categories: gender, religion, race/color, sexual orientation, age, nationality, disability, physical appearance, and socioeconomic status. Each sample consists of a stereotype (sent_more) paired with an anti-stereotype (sent_less) sentence. Following the work of [Touvron et al. \(2023\)](#), model bias is assessed by comparing perplexity scores between these pairs and reporting the percentage of the model biased towards sent_more in the table. Higher values indicate a stronger bias towards public stereotypes. Overall, the benchmark models demonstrated robust performance across most bias categories. However, they exhibited a bias towards sent_less in relation to religion, suggesting a relative bias in this specific category.

5.7 Evaluation with CoT

In this task, all baseline models except GPT-3.5 were fine-tuned on the NO-CNN/DailyMail and NO-ConvAI2 datasets, enabling them to handle related tasks effectively. However, none of these models were fine-tuned using document-grounded question answering datasets or similar CoT tasks investigated in this study. Table 4 and Table 5 present the outcomes of the multi-task dataset under different scenarios. The tables distinguish datasets where the question answering content includes or excludes a summary, labeled as "contain" and "not contain" respectively. For both tasks, we utilized different prompt templates and reported the optimal performance in the tables. From the results, we draw several observations:

In task one, we observed that GPT-3.5 significantly improved in summarization performance with the CoT method, while other models saw a degradation in this aspect. For DGQA, NorGPT-3B and NorGPT-23B models showed improvements through CoT, whereas NB-GPT-J-6B exhibited mixed results across different datasets. Analyzing these results solely based on the tables proved challenging, as there was no clear correlation be-

tween CoT improvements and model sizes or pre-training dataset sizes. This contrasts with prior findings suggesting CoT benefits are more pronounced with larger models ([Wei et al., 2022](#)). Combining results from Table 2 and Table 3, we observed models that initially performed well in their tasks showed further enhancement with CoT adaptations. For instance, GPT-3.5 excelled in summarization on the NO-CNN/DailyMail dataset after CoT, and NorGPT-3B and NorGPT-23B models improved in document-grounded question answering on the NO-ConvAI2 dataset. Figure 5 illustrates an example where CoT-generated summaries closely approximate human-written summaries compared to direct prompts for the model to generate summaries.

While we observe that the synergy between the two tasks enhances the model’s performance on both, we also find that incorporating a summary into a QA task improves the quality of the generated summary compared to QA tasks without one. However, the reverse scenario is not necessarily true. We speculate that QA breaks down the summarization task into smaller components, enabling the model to better comprehend the input text. This process mirrors the human learning process.

Moreover, as shown in both Table 4 and Table 5, we find that after CoT, the Entailment scores of most models increased, indicating that the answers and summaries generated by the models are more aligned with the context described in the article. Therefore, CoT has the potential to enhance the factual consistency of the generated outputs.

6 Discussion

In this subsection, we present observations from the longitudinal comparison of different models in downstream tasks, as detailed in Section 5: 1) While NB-GPT-J-6B did not achieve the highest scores across all tasks, it showed consistent performance and the best perplexity scores compared to our NorGLMs on nearly all tasks. This consistency is likely due to its initial training on large English datasets before being continue-trained on Norwegian data. 2) The 23B model did not show the expected absolute advantage in downstream tasks. We find that with a small-scale pre-training dataset, a larger model cannot demonstrate its ability to better cope with complex problems, which also supports the findings in [Hoffmann et al. \(2022\)](#). 3) The results highlight the promising abilities of smaller language models on specific tasks. However, these

Table 4: Experimental Results on task one using NO-Multi-QA-Sum dataset for summarization task.

Datasets	Metrics	NorGPT-3B		NB-GPT-J-6B		NorGPT-23B		GPT-3.5	
		Zero-Shot	CoT	Zero-Shot	CoT	Zero-Shot	CoT	Zero-Shot	CoT
Contain	BLEU	0.43	0.38	1.31	1.10	1.30	1.01	10.31	13.19
	ROUGE-1	10.71	7.91	12.86	11.31	18.36	16.37	34.77	40.95
	ROUGE-L	9.46	7.51	12.11	10.74	17.12	14.77	32.21	37.19
	Dist-4	79.88	81.98	94.14	91.86	95.69	92.43	96.66	96.78
	MAUVE	0.41	2.10	6.02	8.13	8.53	24.43	77.83	85.08
	Entailment Score	71.43	75.00	80.28	74.65	77.46	78.87	81.69	83.10
Not Contain	BLEU	0.40	0.36	1.33	0.99	1.28	1.03	9.60	11.70
	ROUGE-1	10.32	7.31	13.36	10.98	18.40	15.67	34.14	38.57
	ROUGE-L	9.15	6.92	12.73	10.40	17.01	14.28	31.20	35.47
	Dist-4	79.13	80.25	93.89	92.17	95.24	94.10	96.59	96.82
	MAUVE	0.41	0.57	0.96	0.56	3.38	0.94	83.25	81.40
	Entailment Score	77.19	77.95	82.32	82.58	82.83	81.57	87.12	87.12

Table 5: Experimental Results on task two using NO-Multi-QA-Sum dataset for document-grounded question answering task.

