FoQA: A Faroese Question-Answering Dataset

Annika Simonsen
University of Iceland
ans72@hi.is

Dan Saattrup Nielsen
Alexandra Institute
dan.nielsen@alexandra.dk

Hafsteinn Einarsson
University of Iceland
hafsteinne@hi.is

Abstract

We present FoQA, a Faroese extractive question-answering (QA) dataset with 2,000 samples, created using a semi-automated approach combining Large Language Models (LLMs) and human validation. The dataset was generated from Faroese Wikipedia articles using GPT-4-turbo for initial QA generation, followed by question rephrasing to increase complexity and native speaker validation to ensure quality. We provide baseline performance metrics for FoQA across multiple models, including LLMs and BERT, demonstrating its effectiveness in evaluating Faroese QA performance. The dataset is released in three versions: a validated set of 2,000 samples, a complete set of all 10,001 generated samples, and a set of 2,395 rejected samples for error analysis.

1 Introduction

Recent NLP advancements, driven by the transformer architecture (Vaswani et al., 2017), have led to large-scale models that excel in understanding (Devlin et al., 2018) and generating (Brown et al., 2020) human language. While many models are "massively multilingual" (Conneau et al., 2019; He et al., 2021a; Brown et al., 2020) they often perform better on high-resource languages, leaving low-resource languages undersupported. Furthermore, low-resource languages typically have limited access to native speakers who can serve as data annotators, making it difficult to create high-quality evaluation datasets. High-quality evaluation datasets are crucial for assessing and improving models for these languages, helping to measure performance and guide language technology development.

Extractive QA datasets (Srivastava and Memon, 2024) are especially useful, as they simulate real-world applications like retrieval-augmented generation (Gao et al., 2023). Creating these datasets traditionally requires substantial human effort, often involving multiple annotators for question generation and answer validation. Standardising methods for creating these datasets can significantly advance technology for low-resource languages.

Our research addresses these challenges and makes the following key contributions:

- **An efficient, single-annotator methodology for producing high-quality extractive QA datasets** using a semi-automated approach that significantly reduces the human effort required for dataset creation, provided as an open-source Python codebase[1]
- **The first extractive QA dataset for Faroese** using this method[2].

2 Related Work

QA systems are divided into extractive and abstractive types (Fan et al., 2019). This work focuses on extractive QA, also known as reading comprehension, where text passages are paired with questions, and answers are directly extracted from the text. A well-known example of an extractive QA dataset is the Stanford Question Answering Dataset (SQuAD), which includes over 100,000 QA pairs from Wikipedia articles (Rajpurkar et al., 2016). In the case of Icelandic, a language closely related to Faroese, several QA datasets have been developed. Snæbjarnarson and Einarsson (2022a) introduced a cross-lingual open-domain QA system using machine-translated data, and the Natural Questions in Icelandic, an extractive QA

[1] https://github.com/alexandrainst/foqa
[2] https://huggingface.co/datasets/alexandrainst/foqa

Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL 2025), pages 48–57
March 2, 2025

dataset, which demonstrates approaches applicable to other low-resource languages such as Faroese (Snæbjarnarson and Einarsson, 2022b). Similarly, Skarphedinsson et al. (2023) developed a method to gamify QA dataset creation. However, both approaches relied heavily on human question generation, which bottlenecked the dataset creation process.

At the time of writing, few benchmark datasets exist for Faroese. Snæbjarnarson et al. (2023) introduced named entity recognition[3] and semantic text similarity datasets[4]. The FLORES-200 dataset (Costa-Jussà et al., 2022) is another significant contribution to Faroese benchmarks, being a multilingual parallel corpus covering over 200 languages, including Faroese. Additionally, Nielsen (2023) introduced ScaLA-Fo, a linguistic acceptability dataset for Faroese. Despite these resources, a dedicated Faroese QA dataset is still lacking, which this work aims to address.

