Rethinking Low-Resource MT: The Surprising Effectiveness of Fine-Tuned Multilingual Models in the LLM Age

Barbara Scalvini
University of the Faroe Islands
barbaras@setur.fo

Iben Nyholm Debess
University of the Faroe Islands
ibennd@setur.fo

Annika Simonsen
University of Iceland
ans72@hi.is

Hafsteinn Einarsson
University of Iceland
hafsteinne@hi.is

Abstract

This study challenges the current paradigm shift in machine translation, where large language models (LLMs) are gaining prominence over traditional neural machine translation models. We focus on English-to-Faroese translation. We compare the performance of fine-tuned multilingual models, LLMs (GPT-SW3, Llama 3.1), and closed-source models (Claude 3.5, GPT-4). Our findings show that a finetuned NLLB model outperforms most LLMs, including larger models, in both automatic and human evaluations. We also demonstrate the effectiveness of using LLM-generated synthetic data for finetuning. While closed-source models like Claude 3.5 perform best overall, the competitive performance of smaller, finetuned models suggests a nuanced approach to low-resource machine translation. Our results highlight the potential of specialized multilingual models and the importance of language-specific knowledge. We discuss implications for resource allocation in low-resource settings and suggest future directions, including targeted data creation and comprehensive evaluation methods.

1 Introduction

The recent rise of LLMs has introduced new possibilities in machine translation (Lyu et al., 2024, 2023). LLMs demonstrated impressive performance across various language pairs, often through the use of in-context learning (Brown et al., 2020). These new opportunities often come at a price in terms of computational resources: LLMs have massive requirements in terms of pretraining data and high-end hardware. Hardware requirements can sometimes be mitigated by using closed-source LLM APIs (e.g., OpenAI API). However, this approach introduces issues related to transparency and license limitations.

These limitations and high requirements disproportionately affect low-resource languages and communities. For such languages, lack of resources can often extend beyond data scarcity and effectively imply lack of computational infrastructure and expertise, rendering the use of APIs offered by tech giants the only available option. This is the case for Faroese, an Insular Scandinavian language and official language of the Faroe Islands.

Neural machine translation (NMT) models are less demanding in terms of computational resources. However, due to their more limited reasoning capabilities compared to LLMs, they often underperform in low-resource settings. Nonetheless, there are potential strategies to leverage the linguistic knowledge of an LLM in conjunction with lightweight MT models to optimize performance while minimizing resource requirements. One such approach is to use LLMs to augment parallel datasets, allowing a lighter MT model to be trained on this synthetic data (Yang and Nicolai, 2023).

In NLP, efficiency encompasses various factors like data requirements, model size, training costs, and performance metrics. This paper focuses on the relationship between model performance and size, a crucial consideration for real-world applications. We explore different approaches to English-to-Faroese machine translation, investigating how various techniques balance translation quality with model compactness. Our research aims to shed light on the trade-offs between performance and model size in this specific language pair. We will compare the following approaches, in the context of English to Faroese MT:

- Using LLMs in a few-shot learning setting.
- Fine-tuning LLMs for translation (English-

Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), pages 609–621

to-Faroese).

- Using a multilingual NMT out of the box.
- Fine-tuning a multilingual model on English-Faroese parallel data.
- Fine-tuning a multilingual model on English-to-Faroese parallel data and LLM-generated synthetic parallel data.

These strategies will be compared based on automatic and human evaluation.We will be comparing the following open-source LLMs: Llama 3.1, (Meta) (Dubey et al., 2024)in its 8B version, and GPT-SW3, a generative model for the Nordic languages, primarily Swedish, (Ekgren et al., 2022, 2024), in its 1.3, 6.7 and 40B version.

Their performance will be compared to closed-source models such as Claude 3.5 Sonnet (Anthropic, 2024) by Anthropic, GPT-4 Turbo (OpenAI et al., 2024) and GPT-4o (OpenAI, 2024) by OpenAI. We compare the LLMs with No Language Left Behind (NLLB)(Team, 2024), an open-source NMT multilingual model covering, among other under-resourced languages, Faroese. All new models produced via fine-tuning in this paper are now publicly available.[1]

