FoQA: A Faroese Question-Answering Dataset

Annika Simonsen, Dan Saattrup Nielsen & Hafsteinn Einarsson

RESOURCEFUL-2025, Tallinn, Estonia

TrustLLM

UNIVERSITY OF ICELAND

ALEXANDRA INSTITUTE

Funded by the European Union

Why Do We Need QA for Faroese?

- Recent advances in NLP have been tremendous

- Current challenge:
  - Performance gap between high-resource and low-resource languages
  - Lack of annotators for low-resource languages with few speaker

- Can the cost of data production be lowered with LLMs?

Funded by the European Union

Our Contribution

1. Efficient Dataset Creation Method
   - Single annotator methodology for extractive QA datasets using semi-automated approach
   - Valuable for retrieval-augmented generation applications (Gao et al.,2023)

Funded by the European Union

Our Contribution

2. First Faroese QA Dataset
   - Practical implementation of our methodology, creating a vital resource for Faroese language technology
   - Full codebase and dataset available as open-source resources

GitHub

Link to code: https://github.com/alexandrainst/foqa

Hugging Face

Link to dataset: https://huggingface.co/datasets/alexandrainst/foqa

Funded by the European Union

Question Answering Systems Overview

- QA systems are divided into **extractive** and **abstractive types** (Fan et al., 2019).
  - Our focus is on **extractive QA**, where answers are directly pulled from the text.

- The Stanford Question Answering Dataset, or SQuAD, created by Rajpurkar et al. in 2016, is a prime example of this approach, containing over 100,000 QA pairs. (Rajpurkar et al., 2016)

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

Funded by the European Union

Dataset Creation Methodology

Key components:

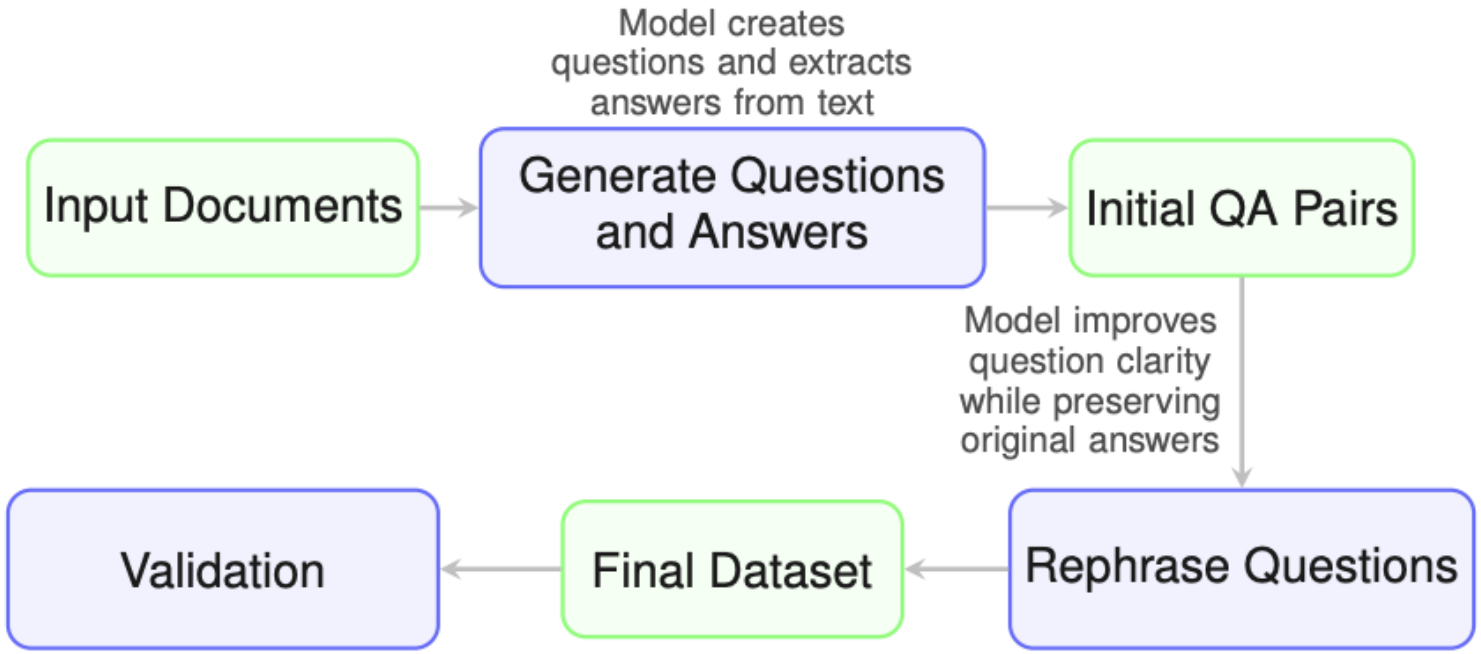
- **Text Corpus**
- **Generative Model**
- **Specialized QA generation functions**

Two-step process:

1. Initial generation (create initial QA pairs from corpus)
2. Question Refinement (rewrite questions while preserving answers)

Funded by the European Union

Dataset Creation Methodology




```
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.''
  }
]
}
```


Funded by the European Union

Why Rephrase?