3 Methodology

3.1 Generation of Tentative Dataset

The process of generating an extractive question-answering dataset begins with several key components: a vocabulary, a text corpus, a generative model, and specialised functions for generating questions and answers and for question reformulation. Using these components, we create a tentative dataset through a two-step process. First, we apply a QA generation function to our text corpus to create initial QA pairs. Then, we refine these pairs by rewriting the questions while keeping the answers unchanged.

The QA generation function operates by utilising our generative model to create multiple questions for each document in the corpus, along with corresponding answers that must be found verbatim within the source document. To ensure consistency and maintainability, we implement strict formatting requirements for the model's output. Specifically, we require the model to generate responses in a structured JSON format, following the approach described by Willard and Louf (2023). Each output must be a dictionary containing a "results" key, which maps to a list of dictionaries. These inner dictionaries must contain exactly two keys: "question" and "answer." Any outputs that deviate from this precise format are automatically filtered out of the dataset.

A significant limitation of the initially generated questions lies in their close adherence to the source documents' original phrasing. These questions often merely restructure existing statements from the text into interrogative forms, diminishing their effectiveness as evaluation tools. Consider a document containing the statement "Jane Smith is an executive and her bike is red." The initial generation might produce "What colour is Jane Smith's bike?"—a question that could be answered through simple text matching algorithms, requiring minimal linguistic or reasoning capabilities. To address this limitation, we employ a question reformation process that introduces additional complexity. By transforming the previous example to "What colour is the executive's bicycle?", we create questions that demand more sophisticated comprehension abilities, including synonym recognition and multi-hop reasoning in this example. This reformulation process is implemented through our question-rewriting function, which utilises the generative model to produce modified questions.

We release our code base implementing this generation process open-source[5].

3.2 Manual Filtering of Tentative Dataset

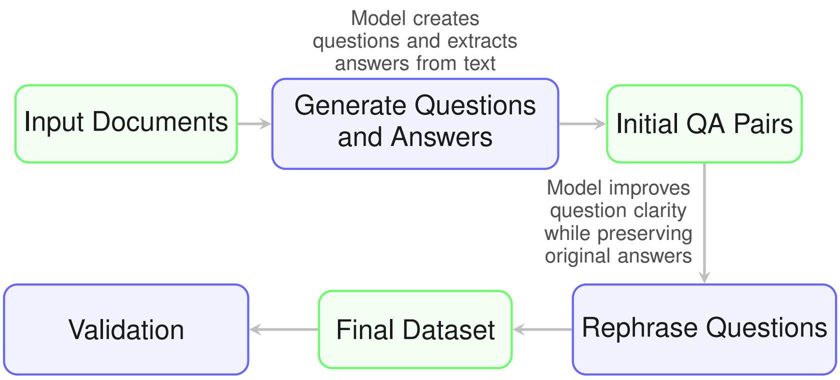
To ensure high-quality dataset creation, we implemented a human validation phase using a custom annotation interface built with Gradio (Abid et al., 2019), a Python library for web-based interfaces. The tool presents annotators with each generated question and its answer, offering three classification options: CORRECT (both question and answer are grammatically and contextually appropriate), INCORRECT (question is grammatically incorrect or contextually inappropriate), and INCORRECT ANSWER (answer is irrelevant, inaccurate, or grammatically incorrect). An annotator reviews each QA pair and assigns the appropriate classification, ensuring linguistic quality and filtering out inadequate samples. The annotation tool is available open-source[6].