2 Background and related work

2.1 LLMs for translation

The emergence of LLMs has challenged the dominance of sequence-to-sequence transformer-based models in the field of machine translation (MT) (Lyu et al., 2024; Hendy et al., 2023; Robinson et al., 2023). LLMs like initially observed for GPT-3 can perform translations with minimal input through in-context learning (ICL), significantly reducing the data requirements typically needed for the training process. This ability to achieve state-of-the-art results with minimal data has highlighted the potential of LLMs as a promising solution for low-resource translation. A few studies have investigated methods to enhance LLMs' MT capabilities in low-resource settings, employing techniques such as layer adaptation and fine-tuning (Tran et al., 2024), retrieval-augmented prompting (Merx et al., 2024), integration with rule-based systems (Coleman et al., 2024), and synthetic parallel data generation with an LLM (Yang and Nicolai, 2023). Additionally, LLMs have demonstrated remarkable performance as evaluators of translation quality, achieving near-human accuracy, although these results have been primarily studied in high-resource languages (Karpinska and Iyyer, 2023; Fernandes et al., 2023; Huang et al., 2024; Kocmi and Federmann, 2023). However, the effectiveness of LLMs in low-resource contexts, such as Faroese, remains relatively underexplored. Some studies suggest that LLM-driven translation may be less competitive for low-resource languages (Robinson et al., 2023), when compared to their higher resource counterparts.

2.2 Machine Translation for Faroese

In recent years, a few notable efforts have focused on improving coverage for Faroese in machine translation (MT). A key initiative was the creation of Sprotin's parallel corpus (Mikkelsen, 2021), which includes around 100,000 short human-translated English-Faroese sentences. This corpus supported Faroese's integration into Microsoft Translator and an Icelandic Machine Translation platform called Vélþýðing, by the Icelandic company Miðeind. The rise of multilingual MT models has led to initiatives like Google's MADLAD 400 (Kudugunta et al., 2023) and Meta's No Language Left Behind (NLLB) (Team, 2024), targeting low-resource languages such as Faroese. Since July 2024, Faroese has also been included in Google Translate (Bapna et al., 2022). The linguistic proximity of Faroese to its higher-resource relatives, the Scandinavian languages, makes it an ideal candidate for transfer learning (Snæbjarnarson et al., 2023). GPT-SW3, an LLM trained on English and Scandinavian languages, has demonstrated significant potential for understanding Faroese (Scalvini and Debess, 2024). Likewise, GPT-4 has shown promising results in Faroese sentiment analysis (Debess et al., 2024) and Faroese-to-English translation (Simonsen and Einarsson, 2024).

[1] https://huggingface.co/barbaroo/llama3.1_translate_8B, https://huggingface.co/barbaroo/gptsw3_translate_1.3B, https://huggingface.co/barbaroo/gptsw3_translate_6.7B, https://huggingface.co/barbaroo/nllb_200_1.3B_en_fo, https://huggingface.co/barbaroo/nllb_200_600M_en_fo

3 Methods

3.1 Experiments

In this study, we evaluate machine translation performance for English into low-resource Faroese of various models: 5 LLM models (GPT-SW3, Llama 3.1, GPT-4 Turbo, GPT-4o, Claude 3.5 Sonnet) and one multilingual MT model covering Faroese in its pre-training phase, NLLB. We chose NLLB as representative of multilingual MT because it demonstrated the highest potential in earlier studies (Simonsen, 2024). Since the goal of this paper is to analyze which settings are best for open-source MT in a low-resource scenario, we mostly preferred smaller, less computationally costly versions of the models. We utilize NLLB in its 600M and 1.3B parameters, and fine-tune LLMs that have sizes below 10B parameters, as these would be the ones most likely to be fine-tuned and deployed on common, commercial hardware. In order to investigate different modalities to exploit LLM language capabilities in machine translation, we fine-tune the MT model, NLLB, on LLM generated parallel sentences, in addition to the available human made corpus. This approach is presented as an alternative to either directly deploying the LLM in a few-shot manner, or instruct fine-tuning it directly for the desired translation direction. We evaluate these models both automatically and by human evaluation, for which we build an openly available evaluation platform online[2]. The performance of these open-source models is also benchmarked against that of three of the most popular closed-source models (GPT-4 Turbo, GPT-4o and Claude 3.5 Sonnet), for comparison.

3.2 Datasets

Faroese, as a low-resource language, lacks substantial parallel datasets for machine translation. The most comprehensive resource is the Sprotin corpus (Mikkelsen, 2021), though it may miss Faroese-specific cultural elements since it was translated from English. Recent studies have explored using LLMs to generate synthetic parallel datasets, like the `fo_en_synthetic` [3] dataset (Scalvini and Debess, 2024), created through back-translation with GPT-SW3, containing 70,000 sentences from the BLARK corpus (Simonsen et al., 2022).

