Synthetic Instruction Generation for Low-Resource Nordic Languages: Viability and Limitations in LLM Instruction-Tuning

Mathias Stenlund[1], **Annika Simonsen**[1], **Lars Bungum**[2], **Jan Ebert**[3], **Jiangtao Wang**[3], **Oleg Filatov**[3], **Hemanadhan Myneni**[1], **Morris Riedel**[13], **Hafsteinn Einarsson**[1]

[1]University of Iceland, Sæmundargata 2, 102 Reykjavík, Iceland
[2]Norwegian University of Science and Technology, Høgskoleringen 1, 7034 Trondheim, Norway
[3]Forschungszentrum Jülich, Wilhelm-Johnen-Straße, 52428 Jülich, Germany
hem29@hi.is, ans72@hi.is, lars.bungum@ntnu.no, ja.ebert@fz-juelich.de,
jian.wang@fz-juelich.de, o.filatov@fz-juelich.de, myneni@hi.is,
morris@hi.is, hafsteinne@hi.is

Abstract

Pretrained large language models (LLMs) gain instruction-following abilities through instruction-tuning, a method which relies on datasets of instruction–response pairs. However, for low-resource languages, collecting human-authored instructions is costly, raising the question of whether synthetic instructions can substitute human-authored instructions for non-English languages. We compare instruction-tuning of a smaller pretrained LLM in four Nordic languages using (a) human-authored instructions paired with synthetic responses and (b) fully synthetic instruction–response pairs generated with a minimal-effort pipeline. Native-speaker evaluations show that models instruction-tuned on synthetic instructions perform on par with those trained on human-authored instructions for the largest Nordic languages, suggesting that minimal-effort synthetic instructions can serve as a practical alternative. In contrast, response quality deteriorates sharply for Icelandic, underscoring the limitations of current synthetic data generation pipelines when the LLM competence in the target language is weak. Overall, our results highlight that while synthetic instructions can enable cost-efficient instruction-tuning for the largest Nordic languages, they remain insufficient for Icelandic, clarifying when minimal-effort synthetic approaches suffice and when they fall short.

Keywords: instruction-tuning, synthetic data, low-resource languages, Nordic languages, multilingual NLP

1. Introduction

While instruction-tuning has proven crucial for aligning large language models (LLMs) with user intent (Wei et al., 2022; Ouyang et al., 2022; Sanh et al., 2022), the creation of high-quality human-authored instruction datasets remains costly and time-consuming (Gilardi et al., 2023; Liu et al., 2024c), limiting progress and widening the gap between high- and low-resource languages. For such languages, researchers often rely on translating existing English datasets, such as Alpaca (Taori et al., 2023), risking missing cultural subtleties (Hershcovich et al., 2022; Cao et al., 2024), or assembling small-scale human-authored collections of instructions (Liu et al., 2024b). In response, synthetic data generation has emerged as a practical and cost-efficient alternative (Nikolenko et al., 2021; Microsoft Research Team, 2025). However, its effectiveness relative to human-authored instructions remains underexplored, particularly for non-English languages. This raises the question: *Can synthetic instructions, generated with minimal human effort, substitute for human-authored ones in low-resource settings?*

We address this question by comparing instruction-tuning with human-authored instructions versus fully synthetic instructions across four Nordic languages of varying resource levels: Swedish, Danish, Norwegian Bokmål, and Icelandic. Using the same pretrained 7.8B base model, we fine-tune on (a) human-written instructions paired with synthetic responses, and (b) fully synthetic instruction-response pairs at two scales. We evaluate instruction-following capability using native-speaker preference rankings, where annotators assess both linguistic quality and instruction adherence. Our controlled comparison across languages with different resource levels suggests when synthetic instructions could potentially substitute for human-authored ones, and where current approaches fall short.

Specifically, we investigate three research questions: **RQ1:** Can fully synthetic instruction-response pairs achieve comparable instruction-following performance to human-authored instructions paired with synthetic responses when fine-tuning pretrained models for low-resource languages? **RQ2:** How does the performance of instruction-tuned models scale with increased synthetic data size? **RQ3:** What are the critical limitations of current synthetic data generation approaches across different resource levels?

Our main contributions are as follows:

- A direct comparison of human- vs. synthetic-instruction fine-tuning across varying re-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 10688–10698
11–16 May 2026.

source levels, examining when minimal-effort synthetic data suffices.