- *“Jane Smith is an executive and her bike is red.”*

Funded by
the European Union

Why Rephrase?

- *"Jane Smith is an executive and her bike is red."*
- *"What colour is Jane Smith's bike?"*

Funded by the European Union

Why Rephrase?

- *"Jane Smith is an executive and her bike is red."*

- *"What colour is the executive's bicycle?"*
  - more sophisticated comprehension abilities, including synonym recognition and multi-hop reasoning in this example

Funded by the European Union

Annotation interface

FoQA Validation

This app automatically fetches examples from the Faroese Question Answering dataset (FoQA), allowing you to annotate whether the question and answer are correct Faroese or not.

Sample ID

0

Question

Hvørjar trupulleikar hevði Jeremia í síni tænastu?

Answer

Mótstøðu frá gomlu vinum sínum, frá starvsbrøðrum sínum, frá fólkinum sum heild og frá valdsharrunum.

Correct

Incorrect

Incorrect Answer

Save results

Funded by the European Union

Three-tier Annotation Guidelines for Quality Assurance

Tier 1: Grammatical Assessment

Evaluate grammatical correctness in Faroese

Tier 2: Semantic Assessment

Check question-answer relationship and context

Tier 3: Final Classification

Assign final status and optional secondary review

Funded by the European Union

The FoQA setup

- Data Source & Model
  - Faroese Wikipedia corpus (1,675 articles)
  - Articles > 1,000 characters (655 for validated dataset)
  - Model: **gpt-4-turbo-2024-04-09** (OpenAI, 2023)
    - Selected based on ScandEval benchmark performance (Nielsen, 2023; Nielsen et al., 2024)

- Configuration
  - Temperature: 1.0
  - Max tokens: 1,024
  - Random seed: 4242

Funded by the European Union

The FoQA setup

- System Prompt

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

- 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".

- Rewrite Question

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".

Funded by the European Union

Dataset Statistics & Analysis

- Tentative dataset of **10,001 samples**
- 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** and **222** had an **INCORRECT ANSWER**.
  - Additionally, the human annotator manually corrected 241 samples as **CORRECTED**, leading to **2,000** correct samples total
- 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%).

Funded by the European Union

Qualitative Error Type Analysis

- Most 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 svimjihylið.NEUT í kappingunum*")

- Icelandicisms appear as words that are partially or fully Icelandic (e.g., the use of "*hraði*" (speed) inflected as a Faroese noun in "*Hvør er hraðin á jørðini í kilometrum hvønn tíma*?")

Funded by the European Union

Qualitative Error Type Analysis

- 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 "*høvuðsstaður*" (capital) when asking about capital cities)