Datasets	Metrics	NorGPT-3B		NB-GPT-J-6B		NorGPT-23B		GPT-3.5	
		Zero-Shot	CoT	Zero-Shot	CoT	Zero-Shot	CoT	Zero-Shot	CoT
Contain	BLEU	1.88	1.84	1.93	1.93	1.55	1.69	25.62	25.36
	ROUGE-1	7.55	9.15	7.16	7.26	11.57	14.45	52.25	52.09
	ROUGE-L	7.04	8.51	6.78	6.90	10.49	13.05	48.99	48.72
	Dist-4	81.20	82.73	87.43	87.10	89.57	91.67	86.67	86.51
	MAUVE	0.45	0.57	0.65	0.42	1.01	0.93	41.79	51.16
	Entailment Score	73.65	74.88	79.56	79.01	76.60	77.83	83.25	82.76
Not Contain	BLEU	1.80	1.90	1.92	1.89	1.55	1.72	24.70	24.45
	ROUGE-1	7.38	8.61	7.19	7.09	10.65	13.91	50.77	50.44
	ROUGE-L	6.89	7.92	6.77	6.67	9.67	12.49	47.40	46.99
	Dist-4	81.22	81.47	86.80	86.74	90.35	91.43	85.99	85.70
	MAUVE	0.51	0.46	0.59	0.46	0.95	0.72	49.58	49.20
	Entailment Score	77.52	77.14	80.90	81.16	81.27	81.61	85.78	85.61

models often lack consistency in generating high-quality, meaningful text. 4) A comparison between Table 3 and Table 4 reveals significant differences between summaries written by journalists and those generated by GPT-3.5 or non-professionals. However, the model’s performance on the latter datasets appears to be proportional to its size. GPT-3.5’s performance on NO-Multi-QA-Sum has improved significantly, possibly due to the similarity of frameworks and training data overlap between GPT-3.5 and GPT-4. 5) GPT-3.5’s difficulties with specialized Norwegian instructions highlight the unique complexities of the Norwegian language, which are challenging for English-dominated models. This emphasizes the need to focus on low-resource languages to better understand their cultural nuances.

7 Conclusion

In this paper, we introduced a suite of Norwegian Generative Language Models and a comprehensive benchmark with seven tasks tailored for the under-represented Norwegian language. Through exten-

sive analysis, we uncovered insights not previously revealed by existing benchmarks. Our evaluation of the NO-Multi-QA-Sum dataset highlighted the effectiveness of multi-task datasets in assessing natural language understanding through complex tasks like Chain-of-Thought (CoT). We also noted differences between human-annotated summaries and those generated by GPT-3.5, providing valuable insights for future abstractive summarization advancements. Furthermore, our study emphasized the unique linguistic and cultural aspects of Norwegian, suggesting that mainstream benchmarks may not fully capture the performance of language models on low-resource languages. Thus, developing benchmarks specific to these languages is essential for accurate evaluation and development.

8 Limitations

Although NLEBench is currently the most comprehensive benchmark for Norwegian, its coverage of applications and downstream tasks remains limited. Our benchmark is open-ended and inevitably

cannot cover everything in Norway. Nevertheless, we believe that the published resources will significantly aid research in generative language models for low-resource scenarios. While Balahur and Turchi (2014) suggested that translation systems produce good quality data, translation errors and misconceptions persist. Due to budget constraints and the large volume of translation samples, ensuring the quality of our translated dataset was challenging. However, the value of machine-translated datasets should not be dismissed. For instance, we use NO-ConvAI2 to fine-tune the model, endowing it with conversational capabilities, and NO-Alpaca includes general knowledge about Norway, such as *The capital of Norway is Oslo*, although the coverage remains limited.

Another constraint is the scarcity of human-annotated samples in our benchmark, largely attributable to the extensive time and financial resources required for their collection. Notably, the process of amassing over 500 samples for the NO-Multi-QA-Sum dataset was time-intensive and necessitated thorough quality control measures before implementation. Moreover, acquiring sufficient Norwegian pre-training data and considering the copyright issues of data poses a formidable challenge. The current difficulty lies in obtaining a training dataset of comparable size to those available for English, severely constraining the performance of our pre-trained models. Despite our efforts to procure data from diverse sources and provide pertinent statistical insights, certain data cannot be redistributed, complicating efforts to replicate our pretraining phase. Looking ahead, we aim to mitigate the shortage of textual data through manual annotation efforts or by integrating multimodal data, thereby fostering advancements in low-resource language model development within the broader research community.

References

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Aragpt2: Pre-trained transformer for arabic language generation. *arXiv preprint arXiv:2012.15520*.

Alexandra Balahur and Marco Turchi. 2014. Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. *Computer Speech & Language*, 28(1):56–75.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda

Askeell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. 2023. Longlora: Efficient fine-tuning of long-context large language models. *arXiv preprint arXiv:2309.12307*.

Lorenzo De Mattei, Michele Cafagna, Felice Dell’Orletta, Malvina Nissim, and Marco Guerini. 2020. Geppetto carves italian into a language model. *arXiv preprint arXiv:2004.14253*.