- Results suggesting that synthetic instructions can replace human-authored ones for the largest Nordic languages, but fall short for Icelandic, highlighting critical limitations.
- An empirical insight that response quality has a stronger impact than instruction quality.

2. Related Work

Recent research has put much focus on how to obtain the data required to effectively instruction-tune LLMs. Previous efforts range from human and hybrid approaches to large-scale synthetic data generation using existing LLMs.

2.1. Instruction-Tuning Dataset Creation

On the human side, datasets can be constructed through large-scale annotation efforts (Köpf et al., 2023; Singh et al., 2024) or consensual data collection via interactive chat-style platforms (Zheng et al., 2024; Zhao et al., 2024). Hybrid approaches, such as Zhou et al. (2023), combine high-quality web-sourced data with carefully curated human-written examples focusing on quality over quantity. In contrast, synthetic approaches rely on powerful LLMs to generate instruction-response pairs, sometimes starting from small sets of human-written seed instructions to iteratively expand datasets (Wang et al., 2023; Taori et al., 2023). Other methods reverse-engineer instructions from online documents for long-context generation (Köksal et al., 2024), employ taxonomy-guided generation for niche domains (Li et al., 2024a; Sudalairaj et al., 2024), or use *generator prompts* (Chen et al., 2024) that encourage diversity in generated data by having an LLM first produce a list of topics and then sample from it. Whereas the aforementioned methods vary in complexity and the initial need for at least some human labor, Xu et al. (2024b) show that it is possible to create large and diverse instruction-tuning datasets from scratch without a human-in-the-loop. This is done by leveraging already instruction-tuned models and their respective pre-query template to have them self-synthesize instructions.

2.2. Synthetic Data for Low-Resource Languages

Research focusing on synthetic data generation for low-resource languages is still scarce. However, some recent studies show promising results for non-English languages. For example, MURI (Köksal et al., 2025) generates instruction-response-pairs for hundreds of low-resource languages by reversing multilingual documents into synthetic instructions, M2Lingual (Maheshwary et al., 2025) creates large-scale multilingual, multi-turn synthetic conversations across 70 languages, and Pengpun et al. (2024) introduce a seed-free framework for Thai that sources relevant documents from Wikipedia as a basis for generating instructions covering various tasks.

3. Methodology

We conduct a controlled comparison of instruction-tuning approaches across four Nordic languages. Using the same pretrained base model (Section 3.1), we instruction-tune it on both human-authored instructions paired with synthetic responses and fully synthetic instruction-response pairs respectively (Section 3.2). These models are then evaluated based on native-speaker preference rankings (Section 4).

3.1. Pretrained Model

The pretrained model used for instruction-tuning in this work is a μP-parametrized (Yang et al., 2022) 7.8B parameter, decoder-only Transformer model (Vaswani et al., 2017; Radford et al., 2018) that largely follows the *Llama-3* architecture (Grattafiori et al., 2024). It was trained on a total of 2.3T tokens. The tokenizer is a SentencePiece-BPE tokenizer (Kudo and Richardson, 2018; Sennrich et al., 2016), optimized for uniform compression across languages. For pre-training data, a subset of the OpenGPT-X dataset (Brandizzi et al., 2025), including Germanic languages and code data, was used. This dataset was divided into domains classified into 6 levels of resource availability. A modified UniMax algorithm (Chung et al., 2023) was used to sample from these domains during training, with sub-datasets assigned between 12 and 2 epochs depending on their resource availability class, resulting in upsampling of low-resource sub-datasets.[1]

3.2. Data

All of the datasets used for instruction-tuning the pretrained model follow the standard instruction-

[1]Training details: A WSD learning rate schedule (Hu et al., 2024) was used with warmup covering 560B tokens, constant learning rate until 2.1T total tokens, then a decay phase over 210B tokens. μP enabled transfer of optimal hyperparameters from a smaller model. For HuggingFace compatibility, the μP base model was chosen to match the 7.8B model, allowing use of existing *Llama-3* code without modifications.

fine-tuning format consisting of instruction-response pairs, where each instruction corresponds with a target response. All datasets use synthetic responses generated by *DeepSeek-v3-0324* (685B parameters) (Liu et al., 2024a). We selected *DeepSeek-v3-0324* for response generation due to its open accessibility that allows us to use its output to improve our models. While other models like GPT-4 might produce higher-quality responses in our target languages, their terms of service restrict using outputs for model training, making them unsuitable for this purpose. Crucially, this design choice, using synthetic responses even for human-authored instructions, allows us to isolate the effect of the impact of the instruction quality.