[3] https://huggingface.co/datasets/vesteinn/sosialurin-faroese-ner

[4] https://huggingface.co/datasets/vesteinn/faroese-sts

[5] https://github.com/alexandrainst/foqa

[6] https://huggingface.co/spaces/saattrupdan/foqa-validation




```
Example Output Format:
{
``results'': [
  {
    ``question'': ``What role did
Alan
    Turing play in the development
    of computer science?'',
    ``answer'': ``Alan Turing laid
the
    theoretical foundation for
    computer science through his
    work on computability.''
  },
  {
    ``question'': ``What was the
    Turing machine?'',
    ``answer'': ``A mathematical
model
    of computation that manipulates
    symbols on a tape.''
  }
]
}
```


Figure 1: Overview of the QA dataset generation pipeline. The system processes input documents to generate initial QA pairs, followed by a question rewriting phase that improves clarity while maintaining the original answers. All outputs follow a structured JSON format to ensure consistency. Note that while the outputs are in Faroese, the example shown in this figure uses an English example for illustrative purposes.

3.3 Annotation Guidelines

This section outlines the complete annotation guidelines for evaluating QA pairs in Faroese. The annotator will follow a three-tier classification system when analysing each sample.

Tier 1: Grammatical Assessment The annotator should begin by evaluating the grammatical correctness of both the question and answer in Faroese. The annotator must check for proper agreement between subjects and verbs, correct case marking on nouns and pronouns, standard Faroese word order, accurate spelling and so on. If any grammatical errors are found in the question, the annotator should mark the entire sample as INCORRECT. If the grammatical errors only appear in the answer, the sample should be marked as INCORRECT ANSWER.

Tier 2: Semantic and Contextual Assessment After confirming grammatical correctness, the annotator should examine the relationship between the question and answer, as well as their connection to the source text. The answer must directly address the question being asked. Additionally, the annotator should ensure the answer demonstrates logical consistency within its context. If any issues with relevance, accuracy, or consistency are found, the sample should be marked as INCORRECT ANSWER.

Tier 3: Final Classification When a sample passes both the grammatical and semantic assessments, the annotator should mark it as CORRECT. The annotator will also be asked to correct a selection of questions marked as INCORRECT. When performing these corrections, the annotator should focus only on samples where the question itself contains errors, not the answer. This is crucial because modifying answers would compromise the extractive nature of the QA task, as answers should appear verbatim in the source text.

Quality Control Process All samples marked as CORRECT can undergo a secondary review by another annotator who is also a native Faroese speaker. This second annotator will apply the same three-tier evaluation process described above.

4 Faroese Setup

We applied our methodology (Section 3) to the Faroese Wikipedia[7] as the text source and `gpt-4-turbo-2024-04-09` (OpenAI, 2023) as the generative model, selected for its top performance on Faroese tasks in the ScandEval benchmark (Nielsen, 2023; Nielsen et al., 2024). To ensure non-trivial contexts, only articles with over 1,000 characters were included, i.e., 1675 articles in total and 655 articles used for the validated dataset. We set the model temperature to 1.0 and generated a maximum of 1,024 tokens, with a consistent random seed (4242) to maintain re-

[7]This dump: `https://hf.co/datasets/alexandrainst/scandi-wiki`.

producibility. The system prompt we use is the following:

> You are a helpful Faroese question answering dataset generator. The only language you know is Faroese.

While we did not explore Faroese-language prompting or prompt variations in this study, such modifications could potentially improve the effectiveness of our approach. As our primary focus was developing a question-answering dataset for Faroese, we leave prompt optimisation for future work. The following prompt was used for generating QA pairs:

> The following is a Wikipedia article in Faroese.
>
> ¡article¿
> {article}
> ¡/article¿
>
> Generate 2 to 10 questions about the article, depending on the length of the article, all of which answered in the article.
>
> You also have to supply answers to the questions, and the answers have to appear exactly as written in the article (including same casing).
>
> The answers should only contain the answers themselves, and not the surrounding sentence - keep the answers as short as possible.
>
> The answers have to be different from each other.
>
> All your questions and answers must be in Faroese.
>
> Your answer must be a JSON dictionary with the key "results", with the value being a list of dictionaries having keys "question" and "answer".