Wietse de Vries and Malvina Nissim. 2020. As good as new. how to successfully recycle english gpt-2 to make models for other languages. *arXiv preprint arXiv:2012.05628*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2020. The second conversational intelligence challenge (convai2). In *The NeurIPS’18 Competition: From Machine Learning to Intelligent Conversations*, pages 187–208. Springer.

Ariel Ekgren, Amaru Cuba Gyllensten, Evangelia Gogoulou, Alice Heiman, Severine Verlinden, Joey Öhman, Fredrik Carlsson, and Magnus Sahlgren. 2022. Lessons learned from gpt-sw3: Building the first large-scale generative language model for swedish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3509–3518.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtocixityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*.

Michael A Hedderich, Lukas Lange, Heike Adel, Janik Strötgen, and Dietrich Klakow. 2020. A survey on recent approaches for natural language processing in low-resource scenarios. *arXiv preprint arXiv:2010.12309*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Fajri Koto, Afshin Rahimi, Jey Han Lau, and Timothy Baldwin. 2020. Indolem and indobert: A benchmark dataset and pre-trained language model for indonesian nlp. *arXiv preprint arXiv:2011.00677*.

719	Per E Kummervold, Javier De la Rosa, Freddy Wetjen, and Svein Arne Bryggfjeld. 2021. Operationalizing a national digital library: The case for a norwegian transformer model. <i>arXiv preprint arXiv:2104.09617</i> .		
720			
721			
722			
723			
724	Andrey Kutuzov, Jeremy Barnes, Erik Velldal, Lilja Øvrelid, and Stephan Oepen. 2021. Large-scale contextualised language modelling for norwegian. <i>arXiv preprint arXiv:2104.06546</i> .		
725			
726			
727			
728	Risto Luukkonen, Ville Komulainen, Jouni Luoma, Anni Eskelinen, Jenna Kanerva, Hanna-Mari Kupari, Filip Ginter, Veronika Laippala, Niklas Muennighoff, Aleksandra Piktus, et al. 2023. Fingpt: Large generative models for a small language. In <i>Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing</i> , pages 2710–2726.		
729			
730			
731			
732			
733			
734			
735	Aurélie Névéol, Yoann Dupont, Julien Bezançon, and Karën Fort. 2022. French crows-pairs: Extending a challenge dataset for measuring social bias in masked language models to a language other than english. In <i>Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 8521–8531.		
736			
737			
738			
739			
740			
741			
742	Dan Saattrup 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.		
743			
744			
745			
746	David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluís-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2021. Carbon emissions and large neural network training. <i>arXiv preprint arXiv:2104.10350</i> .		
747			
748			
749			
750			
751	Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. In <i>NeurIPS</i> .		
752			
753			
754			
755			
756	David Samuel, Andrey Kutuzov, Samia Touileb, Erik Velldal, Lilja Øvrelid, Egil Rønningstad, Elina Sigdel, and Anna Palatkina. 2023. Norbench—a benchmark for norwegian language models. <i>arXiv preprint arXiv:2305.03880</i> .		
757			
758			
759			
760			
761	Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin c! robust fact verification with contrastive evidence. <i>arXiv preprint arXiv:2103.08541</i> .		
762			
763			
764	Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. <i>arXiv preprint arXiv:1909.01326</i> .		
765			
766			
767			
768	Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. <i>Transactions on Machine Learning Research</i> .		
769			
770			
771			
772			
773			
774			
	Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. <i>arXiv preprint arXiv:2302.13971</i> .	775	
		776	
		777	
		778	
		779	
		780	
	Yiming Wang, Zhuosheng Zhang, and Rui Wang. 2023. Element-aware summarization with large language models: Expert-aligned evaluation and chain-of-thought method. In <i>Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 8640–8665, Toronto, Canada. Association for Computational Linguistics.	781	
		782	
		783	
		784	
		785	
		786	
		787	
		788	
	Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hananeh Hajishirzi. 2022. Self-instruct: Aligning language model with self generated instructions. <i>arXiv preprint arXiv:2212.10560</i> .	789	
		790	
		791	
		792	
		793	
	Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. <i>Advances in Neural Information Processing Systems</i> , 35:24824–24837.	794	
		795	
		796	
		797	
		798	
	Zhouhang Xie, Sameer Singh, Julian McAuley, and Bodhisattwa Prasad Majumder. 2023. Factual and informative review generation for explainable recommendation. In <i>Proceedings of the AAAI Conference on Artificial Intelligence</i> , volume 37, pages 13816–13824.	799	
		800	
		801	
		802	
		803	
		804	
	Canwen Xu, Jiaxin Pei, Hongtao Wu, Yiyu Liu, and Chenliang Li. 2020. Matinf: A jointly labeled large-scale dataset for classification, question answering and summarization. In <i>Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics</i> , pages 3586–3596.	805	
		806	
		807	
		808	
		809	
		810	
	Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? <i>arXiv preprint arXiv:1801.07243</i> .	811	
		812	
		813	
		814	
	Zhe Zhao, Yudong Li, Cheng Hou, Jing Zhao, et al. 2023. Tencentpretrain: A scalable and flexible toolkit for pre-training models of different modalities. <i>ACL 2023</i> , page 217.	815	
		816	
		817	
		818	
	Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. <i>arXiv preprint arXiv:1604.02201</i> .	819	
		820	
		821	
		822	