3.3. Data: Human-Authored Instructions

We make use of the *TrustLLM Prompt Reformulation Dataset*[2] (Simonsen et al., 2026), a dataset consisting of native-speaker-authored prompts in several Germanic languages including Swedish, Danish, Norwegian Bokmål, and Icelandic. The prompts were collected through a prompt reformulation task in which each contributor was shown a randomly selected English prompt from the OpenAssistant (Köpf et al., 2023) dataset (OASST2) and were asked to reformulate the prompt in their native language, so that the prompt is culturally appropriate, while keeping the same task type. Alternatively, the prompt authors were given the option to produce their own original prompt without any reference to any pre-existing ones. In total, the dataset contains approximately 1k prompts per language, for which we generate synthetic responses using *DeepSeek-v3-0324* to create our instruction-tuning datasets. We divide the datasets into a train, validation, and evaluation split (see Table 1).

Dataset	Total	Train	Val.	Eval.
human_swe	995	795	100	100
human_dan	999	799	100	100
human_nob	1024	824	100	100
human_ice	994	794	100	100

Table 1: Dataset splits for human-authored instructions from the *TrustLLM Prompt Reformulation Dataset* paired with synthetically generated responses used for fine-tuning and evaluation.

[2] https://huggingface.co/datasets/AnnikaSimonsen/TrustLLM-reformulation-prompts

3.4. Data: Synthetic Instructions

Instruction generation using Magpie: In order to generate new and diverse instructions in our chosen Nordic languages, we make use of *Magpie* (Xu et al., 2024b) which offers a minimal-human-effort method of creating synthetic instruction-response datasets and only requires access to an already instruction-fine-tuned LLM, in our case *Llama-3.3-70B-Instruct*[3], and its pre-query template (see Figure 1). We make use of *Llama-3.3-70B-Instruct* for instruction generation as it represents a strong open-source multilingual model with documented performance across general Nordic language tasks on the EuroEval[4] benchmark (Smart, 2023)[5][6]. Additionally, its instruction-tuning makes it suitable for the *Magpie* approach which relies on instruction-tuned LLMs. The pre-query template, along with the post-query template, are pre-defined templates that ensure correct prompting of the LLM. The pre-query template directly precedes the user query marking the start of the user's turn in the conversation and the post-query template directly succeeds the query, marking the end of the user's turn. However, by only prompting the LLM with the pre-query template without any additional input, its autoregressive nature forces it to self-synthesize a "user" query on its own. The Magpie method leverages this behavior of instruction-tuned LLMs by repeatedly applying this step, allowing the LLM to generate surprisingly diverse instructions at a large scale. These instructions are then collected and can be used as the instruction component of instruction-tuning datasets.

```
<|start_header_id|>user<|end_header_id|>
```

Figure 1: Llama 3's predefined pre-query template that can be used to extract queries from the model.

Language adaptation: However, this strategy using the designated Llama model ends with instructions generated in English. For adaptation to other languages, Xu et al. (2024b) suggest providing a basic system prompt at inference time in the target language to incentivize the LLM to generate instructions in that language. We use translations of the default Vicuna (Chiang et al., 2023) system prompt: "A chat between a curious user

[3] https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct

[4] Previously ScandEval.

[5] https://euroeval.com/leaderboards/Multilingual/mainland-scandinavian/

[6] https://euroeval.com/leaderboards/Monolingual/icelandic/

```
Input:
<|begin_of_text|><|start_header_id|>system<|end_header_id|>
En konversation mellan en nyfiken användare och en AI-assistent.
Assistenten ger användbara, detaljerade och vänliga svar på användarens frågor.
<|eot_id|><|start_header_id|>user<|end_header_id|>

Output:
Hur skiljer sig däggdjur från blötdjur?
```

Figure 2: Full *Magpie* input sequence for Llama models containing a Swedish translation of the Vicuna system prompt: "A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions.". The system prompt is wrapped in Llama's predefined system prompt template (blue), followed by the pre-query template (red) triggering Llama to self-synthesize an instruction in Swedish that fits the context described in the system prompt. The Swedish output translates to: "How do mammals differ from molluscs?"

and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions." Using this system prompt in our target languages along with the pre-query template (see Figure 2), we then proceed to generate tens of thousands of instructions per language.