Lastly, we use the following prompt to re-write the questions:

> The following is a Faroese question.
>
> ¡question¿
> {question}
> ¡/question¿
>
> Re-write the question, preserving the meaning, using synonyms or a different (valid) word order.
>
> Your question must be in Faroese.
>
> Your answer must be a JSON dictionary with the key "question".

In both prompts, we replace {article} and {question} with the actual Wikipedia article and the generated question, respectively.

5 The Dataset

5.1 Format

The validated QA pairs are stored in a structured format, with each entry containing a unique identifier (`id`), the source article's URL (`url`), the article title (`title`), the full text (`context`), the generated and rephrased question (`question`), and an answers dictionary (`answers`) that includes the answer text and its character index (`answer_start`) within the context. This structure ensures compatibility with standard extractive QA formats like SQuAD (Rajpurkar et al., 2016), enabling seamless integration with existing NLP frameworks and models.

5.2 Statistics

The tentative dataset in our Faroese case consisted of **10,001** samples, which were randomly selected from the Wikipedia articles meeting our length criteria (¿1,000 characters). From these samples, **4,130** were annotated by a human annotator. Out of the annotated samples, **1,759** were annotated as CORRECT, **1,908** were INCORRECT[8] and **222** had an INCORRECT ANSWER. While the initial validation was performed by a single annotator, we conducted a second validation phase specifically for the samples marked as CORRECT, where these samples were evenly split between two annotators: the original annotator and a second native Faroese speaker. During this step, **41** out of the **1,759** CORRECT samples were found to have been incorrectly labelled as CORRECT by the annotator, which was then corrected. Additionally, **241** samples have the label CORRECTED where the original question has been corrected by the human annotator (this includes the 41 incorrectly labelled samples which were corrected). These corrected samples are intended to both measure and mitigate potential biases introduced by GPT-4-turbo during the initial sample creation. By comparing model performance on the corrected versus uncorrected samples of the dataset, we can assess whether the model exhibits any bias toward its own generated questions.

5.3 Dataset Versions

We are releasing three versions of the FoQA dataset on the Hugging Face Hub[9]. The format of the dataset is compatible with standard extractive QA formats like SQuAD. The primary version, **default**, contains 2,000 validated examples (comprising 1,759 initially correct examples and 241 examples that were corrected during re-

[8] While more than half of all generated question/answer pairs were marked as incorrect, we release the full dataset to enable researchers to study GPT-4's error patterns in Faroese.

[9] Available at `https://huggingface.co/datasets/alexandrainst/foqa`.

view), including 848 for training, 128 for validation, and 1,024 for testing, with shortened contexts for improved usability. The second version, **all-samples**, includes all 10,001 examples from the initial dataset, retaining full, unshortened contexts, even those that were rejected or not validated. The final version, **incorrect-samples**, comprises 2,395 examples that were rejected during the manual review process.

5.3.1 Question Types

We used the `gpt-4o-2024-05-13` model from OpenAI to annotate the questions into categories and we used the following system prompt:

> Categorize the question (written in Faroese) based on the type of question it is. The question types are "time" for questions that ask about the time of something, "place" if they ask for a place, "people" if they ask about a person, "object" if they ask about an object or a non-person entity. If the question does not fit any of these categories, respond with "other".

Most questions received the people label (679, 33.95%), followed by object (516, 25.80%), time (367, 18.35%), place (290, 14.50%) and other (148, 7.40%).

To assess the quality of the automatic question categorisation, the annotator manually validated 200 randomly sampled questions from the dataset. The validation methodology included assigning binary scores: 1 for correct categorization and 0 for incorrect categorisation. The validation followed an inclusive approach, accepting multiple valid category assignments where applicable. For instance, questions about a person's birthplace (e.g., "Where was Turi Sigurardóttir born?") were considered correctly categorised if labelled as either "person" or "place," as both categories are contextually relevant to the question's intent. This flexible validation framework acknowledges the inherent ambiguity in question categorisation, where multiple interpretations may be equally valid.