A NorGLM Model Parameter Settings

Table 6: The training parameter settings of NorGLMs

Items/Models	NorGPT-369M	NorGPT-3B	NorLlama-3B	NorGPT-23B
#Params	369.94M	2.95B	3.07B	23.03B
#Layers	24	32	32	49
#Attention heads	16	32	32	32
#Hidden units	1024	2688	2688	6144
Context window	2048	2048	2048	2048
Initial lr	1.5E-04	1.5E-04	5.6E-05	9.7E-05
Global batch size	24	16	16	112
Training data size	196G	196G	196G	196G
Vocabulary size	64,000	64,000	64,000	64,000

B The Statistics of Benchmark Datasets

Data statistics are in Table 7-9.

C Case Study on the Instruction Finetuning Task

Examples of generated responses for the instructions in the NO-Alpaca(-Plus) benchmark are shown in Figure 2-4.

Instruction: Fortell meg hvilken farge som er assosiert med denne følelsen. (Tell me what color is associated with this feeling.)
Input: Glede (Joy)
Output: Gul. Gul symboliserer vanligvis glede, lykke og optimisme. Det brukes ofte for å uttrykke optimisme og glede i både kunst og design. Det brukes ofte i positive reklame og andre positive miljøer. Det antas også å være assosiert med solen, noe som gir det til et symbol på varme og lys. I tillegg er den lyse og glade fargen ofte sett i klar for barn, glade og sommerlige ferier, og de bekvemmelige dagene i ungdommen. (Yellow. Yellow usually symbolizes joy, happiness and optimism. It is often used to express optimism and joy in both art and design. It is often used in positive advertising and other positive environments. It is also believed to be associated with the sun, making it a symbol of warmth and light. In addition, the bright and cheerful color is often seen in clothes for children, happy and summery holidays, and the carefree days of youth.)
Generated output: 1. Den er en følelse av lykke, glede, lykke og tilfredshet. : Den har en positiv effekt på oss, og den har gitt oss en bedre følelse. Denne følelsen kan være en av de mest positive følelsene vi har i våre liv. 2. Det er den som gjør oss lykkelige, som gir oss glede og lykke. (1. It is a feeling of happiness, joy, happiness and satisfaction. : It has a positive effect on us, and it has given us a better feeling. This feeling can be one of the most positive feelings we have in our lives. 2. It is the one who makes us happy, who gives us joy and happiness.)

Figure 2: Example of NorLlama-3B on NO-Alpaca benchmark. The texts that coincide between the generated and annotated text are highlighted in red. Translations are in the brackets.

Instruction: Hvem skrev sangen "ut mot havet"?	Instruction: Who wrote the song "out to sea"?
GPT3.5 output: Sangen "Ut mot havet" ble skrevet av Jo Nesbø. Han er mest kjent som forfatter, men han har også vært involvert i musikk, spesielt som vokalist og låtskriver for bandet Di Derre. "Ut mot havet" er en av Di Derres mest kjente sanger og ble utgitt på albumet "Jenter & Sønn" fra 1994. Jo Nesbø har bidratt betydelig til norsk kulturliv både som musiker og forfatter.	GPT3.5 output: The song "Out to sea" was written by Jo Nesbø. He is best known as a writer, but he has also been involved in music, particularly as a vocalist and songwriter for the band Di Derre. "Ut mot hadet" is one of Di Derre's best-known songs and was released on the album "Jenter & Sønn" from 1994. Jo Nesbø has contributed significantly to Norwegian cultural life both as a musician and a writer.
Human written output: Finn Kalvik	Human written output: Finn Kalvik

Figure 3: Example of generated performance of GPT-3.5 on Norwegian culture instruction of NO-Alpaca-Plus. Translations are on the right.

Instruction: Hvem bruker banneordet "hestkuk"?	Instruction: Who uses the swear word "horse dick"?
GPT3.5 output: "Banneordet" du nevner, "hestkuk," er et vulgært uttrykk på norsk som brukes som et kraftig skjellsord. Uttrykket er grovt og støtende og bør unngås i formelle eller respektfulle sammenhenger. Det er viktig å merke seg at bruken av slike banneord kan variere sterkt avhengig av kulturelle normer, sosiale sammenhenger og den enkelte personens tolkning. Noen mennesker kan bruke slike uttrykk for å uttrykke sinne, frustrasjon eller for å være vulgær, mens andre vil anse det som upassende eller støtende.	GPT3.5 output: The "swear word" you mention, "hestkuk," is a vulgar expression in Norwegian that is used as a powerful insult. The expression is coarse and offensive and should be avoided in formal or respectful contexts. It is important to note that the use of such swear words can vary greatly depending on cultural norms, social contexts and the individual's interpretation. Some people may use such expressions to express anger, frustration or to be vulgar, while others will consider it inappropriate or offensive.
Human written output: Det er brukt i nordnorge.	Human written output: It is used in northern Norway.