Quality filtering: After the instruction-generation step, we take an LLM-as-a-judge approach to automatically assess the quality of each instruction. To do this, we use the *Magpie* quality assessment framework which provides a collection of prompts for evaluating the quality of the generated instruction. Using a modified version of the original assessment prompts, we prompt *Llama-3.3-70B-Instruct* to evaluate the quality of each generated instruction, tagging it as 'very poor', 'poor', 'average', 'good', or 'excellent' quality immediately after generating each batch of instructions. To ensure we keep only the highest quality data, we only keep instructions labelled as being of either 'good' or 'excellent' quality, following the threshold approach described in the original paper (Xu et al., 2024b). We also conduct human evaluation on a small sample of Swedish instructions, which confirmed consistency with the model's quality labels. This threshold effectively balances quality, quantity, and diversity: lowering the threshold (e.g., including instructions labelled as 'average') would increase data volume but risk introducing low-quality examples, whereas higher thresholds would filter away a majority of instructions and might restrict instruction diversity. Despite this quality filtering, we observe many instructions still containing features unfit for instruction-tuning such as only asking about the LLM's willingness or capability to help the user, simply making vague statements without asking for anything, directly answering its own question, and the occasional nonsensical sample. To address these remaining issues, we design an additional quality assurance prompt specifically targeting such cases, which we apply as a second filtering stage[7] directly after the first one. Rather than offering multiple quality tags, this second prompt requires the LLM to make a binary decision on whether to keep or discard each instruction. Finally, following the standard *Magpie* procedure, we remove overly similar instructions by representing them in an embedding space and computing the minimum neighbor distance using the FAISS library (Douze et al., 2025). To embed the Swedish, Danish and Norwegian Bokmål instructions, we use the paraphrase-multilingual-MiniLM-L12-v2 (Reimers and Gurevych, 2019) sentence transformers model, as the embedding model originally used only supports English. For Icelandic, we use LaBSE (Feng et al., 2022) due to its wider coverage of lower-resource languages compared to the paraphrase-multilingual model.

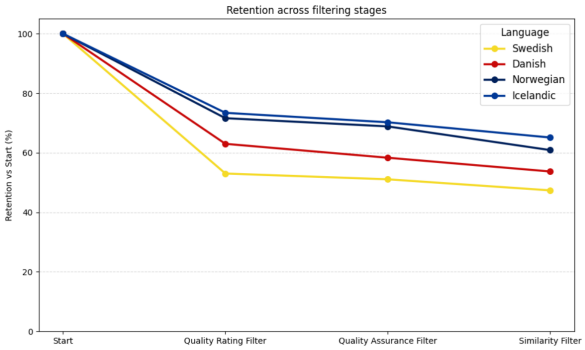

Figure 3: Data retention across filtering stages for the four synthetic datasets (Swedish, Danish, Norwegian, and Icelandic).

Figure 3 shows a breakdown of the percentage-wise data retention after each filtering step.

[7]We experimented with combining the quality assessment prompt and the quality assurance prompt into a single prompt, but saw many bad samples still slipping through.

Finally, to complete the synthetic instruction-tuning dataset, we generate responses to all instructions using *DeepSeek-v3-0324*[8].

3.5. Model Fine-Tuning

Using the pretrained model described in subsection 3.1 and the data outlined in subsection 3.2, we instruction-tune three separate models per language resulting in a total of twelve models. For each language, we fine-tune one model using human-created instructions paired with synthetic responses which we refer to as *human1k*. Additionally, we instruction-tune two models per language using exclusively synthetically generated instruction-response pairs using varying amounts of training samples. The first one, we instruction-tune using a randomly selected subset of synthetic instruction-response pairs from our generated dataset, matching the size of the data containing human-authored instructions for that language (*synth1k*). The second synthetic model for each language (*synth10k*) is instruction-tuned using 10K samples, with an additional 1K held out for validation during training. This allows us to assess whether scaling up synthetic data yields improvements (RQ2).

We instruction-tune our models using a learning rate of 1e-5, effective batch size of 64, and weight decay of 0.01. Every model is fine-tuned until convergence (defined as validation loss plateauing) on four AMD MI250X GPUs, as optimal training length may vary by language and dataset size.

4. Evaluation

We conduct our human evaluation on the 100 held-out instructions described in subsection 3.3 with simple preference rankings and one native-speaker annotator per language. For each language, we generate 3 responses per instruction using the three instruction-tuned language specific models for a total of 300 responses. We present each annotator with the instructions and three responses to each instruction a in random order, asking them to rank the responses from best to worst without knowing which model produced them. We ask them to pay specific attention to two main aspects. The first is natural language use, which concerns whether the response sounds fluent, uses correct grammar and spelling, and employs appropriate vocabulary. The second is instruction-following capability, which concerns how well the response fully addresses the given query while remaining on topic without introducing irrelevant information.