The manual validation revealed an error rate of 7.5% (15 incorrect categorisations out of 200 validated samples), suggesting that the GPT-4o categorisation system achieved 92.5% accuracy on the validated subset.

The annotator also conducted a qualitative error type analysis. Here it was found that common error types in the QA dataset include grammatical gender mistakes, such as using neuter instead of masculine forms in questions about pool length (e.g., "Hvussu langur er svimjihyli.NEUT í kappingunum"). Incorrect phrasing surrounding years, like omitting the preposition "í" (in) when asking about dates (e.g., "Hvørjum ári doyi Stephen Hawking?"), is also prevalent. Icelandicisms appear as words that are partially or fully Icelandic (e.g., the use of "hrai" (speed) inflected as a Faroese noun in "Hvør er hrain á jørini í kilometrum hvønn tíma?"). The questions and answers also contained errors in punctuation, spelling, and capitalization, as seen in the improper capitalization of "Smyril" (merlin) when referring to the bird rather than the ferry (e.g., "Hvat ger Smyril?"). Lastly, some incorrect terms are used consistently (e.g., "høvusbýur" (main city) used instead of "høvusstaur" (capital) when asking about capital cities).

Model Name	F1 Score	Exact Match
GPT-4-turbo[10]	77.6 ± 1.0	55.6 ± 1.8
GPT-4o[11]	77.1 ± 1.0	54.1 ± 1.6
GPT-4o-mini[12]	75.2 ± 1.0	51.2 ± 1.5
Llama-3.1-8B[13]	73.6 ± 1.2	51.9 ± 1.5
GPT-SW3-6.7B[14]	63.4 ± 2.2	45.2 ± 2.1
Mistral-7B[15]	62.4 ± 1.7	45.0 ± 1.6
FoBERT[16]	36.0 ± 1.7	26.8 ± 1.5
mDeBERTa-v3[17]	30.6 ± 1.6	21.0 ± 1.2
ScandiBERT[18]	30.9 ± 2.7	21.9 ± 2.3

Table 1: Evaluation results on FoQA according to F1 scores and exact match.

6 Evaluation

We evaluated several models on the dataset. Since we ensured that all answers appear exactly as in the documents, this allows us to evaluate both encoder models and decoder models on the dataset. We evaluate both Faroese and massively multilingual models on FoQA, the results of which can be found in Table 1.

We also evaluated the model used to generate the dataset, `gpt-4-turbo-2024-04-09`, on

[10] Full OpenAI model ID: `gpt-4-1106-preview`
[11] Full OpenAI model ID: `gpt-4o-2024-05-13`
[12] Full OpenAI model ID: `gpt-4o-mini-2024-07-18`
[13] `https://hf.co/meta-llama/Llama-3.1-8B`
[14] `https://hf.co/AI-Sweden-Models/gpt-sw3-6.7b-v2`
[15] `https://hf.co/mistralai/Mistral-7B-v0.3`
[16] `https://hf.co/vesteinn/FoBERT`
[17] `https://hf.co/microsoft/mdeberta-v3-base`
[18] `https://hf.co/vesteinn/ScandiBERT-no-faroese`

the corrected samples, before and after the correction. This was to test whether the model is biased towards its own generated questions, or whether it generalises to the corrected ones as well. Surprisingly, the model ended up performing significantly better[19] on the corrected samples, rather than the samples it had generated itself.

7 Discussion and Future Work

Our evaluation of the FoQA dataset reveals insights into the performance of various language models on Faroese QA tasks. GPT-4-turbo and GPT-4o achieved the highest performance scores in our evaluation, though further research would be needed to understand whether this indicates genuine Faroese language comprehension or other factors like strong general question-answering capabilities. This finding suggests promising directions for low-resource language processing, while highlighting the need for more detailed investigation into how these models handle Faroese specifically.