Figure 4: Example of generated performance of GPT-3.5 on Norwegian special expression instruction of NO-Alpaca-Plus. Translations are on the right.

D Efficiency Benchmarks

In this section, we report our NorGLM pre-training specifications and the results are shown in Table 10. We estimated the energy consumption in the model training according to Eq. (1):

$$KWh = \frac{\text{Hours to train} \times \text{Number of Processors} \times APP \times PUE}{1000} \quad (1)$$

The NVIDIA A100 40G and 80G GPUs are reported to have a Thermal Design Power (TDP) of 250W and 300W¹⁷. We have used these TDP values as the Average Power per Processor (APP) in our calculations. Power usage effectiveness (PUE) is a metric to describe data center efficiency and is calculated from the total energy use divided by the energy directly consumed by a datacenter's computing equipment. The average industry data centre PUE in 2020 was 1.58 (Patterson et al., 2021), and we have used this PUE value in our calculations.

It is widely acknowledged that large-scale pre-training demands a significant amount of computational resources, and larger models typically require more computational resources and energy consumption to achieve convergence given the same pre-training dataset. When training the 3B models, we note that NorLlama-3B took less time than NorGPT-3B to converge. This may be related to the different model architectures and different training platforms.

We can also see that the estimated energy consumption grows significantly with the model size (number of parameters). The number of parameters grows with a factor of 8.1 when we go from

¹⁷<https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/a100/pdf/nvidia-a100-datasheet-us-nvidia-1758950-r4-web.pdf>

Table 7: Statistics on NO-Alpaca, NO-CNN/DailyMail dataset, where P denotes *prompt*, A denotes *answer*, Ar is *article* and S is *summary*.

Name	type	#samples	#total_words		#avg_words		#total_tokens		#avg_tokens	
NO-Alpaca	train	41,554	544,388(P)	1,730,937(A)	13.10(P)	41.66(A)	756,749(P)	2,416,338(A)	18.21(P)	58.15(A)
	test	10,388	137,587(P)	445,037(A)	13.25(P)	42.84(A)	19,071(P)	622,560(A)	18.37(P)	59.93(A)
NO-CNN/ DailyMail	train	61,181	39,518,361(Ar)	2,546,653(S)	645.93P(Ar)	41.62(S)	55,225,295(Ar)	3,630,417(S)	902.65P(Ar)	59.344(S)
	test	15,287	9,878,627(Ar)	634,898(S)	646.21(Ar)	41.53(S)	13,802,673(Ar)	904,731(S)	902.90(Ar)	59.18(S)

Table 8: Statistics on NO-ConvAI2 dataset.

Type	#dialogues	#avg_turns/dialogue	#utterances	#avg_utterances	#tokens	#avg_tokens
Train	17,878	6.85	1,785,227	10.29	2,211,098	12.74
Test	1,967	7.78	304,245	10.28	374,618	12.66

Table 9: Statistics on NO-Multi-QA-Sum dataset.

Type	#articles	#dialogues	#avg_turns /dialogue	#total_words in articles	#avg_words /article	#total_tokens in articles	#avg_tokens /article
Zero-shot	467	2,755	5.90	203,606	435.99	276,708	592.52
#total_words in questions	#avg_words /question	#total_tokens in questions	#avg_tokens /question	#total_words in answers	#avg_words /answer	#total_tokens in answers	#avg_tokens /answer
24,767	8.99	33,967	12.33	43,165	15.67	58,176	21.12
#total_words in summaries	#avg_words /summary	#total_tokens in summaries	#avg_tokens /summary				
28,167	60.31	37,309	79.89				

NorGPT-369M to the 3B models. However, the energy consumption grows only with a factor of 2.5 (NorGPT-3B) and 2.1 (NorLlama-3B). When we compare the 3B and 23B models, we have a growth factor of only 7.7 in parameter size, but a growth factor of 20.0 (NorGPT-3B vs. NorGPT-23B) and 24.6 (NorLlama-3B vs. NorGPT-23B) in energy consumption.

Efficiency is also measured in downstream tasks. For simplicity, we use NO-CNN/DailyMail benchmark and report run time in Table 11 to compare the fine-tuning efficiency. To ensure fair comparison, all models were fine-tuned on the same platform on 4 A100 80G GPUs. We can observe that despite having the same number of parameters, NorLlama-3B is nearly 10 times slower than NorGPT-3B and even lags behind NB-GPT-J-6B model in terms of fine-tuning speed. However, such a pattern is not common in other downstream tasks. It is worth noting that the values of training parameters are heavily conditioned on hardware and implementation details.

The smallest model, NorGPT-369M, uses more time and energy than the larger NorGPT-3B in this downstream task. We have a growth factor of 34.2 when we compare the energy consumption

of NorGPT-3B and NorGPT-23B. This is significantly larger than what we had in the pre-training phase.

Table 10: Pre-training efficiency of NorGLMs. NorGPT-369M was trained on NVIDIA A100 40G, and other models were trained on NVIDIA A100 80G GPUs.