The inclusion of Faroese in Meta's No Language Left Behind (NLLB) initiative (Team, 2024) enabled the language's integration into the FLORES-200 benchmark for machine translation. Currently, FLORES-200 is the only available evaluation benchmark for Faroese translation, making it our choice for the automatic comparison of model performance. While FLORES-200 is a well-established benchmark in the field, it has known limitations, such as its domain composition and a narrow representation of cultural elements, given that it was originally translated from English (Simonsen and Einarsson, 2024). To address this, we manually compiled a small dataset of 200 English sentences for human evaluation. The dataset consists of 68 sentences sourced from documents produced by the University of the Faroe Islands (Strategic Plan 2025-2030), 56 from the webpage of the Nordic Council [4] and 92 sentences from international news outlets such as BBC, CNN, and Al Jazeera. The dataset is publicly available on Hugging Face, together with all synthetic translations produced in the context of this paper.[5] All sentences were guaranteed to be created within a specific recent time period, ensuring that none of the data had been used in the training of any models included in the study. The inclusion of sentences from Faroese and Nordic-related contexts aimed to better represent Faroese-specific cultural elements, which are typically underrepresented in datasets despite being highly relevant to the end users of Faroese machine translation products. For example, using sentences from locally relevant contexts included concepts and named entities that actually have a Faroese translation, as they are Faroese or Nordic by origin (e.g. the local institution 'Statistics Faroe Islands' - *Hagstova Føroya*). This is opposed to many concepts or entities in sentences from international sources, where the translation of such can be difficult due to the entities not having a direct Faroese translation, as they are often irrelevant to Faroese society (e.g. the concept of a 'US Governor', which has no Faroese equivalent). These foreign concepts make evaluation more complex. Furthermore, using locally or regionally sourced data together with internationally sourced data enables evaluating con-

[2] https://github.com/Haffi112/error-span-labelling

[3] https://huggingface.co/datasets/barbaroo/fo_en_synthetic

[4] https://www.norden.org/en

[5] https://huggingface.co/datasets/barbaroo/news_en_fo

tent for real-use Faroese scenarios.

3.3 Prompting LLMs for English to Faroese translation

All LLMs used in this study were prompted in a few-shot fashion. Each translation query consisted of a prompt presenting the model with 5 randomly selected examples of English to Faroese translation. Examples were selected from a small subset of the Sprotin corpus comprising of 25 manually selected parallel sentences. These sentences were selected by a Faroese linguist based on the following criteria: 1) the meaning of the sentence is fully preserved in its translation 2) all words have unambiguous meaning, 3) they present simple syntax (declarative sentences or interrogative sentences, excluding subordinate clauses or sentences), 4) there are no typographical and inflectional errors. Two different prompting strategies were used for open-source (GPT-SW3 and Llama) and closed-source models (GPT-4o, Claude 3.5 Sonnet). These distinction was made in order to provide each model with an optimal prompting format.

3.4 Open-source models

We used the base versions of the Llama 3.1 and GPT-SW3 models. To facilitate model comprehension, we framed the prompt as a language completion task. Each example was structured as follows:
The English sentence {english_sentence} is translated to Faroese as {faroese_sentence}

The query followed the same format but omitted the Faroese translation:
The English sentence {english_sentence} is translated to Faroese as

This approach minimized the number of failed translation outputs.

3.5 Closed-source models

Closed-source models (GPT-4 Turbo, GPT-4o and Claude 3.5 Sonnet) were prompted via their respective APIs. The prompt structure was then adapted to the API format, with a system prompt containing the few-shot examples and the instructions of the task (*When I give you a sentence in English, you translate it into Faroese. Only answer with a translation.*) and a translation prompt containing the translation query.

3.6 Fine-tuning of models for English to Faroese translation

All open-source models in this study, except GPT-SW3 40B, were also fine-tuned for English-to-Faroese translation. For the LLMs, fine-tuning was conducted for three epochs with early stopping, using the Sprotin corpus. We adopted the Alpaca prompting format for both Llama and GPT-SW3, which includes an instruction ("Translate this sentence from English to Faroese"), an input (the English sentence), and an output (the Faroese sentence). Training was performed in 8-bit precision to reduce computational resource requirements. Two versions of NLLB, with 0.6 billion and 1.3 billion parameters, were also fine-tuned for English-to-Faroese translation. The training was carried out in two settings: (1) using only the Sprotin corpus and (2) using a combination of the Sprotin corpus and the `fo_en_synthetic` dataset. These different settings were chosen to demonstrate the potential benefits of incorporating LLM-generated parallel sentences to improve translation quality. The complete training configuration can be found in our GitHub repository.[6]

3.7 Evaluation

Automatic evaluation is performed using the metrics BLEU, ChrF and BERTscore. We do not use more advanced neural metrics, as these are not currently available for Faroese.

For human evaluation, we adopted the recently developed Error Span Annotation (ESA) metric proposed by Kocmi et al. (2024). ESA combines elements from two established methods: the overall scoring approach of Direct Assessment (DA) and the error severity span markings from Multidimensional Quality Metrics (MQM). In their study, Kocmi et al. (2024) compared ESA to MQM and DA across several MT systems. Their findings demonstrated that ESA offers a more cost-effective and time-efficient alternative to MQM without compromising evaluation quality. The ESA operates with a dual error system, which is less complex to the annotator compared to the multiple error categories and subcategories of MQM.