An important observation from our annotator indicates that most errors in the generated questions were grammatical in nature rather than contextual. This suggests a need for dedicated benchmarks specifically measuring grammatical correctness of LLMs in Faroese, which would complement FoQA's focus on QA capabilities.

Early question answering datasets like SQuAD faced criticism that their questions were too simplistic, often directly mirroring the source text structure. Later datasets like TyDi QA (Clark et al., 2020) and Natural Questions in Icelandic (Snæbjarnarson and Einarsson, 2022b) addressed this by having annotators create natural questions first, which were later matched to source material. This approach prevented the tight coupling between question phrasing and source text that can make questions artificially easy. Following this insight, we implemented question rephrasing in our methodology. However, we acknowledge that we did not specifically measure performance differences between original and rephrased questions, which would require separate evaluation sets.

We found that encoder models like mDeBERTa-v3 (He et al., 2021a,b), FoBERT and ScandiBERT (Snæbjarnarson and Einarsson, 2022a) perform significantly worse than the decoder models, but that could simply be explained by the fact that these models differ in sizes by several orders of magnitude. A controlled experiment will need to reveal whether architectural choices are the real cause for difference in performance or whether it is due to other reasons such as parameter count. A performance gap has been observed between encoder-type models and decoder-type models across other languages and (Nielsen et al., 2024) suggests that certain architectures may be inherently better suited for specific language processing tasks.

For future work, we propose evaluating larger open models, such as 70B parameter models and even larger ones like Llama 3.1 405B (Dubey et al., 2024). Additionally, assessing the performance of Claude 3.5 Sonnet (Anthropic, 2024) would be valuable, given its strong performance on Icelandic NLP tasks[20] since Icelandic is a language closely related to Faroese.

8 Conclusion

We introduced FoQA, the first Faroese extractive question-answering dataset, containing 2,000 QA pairs. All samples underwent initial validation by one annotator, followed by a second validation phase where the correct samples were split equally between the original annotator and a second annotator. Our evaluation reveals significant performance gaps between decoder-based LLMs and encoder models, with GPT-4-turbo achieving the highest F1 score of 77.6, while encoder models like mDeBERTa-v3 and ScandiBERT scored around 30. Notably, our analysis of question types shows a diverse distribution across categories, with people-related questions comprising the largest portion at 33.95%. The dataset's manual validation process identified common error patterns, including grammatical gender mistakes and Icelandicisms, providing valuable insights for future Faroese language model development. The FoQA dataset serves as valuable benchmark for evaluating Faroese language understanding. Additionally, our contributions include a semi-automated methodology for creating extractive QA datasets for low-resource languages.

[19] $p = 0.0007$ for F1-score and $p = 0.0185$ for exact match, using a two-tailed t-test.

[20] `https://huggingface.co/spaces/mideind/icelandic-llm-leaderboard`

Limitations

A significant limitation of our dataset is that our current annotation process does not differentiate between grammatical errors and contextual errors in the generated questions. This granular error categorisation would provide valuable insights for improving model performance and understanding specific challenges in Faroese language generation.

The use of GPT-4-turbo for dataset generation introduces potential biases in the linguistic patterns of the generated text. Despite native speaker validation, there remains a risk that the generated questions may not fully capture natural Faroese language patterns and could subtly reflect machine-generated language characteristics.

Our methodology relied on a single annotator for the initial validation phase, which means we could not perform traditional inter-annotator agreement measurements. While this limitation was intentional, as our approach aimed to demonstrate the feasibility of creating useful datasets with minimal human resources, it does impact our ability to measure annotation consistency quantitatively. Another limitation is our lack of evaluation on the non-rephrased questions. This missing comparison makes it difficult to quantify the impact of our question rephrasing strategy and determine whether it actually increased question difficulty as intended.