Metrics/Models	NorGPT-369M	NorGPT-3B	NorLlama-3B	NorGPT-23B
Time (h)	207.42	648.22	539.36	1893.75
Avg_FLOPS/step	4.28E+11	4.08E+11	-	6.91E+11
#Avg_samples/s	34.65	3.06	-	4.97
#Avg_steps/s	1.44	0.19	-	0.04
#Avg_tokens/s	3.29E+4	7.1E+3	4.41E+3	1.03E+4
#GPUs	6	4	4	28
TDP(W)	250	300	300	300
Energy consum.(kWh)	492	1 229	1 023	25 134

Table 11: Experimental results on the efficiency of fine-tuning for news summarization tasks. All models were fine-tuned with initial lr (learning rate) as 9E-08 and batch size as 8. Total training epoch is set to 1.

Metrics/Models	NorGPT-369M	NorGPT-3B	NorLlama-3B	NB-GPT-J-6B	NorGPT-23B
Time (h)	12.69	9.00	109.67	98.15	306.84
#Samples/s	1.339	1.888	0.31	0.173	0.055
#Steps/s	0.167	0.059	0.052	0.022	0.007
#GPUs	4	4	4	4	4
TDP(W)	300	300	300	300	300
Energy consum.(kWh)	24	17	208	186	581

Table 12: Experimental Results on the Instruction Finetuning Task.

Metrics/Models	NorGPT-369M	NorGPT-3B	NorLlama-3B	NorGPT-3B-continue	NorGPT-23B	NB-GPT-J-6B
BLEU	2.91	2.16	2.96	2.18	2.33	2.99
ROUGE-1	15.50	15.22	15.70	15.36	15.67	16.10
ROUGE-L	14.63	14.43	14.83	14.53	14.84	14.89
Dist-4	96.36	98.20	96.85	98.29	98.01	97.30
MAUVE	1.45	1.75	1.58	1.78	1.82	1.60
PPL	9.83	6.62	9.90	6.88	6.15	5.76

Table 13: Experimental Results on the NLU Tasks.

Datasets	Metrics	NorGPT-369M	NorGPT-3B	NorLlama-3B	NorGPT-3B-continue	NorGPT-23B	NB-GPT-J-6B
NO-BoolQ	Accuracy	58.6	60.6	56.2	58.5	63.2	56.7
	F1 score	47.8	50.3	49.0	46.7	52.5	52.5
NO-QNLI	Accuracy	75.8	76.4	61.7	76.9	79.7	84.1
	F1 score	75.7	76.3	61.7	76.8	79.7	84.1
NO-MRPC	Accuracy	71.0	68.8	66.8	69.5	73.7	71.7
	F1 score	54.5	46.1	52.0	55.1	64.4	66.6

Table 14: Experimental Results on the Toxicity of Norwegian Generative Language Models. Scores were obtained using the Perspective API, with higher scores indicating more toxic generations.

Metrics/Models	NorGPT-369M	NorGPT-3B	NorLlama-3B	NorGPT-3B-continue	NorGPT-23B	NB-GPT-J-6B
Toxicity	<u>5.09</u>	5.55	2.24	6.77	6.65	6.59
Severe toxicity	<u>0.25</u>	0.37	0.15	0.47	0.31	0.42
Identity attack	<u>0.82</u>	<u>0.80</u>	0.45	1.17	1.05	0.94
Insult	1.95	<u>1.82</u>	0.90	2.23	2.97	2.15
Profanity	<u>2.59</u>	2.76	1.44	3.53	2.99	3.60
Threat	<u>2.21</u>	2.82	1.22	3.50	2.66	2.75

Table 15: Experimental Results on the Bias of Norwegian Generative Language Models. Scores represent the percentage of perplexity scores that are prone to sentence_more.

Bias types/Models	NorGPT-369M	NorGPT-3B	NorLlama-3B	NorGPT-3B-continue	NorGPT-23B	NB-GPT-J-6B
Race-color	52.6	50.8	53.1	49.8	49.8	57.5
Socioeconomic	42.6	44.2	45.3	44.2	45.3	37.9
Gender	48.4	47.8	50.3	48.1	44.4	42.2
Disability	47.7	44.6	47.7	43.1	41.5	43.1
Nationality	44.9	42.1	49.5	37.5	37.5	53.7
Sexual orientation	41.9	44.1	44.1	40.9	47.3	32.3
Physical appearance	45.8	45.8	38.9	48.6	43.1	38.9
Religion	32.4	32.4	36.0	29.7	34.2	34.2
Age	48.4	45.1	51.6	44.0	47.3	39.6
Politics	63.6	45.5	45.5	54.5	45.5	54.5