We created an annotation user interface based on the task description in Kocmi et al. (2024). Figure 1 shows an example from the interface. The

[6] `https://github.com/barbaroo/finetune_translation`

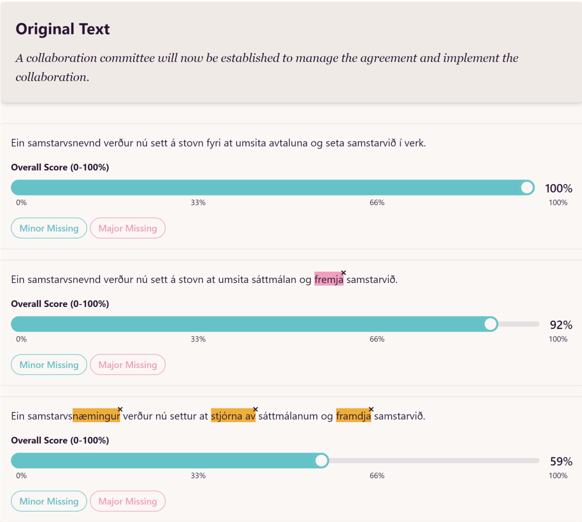

Figure 1: The annotation interface. Annotators were presented with the original text along with four translations (three shown here). The annotators mark any segment and are prompted to label it minor (pink) or major (orange). The annotators assign an overall score (1-100) to each translation (blue). For each translation, the annotators can optionally mark missing elements as major or minor.

annotation process was the following: the annotator is presented with the original English sentence along with four Faroese translations. The annotator then marks all the errors in the Faroese translations and to each error assigns one of the two severity levels: **major** or **minor**. Additionally, there is a label for omission errors, called *minor/major missing*. After marking the errors, the annotators assign each translation with an overall score from 0 to 100. The overall score reflects translation quality in a broad sense, covering adequacy, fluency and comprehension.

3.8 Annotator Guidelines

For the human evaluation, we had two human annotators, both linguists and native speakers of Faroese. The annotators developed the annotation guidelines together, using the original guidelines from Kocmi et al. (2024) as a starting point and adjusting it to fit the specific task. The full guidelines are shown below.

Approach

Annotators identified and marked error spans in translations, assigning severity levels (major or minor) to each. They then provided an independent, holistic overall score that could consider factors beyond marked errors, such as fluency. Major errors include significant meaning changes, mistranslations, foreign words, untranslated named entities, and synthetic words (constructed well-structured and sensible words, that are however not recognized in human language use). Minor errors encompass slight meaning alterations, style issues, grammatical mistakes, spelling errors, and punctuation problems.

Other

- Grammatical errors spanning over multiple words are marked as a single error
- If the source sentence has an error, annotators consider this original error in their evaluation of the translations
- If the source sentence is erroneous to an extent where translation output is completely off, all 4 sentences are given 0% and no errors are marked.

Scoring

This method provides two scores: an ESA overall score (0-100) and the $\text{ESA}_{\text{spans}}$ (number and severity of errors). The $\text{ESA}_{\text{spans}}$ is calculated as segment score, SEG, SCORE = $-1 * N_{\text{MINOR}}$ $-4.8 * N_{\text{MAJOR}}$, as suggested by Kocmi et al. (2024). As the evaluations of overall score and errors are meant to be performed independently, these scores can be treated separately.

4 Results

4.1 Automatic evaluation

The results for automatic evaluation on the FLORES-200 benchmark for all models can be found in Table 1. For all three different scores, we can see how closed-source Claude yields the best results. However, NLLB 1.3B, in its fine-tuned version (Sprotin + `fo_en_synthetic`) scores second overall and first among open-source models. A representation of the CHRF score with respect to model size, for all models under 10B is shown in Figure 2. As we can see the top left corner, representing the best performing models with respect to their hardware requirements, is dominated by fine-tuned NLLB models. NLLB 1.3 fine-tuned with the Sprotin corpus alone does yield a better performance with respect to fine-tuned LLMs, and with respect to GPT-4o as well. The performance is anyway sensibly increased (1

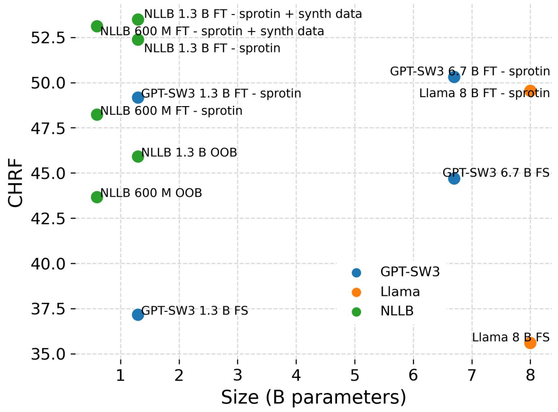

Figure 2: Translation performance for all models (with fewer than 10 billion parameters) in the automatic evaluation, quantified by the CHRF score. The performance is plotted against the model size, expressed in billions of parameters.