Furthermore, Faroese Wikipedia, while a valuable resource, is relatively small and occasionally contains ungrammatical content due to a limited pool of contributors. This occasionally led to incorrect-answer errors, since the answers are extracted directly from the source text. And lastly, the current size of 2,000 validated QA pairs, while a solid starting point, is relatively small compared to QA datasets for high-resource languages, which may limit its capacity to train or fine-tune LLMs effectively.

Ethical Statement

The creation of language resources for low-resource languages like Faroese raises important ethical considerations, particularly when utilising LLMs. Our dataset generation process involved processing approximately 1,675 Faroese Wikipedia articles through GPT-4-turbo. While this automated approach enabled efficient initial data generation, we acknowledge the computational resources required and their environmental impact, and we conservatively estimate that the processing spanned 48 GPU hours. We note that OpenAI's infrastructure runs on Azure, and Azure will be running on 100% renewable energy by 2025 and has been carbon neutral since 2012[21].

A primary ethical concern in using LLMs for low-resource language content generation is the potential introduction of non-native language patterns and cultural misrepresentations. This risk is particularly relevant for Faroese, where preserving authentic linguistic patterns and cultural context is crucial. To address these concerns, we implemented a comprehensive validation protocol requiring native speaker review of all generated content. This human-in-the-loop approach helped identify and correct systematic errors while ensuring linguistic authenticity.

To maximise the dataset's benefit to the Faroese language technology community, we have made it freely available under an open-source license. We are committed to ongoing maintenance and error correction, ensuring the dataset remains a valuable resource for Faroese language technology development while maintaining high standards of linguistic quality and cultural authenticity.

Acknowledgments

AS was supported by the European Commission under grant agreement no. 101135671. We thank the reviewers for their constructive and helpful comments, which have significantly improved the quality and clarity of our manuscript.

[21] See more information on Azure's sustainability page.

References

Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. 2019. Gradio: Hassle-Free Sharing and Testing of ML Models in the Wild. *arXiv preprint arXiv:1906.02569*.

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

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.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised Cross-lingual Representation Learning at Scale. *arXiv preprint arXiv:1911.02116*. *Equal contribution for first two authors.

Marta R. Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. *arXiv preprint arXiv:2207.04672*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

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*.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long Form Question Answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv preprint arXiv:2312.10997*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021a. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. *arXiv preprint arXiv:2111.09543*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021b. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. In *International Conference on Learning Representations*.

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

Dan Saattrup Nielsen, Kenneth Enevoldsen, and Peter Schneider-Kamp. 2024. Encoder vs Decoder: Comparative Analysis of Encoder and Decoder Language Models on Multilingual NLU Tasks. *arXiv preprint arXiv:2406.13469*.

OpenAI. 2023. New models and developer products announced at DevDay.

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

Njall Skarphedinsson, Breki Gudmundsson, Steinar Smari, Marta Kristin Larusdottir, Hafsteinn Einarsson, Abuzar Khan, Eric Nyberg, and Hrafn Loftsson. 2023. GameQA: Gamified Mobile App Platform for Building Multiple-Domain Question-Answering Datasets. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 152–160, Dubrovnik, Croatia. Association for Computational Linguistics.

Vésteinn Snæbjarnarson and Hafsteinn Einarsson. 2022a. Cross-Lingual QA as a Stepping Stone for Monolingual Open QA in Icelandic. In *Proceedings of the Workshop on Multilingual Information Access (MIA)*, pages 29–36, Seattle, USA. Association for Computational Linguistics.

Vésteinn Snæbjarnarson and Hafsteinn Einarsson. 2022b. Natural Questions in Icelandic. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4488–4496, 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.

Akchay Srivastava and Atif Memon. 2024. Towards Robust Evaluation: A Comprehensive Taxonomy of Datasets and Metrics for Open Domain Question Answering in the Era of Large Language Models. *IEEE Access*.

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.

Brandon T Willard and Rémi Louf. 2023. Efficient Guided Generation for Large Language Models. *arXiv preprint arXiv:2307.09702*.