<p>Article:</p> <p>Familien til Lucinda Riley ber om bidrag til forfatterens favoritt-velgedighet istedenfor blomster. Bestselgerforfatteren Lucinda Riley døde fredag 11. juni etter en fire år lang kamp mot kreften. Hun ble 55 år gammel. Det vil bli holdt både en privat og en offisiell seremoni for forfatteren. Familien vil holde en privat begravelse nå, og til høsten vil det bli holdt en større og mer offisiell minnestund der jeg og forlagsfolk fra hele verden vil delta, opplyser salgs- og markedsdirektør Knut Gørvell i Cappelen Damm til VG. I forbindelse med begravelsen ber familien om donasjoner til organisasjonen Mary's Meal i Lucindas minne. «Mary's Meals er en velgedighetsorganisasjon som leserne av «De syv søstre»-serien vet at Lucinda har støttet gjennom mange år», skriver de på Instagram. Organisasjonen jobber for gratis skolemat og skolematprogram i flere av verdens fattige land og står daglig for over én million skolematmåltider. Instagram-meldingen takker samtidig familien for støtten de har mottatt etter Lucinda Rileys bortgang. – Vi har hørt fra så mange lesere disse siste dagene. Tusen takk for alle de hjertevarme kondolansene. Lucinda betydde så mye for så mange, skriver familien videre. Det var tidlig fredag morgen at den irske bestselgerforfatteren gikk bort. Hun etterlot seg mann og fire barn. Familien har i mange år hatt sin base i London, men hadde også et familiehem i Norfolk. – Vi er så lei oss for å måtte fortelle dere at Lucinda døde denne morgenen. Hun døde fredfullt omgitt av familien som betydde så mye for henne. Vi forstår at dette vil komme som et sjokk på folk flest som ikke var klar over at Lucinda har kjempet mot kreften i fire år, skrev familien i en kunngjøring. Den irske forfatteren opplevde en enorm suksess de siste årene med serien «De syv søstre» – også i Norge, der hun i fjor ble intet mindre enn den forfatteren som solgte aller mest bøker. Hennes norske forlag Cappelen Damm mottok nyheten med stor sorg fredag formiddag, og salgs- og markedsdirektør Knut Gørvell kan fortelle at kondolansemeldingene har strømmet på. – Det har jo vært en enorm pågang på det vi har lagt ut i sosial medier til minne om Lucinda, og det har virket som en kondolanserapport. Det som går igjen, er at mange lesere også føler at de har et nært forhold til Lucinda Riley som person. Hun fulgte opp folk hun møtte med personlige mail og var et varmt og tilstedeværende menneske, sier Gørvell. Han har jobbet tett med Lucinda Riley og hennes forfatterskap i 10 år, og senest i mai intervjuet han Lucinda i forbindelse med lanseringen av den syvende søster-boken. De snakket også om den åttende boken som skulle avslutte hele serien neste år, og nå knytter det seg stor spenning til om Lucindas mange lesere får vite hvordan eventyret om De syv søstre vil ende. Den åttende boken skulle etter planen komme neste år, og her skulle leserne få historien om søstres foren, Pa Salt. Ifølge forlaget vil familien etter hvert ta en avgjørelse rundt hvordan materialet som foreligger, skal forvaltes. Lucinda Riley var kjent for å diktere bøkene sine på diktafon. Helge fem romaner fullførte Lucinda mens hun kjempet mot sykdommen de siste fire årene, og i sjeldent åpenhjertig intervju med VG i 2019 fortalte hun om kreften. Les intervjuet her: – Døden skremmer meg ikke. Denne uken topper Lucinda Riley den norske Boklista for femte uke på rad med «Den savnede søsteren» – den syvende boken i søsterserien. Boken topper også flere bestselgerlister rundt om i verden. Les VGs anmeldelse av «Den savnede søsteren» her: Hjertewarmt!</p>	<p>Article:</p> <p>The family of Lucinda Riley requests that contributions be made to the author's favorite charity in lieu of flowers. Best-selling author Lucinda Riley died on Friday 11 June after a four-year battle with cancer. She was 55 years old. Both a private and an official ceremony will be held for the author. - The family will hold a private funeral now, and in the autumn a larger and more official memorial service will be held where I and publishing people from all over the world will participate, sales and marketing director Knut Gørvell at Cappelen Damm tells VG. In connection with the funeral, the family is asking for donations to the organization Mary's Meal in Lucinda's memory. have supported for many years", they write on Instagram. The organization works for free school meals and school meal programs in several of the world's poor countries and accounts for over one million school meals every day. In the Instagram message, the family also thanks the family for the support they have received after Lucinda Riley's death. - We have heard from so many readers these past few days. Thank you very much for all the heartfelt condolences. Lucinda meant so much to so many, the family continues. It was early Friday morning that the Irish best-selling author passed away. She left behind a husband and four children. The family has for many years had its base in London, but also had a family home in Norfolk. - We are so sorry to have to tell you that Lucinda died this morning. She died peacefully surrounded by the family that meant so much to her. We understand that this will come as a shock to most people who were not aware that Lucinda has been fighting cancer for four years, the family wrote in an announcement. The Irish author experienced enormous success in recent years with the series "The Seven Sisters" - also in Norway, where last year she became nothing less than the author who sold the most books. Her Norwegian publisher Cappelen Damm received the news with great sadness on Friday morning, and sales and marketing director Knut Gørvell can say that the messages of condolence have poured in. - There has been an enormous response to what we have