Annotation took a few hours per language, primarily due to the length of the responses being evaluated. The annotators volunteered their time for the task, and each received a short training session from the first author, although no formal calibration examples were provided. Ties were not handled according to a fixed protocol, but instead, annotators are asked to resolve them based on their own judgment.

We acknowledge that using a single annotator per language is a limitation, as it prevents measuring inter-annotator agreement and may introduce individual bias. This choice was made due to resource constraints, and future work should employ multiple annotators to strengthen validity.

5. Results

Table 2 presents the preference-ranking results from our native-speaker evaluation across all four languages and three instruction-tuning setups.

Lang.	Model	First	Second	Third	Avg. Rank
Swe	human1k	35	35	30	1.95
	synth1k	34	28	38	2.04
	synth10k	31	37	32	1.97
Dan	human1k	24	41	35	2.11
	synth1k	29	35	36	2.07
	synth10k	47	24	29	1.82
Nob	human1k	40	31	29	1.89
	synth1k	20	38	42	2.22
	synth10k	40	31	29	1.89
Ice	human1k	30	43	27	1.97
	synth1k	34	30	36	2.02
	synth10k	36	27	37	2.02

Table 2: Human evaluation results across four languages. Each row displays the number of times a model was ranked first, second, or third out of 100 prompts, along with the average rank (1=best, 3=worst).

6. Discussion

The annotation results reveal only moderate variation between models, with no single instruction-tuning configuration consistently outperforming the others. Swedish and Icelandic annotations are largely comparable across setups, indicating that neither data origin nor scale provides a clear advantage. In contrast, Danish and Norwegian show a mild but consistent preference for models trained on larger synthetic datasets (*synth10k*), suggesting that scaling synthetic data can yield limited yet measurable benefits when the overall quality

[8]We have made the fully synthetic datasets available on HuggingFace: https://huggingface.co/collections/matsten/magpie-datasets

of the synthetic data is sufficient. Expanding the dataset from 1K to 10K samples yields modest gains for Danish (*synth10k*: average rank 1.82 vs. *synth1k*: average rank 2.07) and Norwegian (*synth10k*: average rank 1.89 vs. *synth1k*: average rank 2.22) but has negligible impact for Swedish and Icelandic.

6.1. Higher-Resource Languages (Swedish, Danish, Norwegian)

Qualitative feedback from our annotators suggests that while the Swedish, Danish, and Norwegian responses are generally grammatically sound, lexically appropriate, and instruction-following, they occasionally include cross-lingual contamination, such as isolated English words or Danish in the Norwegian outputs. Factual hallucinations also remain frequent, particularly in responses to instructions requiring domain-specific or cultural knowledge. This aligns with broader findings that LLMs can struggle with factual accuracy despite linguistic fluency (Huang et al., 2025), underscoring the need for hallucination-mitigating strategies such as retrieval-augmented generation (RAG) (Lewis et al., 2020). The issue is especially pronounced in smaller models like the one used in this study, which typically encode less world knowledge compared to larger ones (Li et al., 2024b). Through knowledge distillation (KD) (Hinton et al., 2015), smaller models commonly (Xu et al., 2024a) mitigate this limitation by leveraging larger teacher models, allowing them to inherit world knowledge indirectly. However, currently, we lack a larger Nordic-language-specific teacher model suitable for such distillation, which constrains our ability to transfer domain-relevant knowledge into the smaller model. Furthermore, our observations are consistent with the findings of Zhou et al. (2023), who argue that most factual knowledge is acquired during pretraining and that only a small number of high-quality instruction examples are needed to achieve strong instruction-following capabilities. In light of this, the limited benefits observed from simply expanding the synthetic datasets suggest that additional samples do little to compensate for at least the current quality of the synthetic instructions using our setup. Rather than simply increasing dataset size, future improvements are more likely to come from improving synthetic data quality and coverage. Nevertheless, the comparable preference between responses from *human1k* and *synth10k* for Swedish, Danish, and Norwegian suggests that synthetic instructions serve as a viable and cost-effective alternative to human-authored ones for these languages.