ChrF point and 3 BLEU points) by adding LLM-generated synthetic data. Llama 3.1 8 B does yield the worst performance in a few-shot setting, demonstrating however great potential for improvement after fine-tuning, beating out of the box NLLB and GPT-SW3 1.3 B.

4.2 Human evaluation

When picking models for human evaluation, we picked the best models from each category according to the automatic evaluation (see Table 1). We picked the following four models: GPT-SW3 6.7B - Sprotin, Llama 3.1 8B - Sprotin, NLLB 1.3B - Sprotin + `fo_en_synthetic` and we also picked the best performing closed-source model, Claude 3.5 Sonnet. The results from the human evaluation, in terms of ESA - overall quality score - and ESA_{spans} scores, are displayed in Table 2. Claude 3.5 Sonnet shows the best performance of the four, with NLLB getting the best results for the open-source models. GPT-SW3, despite the smaller size, does beat Llama 3.1 in both human and automatic evaluation, showing that family language specific knowledge is an advantage for models of comparable sizes.

Figure 3 shows the average ESA score for the two annotators separately, showing that the two annotators agree on how the models should be ranked in terms of translation quality. The ESA_{spans} score can be deconstructed into different error types, as shown in Figure 4. Here we see the

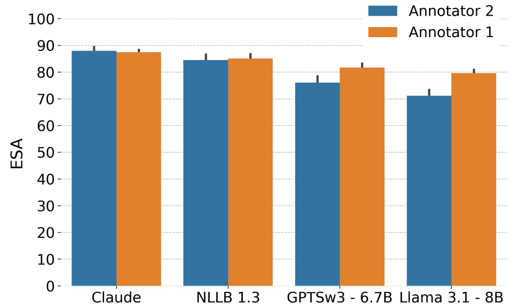

Figure 3: Average overall quality score (ESA) per model, assigned by the two annotators. "Average overall quality score (ESA) per model, as assigned by the two annotators. All models in the plot are shown in their fine-tuned versions (GPT-SW3 6.7B - Sprotin, Llama 3.1 8B - Sprotin, NLLB 1.3B - Sprotin + `fo_en_synthetic`), except for Claude."

two best performing models, Claude and NLLB 1.3, have comparable number of minor and major errors, with Claude performing better when it comes to preserving content (missing content, major and minor). NLLB and Claude do display comparable performance across the metrics. While the ESA scores assigned to the two models are statistically distinct ($p = 0.017$, as calculated by Mann-Whitney U test), the same cannot be said for the ESA_{spans} scores ($p = 0.465$). GPT-SW3 6.7B seems to struggle the most with preserving content due to the greatest number of missing content errors. However, it is performing largely better than Llama 3.1 8B when it comes to number of errors.

4.2.1 Annotator agreement

Figure 5 shows the distribution of ESA scores from both annotators. While mostly overlapping, the distributions have different variances (Levene test, $p = 1.34 \times 10^{-28}$). Krippendorff's alpha indicates moderate to strong agreement for absolute ESA (0.58) and ESA_{spans} (0.67) scores. We also converted scores to rankings for each translation query, assigning equal ranks for tied scores. Kendall's W analysis of these rankings showed moderate to strong inter-annotator agreement (ESA: 0.514, ESA_{spans}: 0.518), further supporting the reliability of our annotations.

4.3 Common Error Patterns

From a qualitative perspective the annotators report some common error patterns that emerged in