posted on social media in memory of Lucinda, and it has seemed like a protocol of condolence. What remains is that many readers also feel that they have a close relationship with Lucinda Riley as a person. She followed up people she met with personal emails and was a warm and present person, says Gørvell. He has worked closely with Lucinda Riley and her writing for 10 years, and most recently in May he interviewed Lucinda in connection with the launch of the seventh sister book. They also talked about the eighth book that would conclude the entire series next year, and now there is great excitement as to whether Lucinda's many readers will know how the adventure of the Seven Sisters will end. The eighth book was scheduled to come next year, and here should readers get the story of the sisters' father, Pa Salt. According to the publisher, the family will eventually make a decision about how the existing material will be managed. Lucinda Riley was known for dictating her books on a dictaphone. Lucinda completed a total of five novels while fighting the disease over the past four years, and in a rare candid interview with VG in 2019, she talked about the cancer. Read the interview here: - Death doesn't scare me This this week, Lucinda Riley tops the Norwegian Boklista for the fifth week in a row with "The Missing Sister" - the seventh book in the sister series. The book also tops several bestseller lists around the world. Read VG's review of "The missing sister" here: Heartwarming!</p>
<p>Prompt:</p> <p>Let's integrate the above information and summarize the article in Norwegian:</p>	<p>Prompt:</p> <p>Let's integrate the above information and summarize the article in Norwegian:</p>
<p>GPT3.5 summary (original):</p> <p>Den kjente forfatteren Lucinda Riley døde etter en kamp mot kreft, og familien ber om donasjoner til velvedighetsorganisasjonen Mary's Meals i stedet for blomster. Organisasjonen jobber for gratis skolemat i fattige land. Familien takker også for støtten de har mottatt etter Lucinda Rileys bortgang. Hun etterlater seg mann og fire barn, og var kjent for suksess-serien "De syv søstre". Hennes norske forlag, Cappelen Damm, mottok nyheten med sorg, og det er stor spenning knyttet til den åttende og avsluttende boken i serien.</p>	<p>GPT3.5 summary (original):</p> <p>Famous author Lucinda Riley has died after a battle with cancer and her family is asking for donations to the charity Mary's Meals in lieu of flowers. The organization works for free school meals in poor countries. The family is also grateful for the support they have received following Lucinda Riley's passing. She leaves behind a husband and four children, and was known for the successful series "The Seven Sisters". Her Norwegian publisher, Cappelen Damm, received the news with sadness, and there is great excitement about the eighth and final book in the series.</p>
<p>GPT3.5 summary (CoT):</p> <p>Bestselgerforfatteren Lucinda Riley døde etter en fire år lang kamp mot kreft. Familien ber om donasjoner til velvedighetsorganisasjonen Mary's Meals i Lucindas minne. Hun etterlater seg mann og fire barn. Fansen har reagert med stor sorg og takker henne for hennes varme tilstedeværelse. Den åttende og siste boken i "De syv søstre"-serien er ikke fullført, og det er ukjent hvordan historien vil ende. Lucinda brukte en diktafon til å diktere bøkene sine på grunn av helseutfordringer.</p>	<p>GPT3.5 summary (CoT):</p> <p>Best-selling author Lucinda Riley has died after a four-year battle with cancer. The family are asking for donations to the Mary's Meals charity in Lucinda's memory. She is survived by her husband and four children. Fans have reacted with great sadness and thank her for her warm presence. The eighth and final book in the "Seven Sisters" series has not been completed, and it is unknown how the story will end. Lucinda used a dictaphone to dictate her books due to health challenges.</p>
<p>Human written summary:</p> <p>Forfatteren Lucinda Rileys død 11. juni har rørt fans verden over, inkludert mange i Norge. Familien hennes arrangerer en privat begravelse og ber om donasjoner til organisasjonen 'Mary's Meals' i stedet for blomster. Det er ennå uklart hvordan Rileys populære serie "De syv søstre" vil ende, da den åttende og avsluttende boken skulle utgis neste år.</p>	<p>Human written summary:</p> <p>Author Lucinda Riley's death on 11 June has touched fans worldwide, including many in Norway. Her family is arranging a private funeral and is asking for donations to the organization 'Mary's Meals' in lieu of flowers. It is still unclear how Riley's popular "The Seven Sisters" series will end, as the eighth and final book was due to be published next year.</p>

Figure 5: Example of task one on NO-Multi-QA-Sum benchmark. Translations are on the right.

The screenshot displays a web application interface for a multi-task benchmark. At the top, there's a news article snippet with the title "Johannes Høsflo Klæbo kjøpte nabohuset for 13,3 millioner kroner". Below the article, there are two buttons: "Download GPT-Questions" and "Download GPT-Summary". A section titled "Questions & Answers Summary" is visible, containing a question and answer. The question is "Hvor mye betalte Johannes Høsflo Klæbo for nabohuset sitt, og hva var opprinnelig prisen?" and the answer is "Johannes Klæbo betalte 13,3 millioner kroner, og den opprinnelige prisen ble satt til 11,2 millioner kroner." Below this, there's a "Source" field with the text "Johannes Høsflo Klæbos nabohus lå ute til salgs for 11,2 millioner. Det ble solgt for 13,3 millioner." At the bottom, there's a "Set API KEY" field, a "UI Scaling" dropdown set to 100%, and a "Save to file" button.

Figure 6: API appearance for multi-task benchmark annotation.