6.2. Lower-Resource Language (Icelandic)

The qualitative feedback from annotators on the Icelandic responses paint a different picture indicating issues beyond just hallucinations such as widespread grammatical errors, inappropriate word choices, and overall nonsensical responses across all Icelandic models. Occasionally, responses are even produced in the wrong language. The particularly weak responses in Icelandic reveal a critical flaw in synthetic data generation in low-resource settings. Because synthetic approaches depend on multilingual LLMs, which are known to underperform in low-resource languages in general (Bang et al., 2023), the resulting data might be of lacking quality. The instruction-response generation pipeline in this study was designed for languages with robust LLM support and depends on LLMs for three steps: instruction generation, quality filtering, and response generation. When the LLMs underperform at any step, data quality deteriorates, and when they underperform across several, as is likely the case for Icelandic, the errors compound and yield synthetic data that fails to represent genuine Icelandic. In our setup, we used the same model, *Llama-3.3-70B-Instruct*, for both instruction-generation and quality assessment which might introduce self-preference-biases leading LLMs to give more favorable quality assessments to instructions that match their own style (Panickssery et al., 2024; Wataoka et al., 2024). In this sense, the LLM may generate low-quality instructions, and then, due to its limited grasp of Icelandic and possible self-bias, might incorrectly assess instructions as being of high-quality, creating a double negative effect that contributes to lower overall quality. Consequently, low-quality instructions are later fed to the response-generating LLM, further degrading the overall quality of the final dataset.

6.3. Influence of Response Quality

All datasets in this study, including those with human-authored instructions, rely on synthetic responses generated by *DeepSeek-v3-0324*. This shared dependency suggests that response quality may be the primary limiting factor, particularly for Icelandic. If instruction quality were the key determinant, the model fine-tuned on human-authored instructions (*human1k*) would likely have outperformed those trained on fully synthetic data in Icelandic. Instead, the similarly poor performance across all Icelandic models suggests that even high-quality instructions cannot compensate for responses of lower quality. This finding further supports the use of synthetic instruction generation in the dataset creation process, as instruction

quality appears to have a relatively minor impact compared to the quality of the responses themselves.

6.4. Future Work

Future work should focus on improving the quality and linguistic robustness of the synthetic data generation process, as well as exploring alternative options for generation, particularly for low-resource languages such as Icelandic. One alternative that might work better in low-resource settings is to introduce more human engagement into the synthetic generation pipeline, for instance, through seed-instruction methods such as *Self-Instruct* by Wang et al. (2023), where small sets of high-quality human-written prompts guide the automatic creation of additional samples. By presenting an LLM with a high-quality seed instruction, this could encourage the LLM to produce a higher quality instruction in turn that is linguistically appropriate. Another key improvement involves using models that demonstrate stronger competence in the target languages. Although the most capable models for Icelandic often come with restrictive licenses that prohibit using their outputs for training, they could still be leveraged during the quality-filtering stage to better identify and remove low-quality samples. Importantly, such filtering could be applied not only to instruction generation, but also to the response generation phase of the pipeline, improving overall dataset quality. To further enhance the linguistic accuracy of Icelandic responses, grammar correction tools could be applied after generation and filtering to ensure basic grammatical validity before instruction-tuning. Additionally, adapting the generation pipeline to incorporate language-specific models, such as NorMistral-7B-Warm[9] for Norwegian, might yield improvements in linguistic and cultural appropriateness compared to multilingual models. Future work could also take inspiration from Karan and Du (2025), who introduce a sampling algorithm at inference time that elicits high-quality data directly from pretrained models, potentially establishing a less restrictive method for creating synthetic instructions that does not rely on already instruction-tuned LLMs. Finally, future work should also aim to strengthen the evaluation methodology. This includes using multi-annotator setups to assess inter-annotator agreement and reduce individual bias, and developing error taxonomies to better capture specific failure types beyond simple preference rankings. Human annotations could also be complemented with assessments from strong LLMs to add a dimension of scalability to the evaluation (Zheng et al., 2023), however, their capabilities of evaluating responses in low-resource languages first need to be compared to human assessment to ensure agreement.