Model	BLEU	CHRF	BERTScore (f1)
GPT-SW3 40 B	0.173 ± 0.005	48.3 ± 0.4	0.9472 ± 0.0005
GPT-SW3 6.7 B	0.119 ± 0.004	44.7 ± 0.4	0.9373 ± 0.0005
GPT-SW3 1.3 B	0.084 ± 0.004	37.1 ± 0.4	0.9279 ± 0.0006
GPT-SW3 6.7 B* - Sprotin	0.183 ± 0.006	50.3 ± 0.4	0.951 ± 0.001
GPT-SW3 1.3 B*- Sprotin	0.179 ± 0.005	49.2 ± 0.4	0.947 ± 0.001
Llama 3.1 8 B	0.062 ± 0.003	35.6 ± 0.3	0.9311 ± 0.0005
Llama 3.1 8 B* - Sprotin	0.175 ± 0.005	49.5 ± 0.4	0.9487 ± 0.0005
NLLB 600 M	0.129 ± 0.005	43.7 ± 0.4	0.9428 ± 0.0005
NLLB 600 M* - Sprotin	0.171 ± 0.005	48.2 ± 0.5	0.9458 ± 0.0006
NLLB 600 M* - Sprotin + `fo_en_synthetic`	0.200 ± 0.006	53.1 ± 0.4	0.9524 ± 0.0005
NLLB 1.3 B	0.161 ± 0.005	45.9 ± 0.4	0.9459 ± 0.0005
NLLB 1.3 B* - Sprotin	0.209 ± 0.006	52.4 ± 0.4	0.9516 ± 0.0005
NLLB 1.3 B* - Sprotin + `fo_en_synthetic`	$\mathbf{0.212 \pm 0.006}$	$\mathbf{53.5 \pm 0.4}$	$\mathbf{0.9530 \pm 0.0005}$
GPT-4 Turbo	0.193 ± 0.006	52.7 ± 0.4	0.9518 ± 0.0005
GPT-4o	0.191 ± 0.005	51.7 ± 0.4	0.9509 ± 0.0005
Claude 3.5 Sonnet	$\mathbf{0.226 \pm 0.006}$	$\mathbf{55.3 \pm 0.4}$	$\mathbf{0.9546 \pm 0.0005}$

Table 1: Model performance metrics, calculated over the FLORES-200 dataset. All scores pertaining to LLMs were obtained in a few shot setting, with the exception of those that were fine-tuned (*). The mention of *Sprotin* and `fo_en_synthetic` indicate which datasets was the model fine-tuned on. The error term represents the standard error of the mean for 1012 translations.

Model	ESA	$\text{ESA}_{\text{spans}}$	N (ESA = 0)
Claude 3.5 Sonnet	$\mathbf{87.7 \pm 0.5}$	$\mathbf{-2.3 \pm 0.1}$	**0**
NLLB 1.3B - Sprotin + `fo_en_synthetic`	84.8 ± 0.7	-2.3 ± 0.1	3
Llama 3.1 8B - Sprotin	75.3 ± 0.6	-6.3 ± 0.2	0.5*
GPT-SW3 6.7B - Sprotin	78.8 ± 0.7	-4.6 ± 0.2	2

Table 2: Comparison of Models based on human evaluation. The table portrays ESA and $\text{ESA}_{\text{spans}}$ scores, and number of failed translations, expressed in terms of number of translations that received a 0 as ESA score, N (ESA = 0). The * indicates that only one of the two annotators assigned a 0 score, therefore we do not assign N = 1, but N = 0.5. The error term represents the standard error of the mean for 215 translations.

the annotation process. Taking a closer look at linguistic errors, morphological errors seem more common with inflectional errors in adjectives being prevalent. Errors in translating named entities were also frequent, as the models struggle with identifying the correct entities in Faroese. An interesting observation is the occurrences of a type of error, where the models make up new words, that are structurally well-formed for Faroese and semantically appropriate to various extents, but are complete neologisms and not recognised in natural Faroese language use, spoken or written. These words were typically compound words, like the example of "artificial intelligence" being translated into *telduheimsniðgóðskapur*. Finally, all models tend to translate word-for-word, which leads to literal translations of idioms and fixed phrases. Error patterns like these can suggest effective focus areas when creating parallel data for improving the models.

5 Discussion

Our study on English to Faroese machine translation reveals several important findings that provide new insights into the relative strengths of different approaches to low-resource language translation, including large language models and specialized multilingual models. Surprisingly, the fine-tuned NLLB model outperformed most LLMs, including GPT-4 and GPT-SW3 40B, in both au-

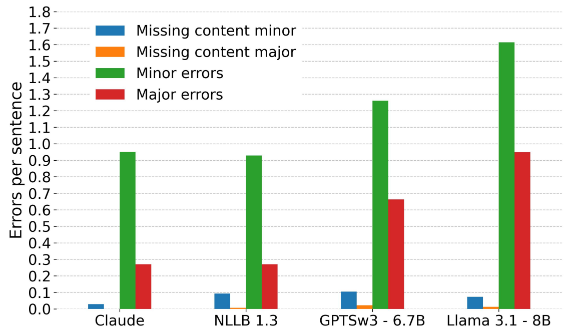

Figure 4: Average error type per model, as defined by the ESA framework: minor error, major error, minor missing content and major missing content. All models in the plot are shown in their fine-tuned versions (GPT-SW3 6.7B - Sprotin, Llama 3.1 8B - Sprotin, NLLB 1.3B - Sprotin + `fo_en_synthetic`), except for Claude.