[9] https://huggingface.co/norallm/normistral-7b-warm

7. Conclusion

Our findings across the four Nordic languages offer insight into the viability of using synthetic instructions for instruction-tuning. We examine whether synthetic instructions could replace human-authored ones (RQ1), how performance scales with data size (RQ2), and what limitations arise (RQ3). For higher-resource Nordic languages such as Swedish, Danish, and Norwegian, synthetic instructions perform on par with human-authored data, suggesting that minimal-effort synthetic generation can serve as a practical substitute. Scaling to 10K samples yields moderate gains for Danish and Norwegian but minimal improvements for Swedish and Icelandic, indicating limited benefits from upscaling. For Icelandic, overall poor responses across all models point to fundamental weaknesses in current synthetic data pipelines when LLM capabilities are limited. While these findings should be interpreted cautiously given our single-annotator setup and reliance on one method for generating synthetic instructions, the consistent cross-lingual patterns reinforce their validity. Overall, the results highlight that synthetic instruction-tuning-data, as created in this study, have the potential to enable cost-efficient instruction-tuning in higher-resource Nordic languages, while also clarifying its limits. Our work outlines when minimal-effort synthetic approaches suffice and when they fall short, offering practical guidance for future research on efficient instruction-tuning in Nordic languages.

8. Limitations

Working with smaller models has been crucial for this study, as it enables a more exhaustive fine-tuning setup within realistic computational limits. However, this choice constrains the broader applicability of our findings, since smaller models may not fully capture the potential of large-scale instruction-tuning. Likewise, each language was evaluated by a single annotator, introducing potential individual bias and limiting the generalizability of the results. Further limitations concern the evaluation procedure itself. The prompt responses were often lengthy, making the annotation process more time-consuming, less engaging, and less extensive than intended. Moreover, ties between model outputs were not permitted during annotation. Annotators occasionally reported difficulty distinguishing between responses of simi-

larly good or poor quality, leading to forced rankings even when none of the options were clearly preferable. Allowing for neutral or "tie" judgments as implemented in frameworks such as Chatbot Arena (Chiang et al., 2024) could provide a more nuanced and realistic assessment of model performance, especially in cases where differences are marginal or qualitative judgments are ambiguous. In addition, our evaluation focused on preference rankings rather than fine-grained error analysis, leaving open questions about specific failure modes such as hallucinations, factual errors, or grammatical inconsistencies. Finally, the study relies on a single synthetic data generation method and a single response model, which may not represent the full range of available synthetic data approaches. As such, the results should be interpreted as indicative rather than definitive.

9. Ethical Statement

From an ethical standpoint, all human annotations were conducted with informed consent by native speakers affiliated with academic institutions who volunteered their time. Annotators remain anonymous, and no personally identifiable information was collected beyond language background verification. While synthetic data generation methods can reduce the costs of building instruction-tuning datasets and thus broaden access to language technology, they also risk perpetuating biases or inaccuracies present in the generating LLMs. The hallucinations and factual inconsistencies observed in this study underscore the importance of rigorous validation and transparency when deploying synthetic data for downstream tasks.

10. Acknowledgements

This work was supported by the European Union's Horizon 2020 research and innovation program under grant agreement No 101135671 (TrustLLM). The authors gratefully acknowledge the Gauss Centre for Supercomputing e.V. (https://www.gauss-centre.eu/) for funding this project by providing computing time on the GCS Supercomputer JUWELS at the Jülich Supercomputing Centre (JSC). We acknowledge the Icelandic LUMI Consortium, Iceland for awarding this project access to the LUMI supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CSC (Finland) and the LUMI consortium through the University of Iceland, Iceland, LUMI UoI Regular Access (Project_465002139). We also thank the reviewers for their insightful comments and constructive feedback.

11. Bibliographical References

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.

Nicolo' Brandizzi, Hammam Abdelwahab, Anirban Bhowmick, Lennard Helmer, Benny Jörg Stein, Pavel Denisov, Qasid Saleem, Michael Fromm, Mehdi Ali, Richard Rutmann, Farzad Naderi, Mohamad Saif Agy, Alexander Schwirjow, Fabian Küch, Luzian Hahn, Malte Ostendorff, Pedro Ortiz Suarez, Georg Rehm, Dennis Wegener, Nicolas Flores-Herr, Joachim Köhler, and Johannes Leveling. 2025. Data processing for the opengpt-x model family.

Yong Cao, Yova Kementchedjhieva, Ruixiang Cui, Antonia Karamolegkou, Li Zhou, Megan Dare, Lucia Donatelli, and Daniel Hershcovich. 2024. Cultural adaptation of recipes. *Transactions of the Association for Computational Linguistics*, 12:80–99.

Jiuhai Chen, Rifaa Qadri, Yuxin Wen, Neel Jain, John Kirchenbauer, Tianyi Zhou, and Tom Goldstein. 2024. Genqa: Generating millions of instructions from a handful of prompts. *arXiv preprint arXiv:2406.10323*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*.

Hyung Won Chung, Xavier Garcia, Adam Roberts, Yi Tay, Orhan Firat, Sharan Narang, and Noah

Constant. 2023. Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining. In *The Eleventh International Conference on Learning Representations*.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2025. The faiss library. *IEEE Transactions on Big Data*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

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

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. 2024. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Aayush Karan and Yilun Du. 2025. Reasoning with sampling: Your base model is smarter than you think.

Abdullatif Köksal, Timo Schick, Anna Korhonen, and Hinrich Schuetze. 2024. LongForm: Effective instruction tuning with reverse instructions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7056–7078, Miami, Florida, USA. Association for Computational Linguistics.

Andreas Köpf, Yannic Kilcher, Dimitri Von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. 2023. Openassistant conversations-democratizing large language model alignment. *Advances in neural information processing systems*, 36:47669–47681.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Abdullatif Köksal, Marion Thaler, Ayyoob Imani, Ahmet Üstün, Anna Korhonen, and Hinrich Schütze. 2025. Muri: High-quality instruction tuning datasets for low-resource languages via reverse instructions. *Transactions of the Association for Computational Linguistics*, 13:1032–1055.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Haoran Li, Qingxiu Dong, Zhengyang Tang, Chaojun Wang, Xingxing Zhang, Haoyang Huang, Shaohan Huang, Xiaolong Huang, Zeqiang Huang, Dongdong Zhang, Yuxian Gu, Xin Cheng, Xun Wang, Si-Qing Chen, Li Dong, Wei Lu, Zhifang Sui, Benyou Wang, Wai Lam, and Furu Wei. 2024a. Synthetic data (almost) from scratch: Generalized instruction tuning for language models.

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen.

2024b. The dawn after the dark: An empirical study on factuality hallucination in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10879–10899, Bangkok, Thailand. Association for Computational Linguistics.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024a. Deepseek-v3 technical report.

Peng Liu, Lemei Zhang, Terje Farup, Even W. Lauvrak, Jon Espen Ingvaldsen, Simen Eide, Jon Atle Gulla, and Zhirong Yang. 2024b. 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.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, and Andrew M. Dai. 2024c. Best practices and lessons learned on synthetic data. In *First Conference on Language Modeling*.

Rishabh Maheshwary, Vikas Yadav, Hoang H Nguyen, Khyati Mahajan, and Sathwik Tejaswi Madhusudhan. 2025. M2Lingual: Enhancing multilingual, multi-turn instruction alignment in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9676–9713, Albuquerque, New Mexico. Association for Computational Linguistics.

Microsoft Research Team. 2025. Synthllm: Breaking the ai data wall with scalable synthetic data. Accessed: 2025-08-25.

Sergey I Nikolenko et al. 2021. *Synthetic data for deep learning*, volume 174. Springer.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Arjun Panickssery, Samuel Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems*, 37:68772–68802.

Parinthapat Pengpun, Can Udomcharoenchaikit, Weerayut Buaphet, and Peerat Limkonchotiwat. 2024. Seed-free synthetic data generation framework for instruction-tuning LLMs: A case study in Thai. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 445–464, Bangkok, Thailand. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Annika Simonsen, Mathias Stenlund, Lars Bungum, Marc Daníel Skipstað Volhardt, and Hafsteinn Einarsson. 2026. Reformulate and create, don't translate: Creating natural prompts for underserved languages. In *Proceedings of the 2026 Language Resources and Evaluation Conference (LREC 2026)*, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Chien, Sebastian Ruder, Surya Guthikonda, Emad Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567, Bangkok, Thailand. Association for Computational Linguistics.

Dan Saattrup Smart. 2023. ScandEval: A Benchmark for Scandinavian Natural Language Processing. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 185–201.

Shivchander Sudalairaj, Abhishek Bhandwaldar, Aldo Pareja, Kai Xu, David D. Cox, and Akash Srivastava. 2024. Lab: Large-scale alignment for chatbots.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. 2024. Self-preference bias in llm-as-a-judge. *arXiv preprint arXiv:2410.21819*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. 2024a. A survey on knowledge distillation of large language models. *arXiv preprint arXiv:2402.13116*.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. 2024b. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. *ArXiv*, abs/2406.08464.

Greg Yang, Edward J. Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. 2022. Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. Wildchat: 1m chatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. 2024. LMSYS-chat-1m: A large-scale real-world LLM conversation dataset. In *The Twelfth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021.