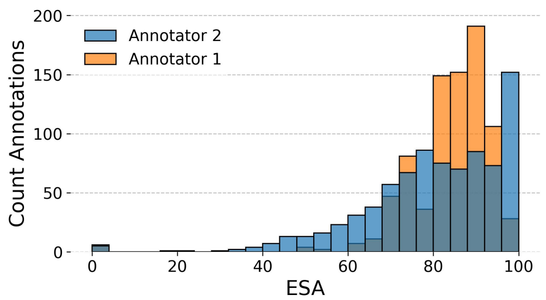

Figure 5: Distribution of overall quality scores (ESA) given by the annotators.

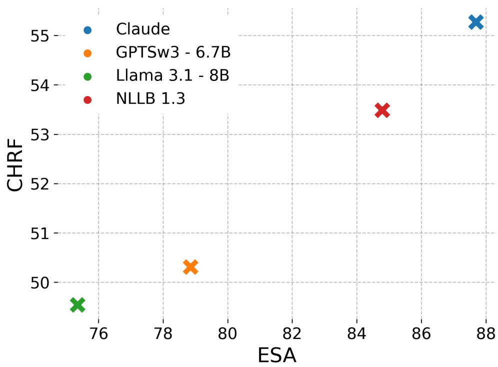

Figure 6: Scatterplot of CHRF scores versus overall quality scores (ESA). All models in the plot are shown in their fine-tuned versions (GPT-SW3 6.7B - Sprotin, Llama 3.1 8B - Sprotin, NLLB 1.3B - Sprotin + `fo_en_synthetic`), except for Claude.

tomatic and human evaluations. This suggests that specialized multilingual models, when fine-tuned appropriately, can be highly effective, often achieving comparable or even superior performance to larger LLMs for specific language pairs. The success of NLLB highlights the importance of domain-specific training and more compact, efficient models, which can be especially valuable in low-resource settings where computational power may be limited. Furthermore, the performance of GPT-SW3, despite its smaller size compared to Llama 3.1, underscores the critical role of language-specific knowledge in translation tasks. These findings have significant implications for resource allocation and model selection in low-resource language translation.

While automatic and human evaluations generally aligned on model rankings, there were key differences in perceived quality. This reveals the limitations of relying solely on automatic metrics, especially for low-resource languages. Human evaluations showed that while Claude 3.5 Sonnet and NLLB 1.3B had similar error counts, Claude performed better in content preservation and received a higher overall ESA score, suggesting that evaluators may prioritize factors like fluency and naturalness beyond just error quantity.

The improvement in NLLB's performance when fine-tuned on both the Sprotin corpus and LLM-generated synthetic data (`fo_en_synthetic`) highlights the potential of leveraging LLMs to augment training data for low-resource languages (Yang and Nicolai, 2023). This strategy could enhance translation quality in resource-constrained settings. However, despite these gains, all evaluated models still exhibit significant errors, falling short of human-quality translation, which calls for further research. These findings suggest that fine-tuning smaller, specialized models may offer a more cost-effective solution than relying on large LLMs, and that targeted data creation, informed by common error patterns, could further boost performance. Additionally, the discrepancies between automatic and human evaluations emphasize the need for more nuanced evaluation methods for low-resource language translation.

Future work should focus on iterative improvement techniques such as back-translation, exploring methods to distill knowledge from larger LLMs to smaller, more deployable models, and creating more diverse and representative parallel datasets for low-resource languages like Faroese.

6 Conclusion

Our study on English to Faroese machine translation offers a nuanced perspective on the effectiveness of different approaches to low-resource language pairs, highlighting how fine-tuned models like NLLB can rival or outperform larger LLMs for low-resource languages. This suggests that focusing on fine-tuning smaller models and creating targeted synthetic datasets may be more effective and resource-efficient. Despite improvements, all models still fall short of human-quality translation, emphasizing the need for further research on error patterns, data augmentation, and better evaluation methods. Advancing low-resource translation likely calls for a tailored combination of specialized models with effective data augmentation strategies.

7 Limitations

One possible limitation of our study is that we did not consider how much Faroese text these models were exposed to during pre-training. We excluded this information because, for some models, it is not publicly available: we do not have access to closed-source training data, and detailed documentation on the data sources for Llama 3.1 had not been released as of December 2024. GPT-SW3 does not officially cover Faroese, although it is possible that some Faroese text was misclassified as Icelandic within the training data. Conversely, NLLB was trained on approximately 2.8 million Faroese–English bitext sentences (Schwenk et al., 2020; Fan et al., 2020), which are now available on Opus (Tiedemann, 2012). The amount of Faroese these models have seen certainly influences their final performance; however, quantifying this exposure is difficult for most LLMs, making such comparisons challenging.

References

Anthropic. 2024. Claude 3.5 sonnet. `https://www.anthropic.com`. Proprietary software, closed-source.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, Theresa Breiner, Vera Axelrod, Jason Riesa, Yuan Cao, Mia Xu Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, Apurva Shah, Yanping Huang, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. 2022. Building machine translation systems for the next thousand languages. *arXiv preprint arXiv:2205.03983*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jared Coleman, Bhaskar Krishnamachari, Ruben Rosales, and Khalil Iskarous. 2024. LLM-assisted rule based machine translation for low/no-resource languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 67–87, Mexico City, Mexico. Association for Computational Linguistics.

Iben Nyholm Debess, Annika Simonsen, and Hafsteinn Einarsson. 2024. Good or bad news? Exploring GPT-4 for sentiment analysis for Faroese on a public news corpora. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7814–7824, Torino, Italia. ELRA and ICCL.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade

Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang,

Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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, Marseille, France. European Language Resources Association.

Ariel Ekgren, Amaru Cuba Gyllensten, Felix Stollenwerk, Joey Öhman, Tim Isbister, Evangelia Gogoulou, Fredrik Carlsson, Judit Casademont, and Magnus Sahlgren. 2024. GPT-SW3: An autoregressive language model for the Scandinavian languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7886–7900, Torino, Italia. ELRA and ICCL.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation. *CoRR*, abs/2010.11125.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1066–1083, Singapore. Association for Computational Linguistics.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Xu Huang, Zhirui Zhang, Xiang Geng, Yichao Du, Jiajun Chen, and Shujian Huang. 2024. Lost in the source language: How large language models evaluate the quality of machine translation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 3546–3562, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. In *Proceedings of the Eighth Conference on Machine Translation*, pages 419–451, Singapore. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024. Error span annotation: A balanced approach for human evaluation of machine translation. *arXiv preprint arXiv:2406.11580*.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: A multilingual and document-level large audited dataset. *arXiv preprint arXiv:2309.04662*.

Chenyang Lyu, Zefeng Du, Jitao Xu, Yitao Duan, Minghao Wu, Teresa Lynn, Alham Fikri Aji, Derek F. Wong, and Longyue Wang. 2024. A paradigm shift: The future of machine translation lies with large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1339–1352, Torino, Italia. ELRA and ICCL.

Chenyang Lyu, Jitao Xu, and Longyue Wang. 2023. New trends in machine translation using large language models: Case examples with chatgpt. *arXiv preprint arXiv:2305.01181*.

Raphaël Merx, Aso Mahmudi, Katrina Langford, Leo Alberto de Araujo, and Ekaterina Vylomova. 2024. Low-resource machine translation through retrieval-augmented LLM prompting: A study on the Mambai language. In *Proceedings of the 2nd*

Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia (EURALI) @ LREC-COLING 2024, pages 1–11, Torino, Italia. ELRA and ICCL.

Jonhard Mikkelsen. 2021. Sprotin sentences. `https://raw.githubusercontent.com/Sprotin/translations/main/sentences_en-fo.strict.csv`. Accessed: October 13, 2023.

OpenAI. 2024. Gpt-4o. `https://www.openai.com`. Proprietary software, closed-source.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.

Barbara Scalvini and Iben Nyholm Debess. 2024. Evaluating the potential of language-family-specific generative models for low-resource data augmentation: A Faroese case study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6496–6503, Torino, Italia. ELRA and ICCL.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, and Armand Joulin. 2020. Ccmatrix: Mining billions of high-quality parallel sentences on the web. *arXiv preprint arXiv:1911.04944*.

Annika Simonsen. 2024. Improving Machine Translation for Faroese using ChatGPT-Generated Parallel

Data. Master's thesis, University of Iceland, Reykjavík.

Annika Simonsen and Hafsteinn Einarsson. 2024. A Human Perspective on GPT-4 Translations: Analysing Faroese to English News and Blog Text Translations. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 24–36, Sheffield, UK. European Association for Machine Translation (EAMT).

Annika Simonsen, Sandra Saxov Lamhauge, Iben Nyholm Debess, and Peter Juel Henrichsen. 2022. Creating a basic language resource kit for Faroese. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4637–4643, Marseille, France. European Language Resources Association.

Vésteinn Snæbjarnarson, Annika Simonsen, Goran Glavaš, and Ivan Vulić. 2023. Transfer to a low-resource language via close relatives: The case study on Faroese. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 728–737, Tórshavn, Faroe Islands. University of Tartu Library.

NLLB Team. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630:841–846.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Khanh-Tung Tran, Barry O'Sullivan, and Hoang Nguyen. 2024. Irish-based large language model with extreme low-resource settings in machine translation. In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 193–202, Bangkok, Thailand. Association for Computational Linguistics.

Wayne Yang and Garrett Nicolai. 2023. Neural machine translation data generation and augmentation using chatgpt. *arXiv preprint arXiv:2307.05779*.