Funded by the European Union

Dataset Versions

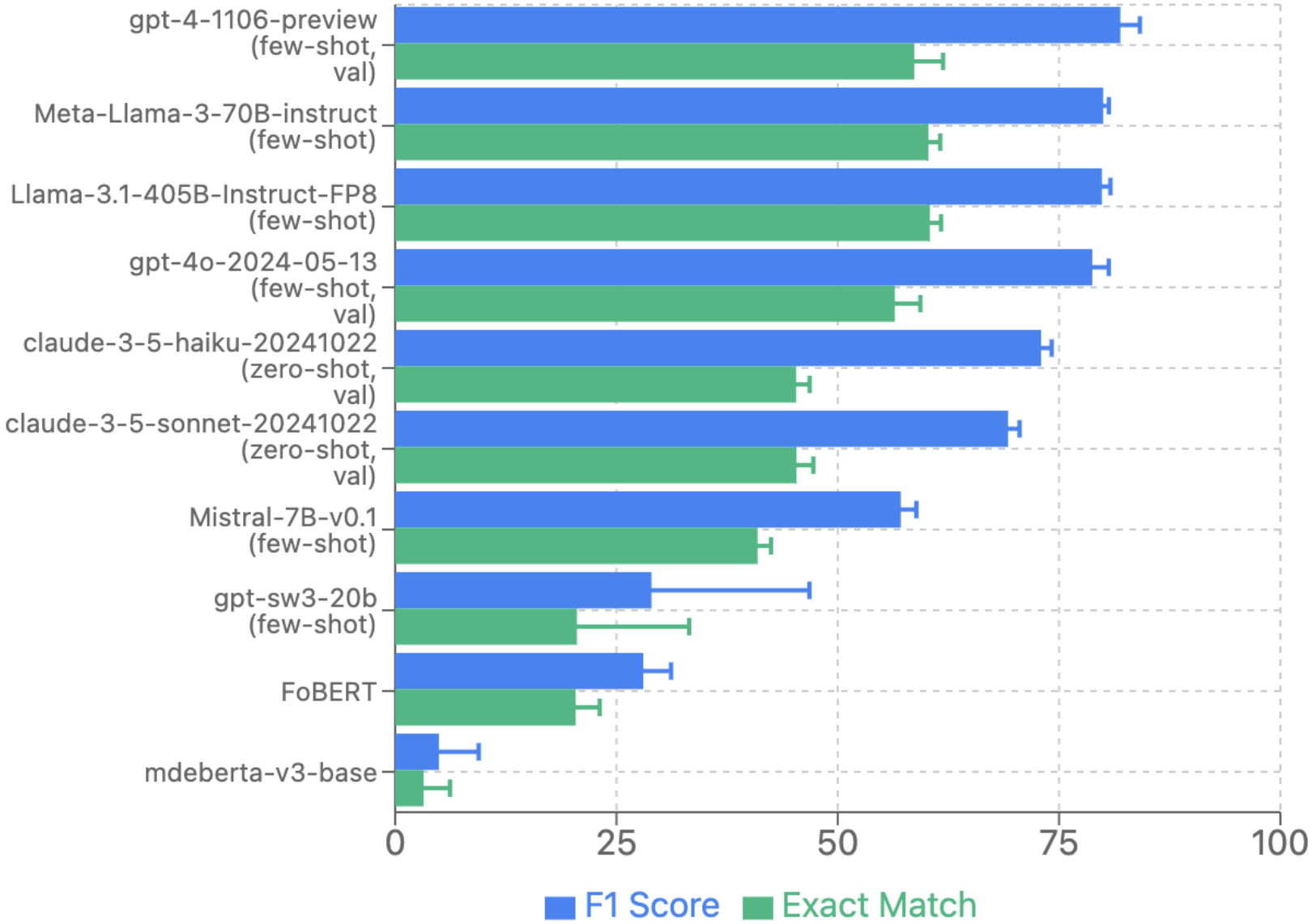
- **Default:** 2,000 human-validated examples (848 for training, 128 for validation, and 1,024 for testing, with shortened contexts for improved usability)
- **All-samples:** all 10,001 examples from the initial dataset, retaining full, unshortened contexts, even those that were rejected or not validated
- **Incorrect-samples:** the 2,395 examples that were rejected during the manual review process

Funded by the European Union

ScandEval Leaderboard FoQA (26 Feb 2025)

Model Name	F1 Score	Exact Match
gpt-4-1106-preview (few-shot, val)	81.93 ± 2.20	58.65 ± 3.25
Meta-Llama-3-70B-instruct (few-shot)	79.98 ± 0.65	60.24 ± 1.33
Llama-3.1-405B-Instruct-FP8 (few-shot)	79.84 ± 0.96	60.41 ± 1.26
gpt-4o-2024-05-13 (few-shot, val)	78.76 ± 1.86	56.45 ± 2.88
claude-3-5-haiku-20241022 (zero-shot, val)	72.97 ± 1.19	45.29 ± 1.54
claude-3-5-sonnet-20241022 (zero-shot, val)	69.24 ± 1.30	45.34 ± 1.90
Mistral-7B-v0.1 (few-shot)	57.12 ± 1.76	40.95 ± 1.51
gpt-sw3-20b (few-shot)	28.96 ± 17.84	20.54 ± 12.67
FoBERT	28.03 ± 3.14	20.40 ± 2.71
mdeberta-v3-base	4.95 ± 4.49	3.22 ± 2.98

Funded by the European Union



Funded by the European Union

Discussion

- GPT-4 variants achieved highest performance, but further research needed to determine if this indicates true Faroese comprehension or general QA capabilities
- Most errors in generated questions were grammatical rather than contextual, suggesting need for dedicated Faroese grammar benchmarks
- Encoder models performed significantly worse than decoder models - controlled experiments needed to determine if architectural choices or parameter count is the cause

Funded by the European Union

Limitations

- No differentiation between grammatical and contextual errors in annotation process, limiting insight into specific challenges
- Potential biases from using GPT-4-turbo for dataset generation - generated questions may not fully capture natural Faroese patterns
- Single annotator approach prevents inter-annotator agreement measurements and quantitative consistency analysis
- Dataset size of 2,000 QA pairs is relatively small compared to high-resource language datasets

Funded by the European Union

Future Work

- Conduct controlled experiments to investigate architectural impact vs parameter count on model performance
- Would a model perform worse on a dataset created by a human than on a dataset created using an LLM?
- Develop dedicated benchmarks for measuring grammatical correctness in Faroese

Funded by the European Union

Conclusion

- Created FoQA: First Faroese extractive QA dataset with 2,000 validated pairs
- Dataset and code is available right now!

GitHub

Link to code:
https://github.com/alexandrainst/foqa

Hugging Face

Link to dataset:
https://huggingface.co/datasets/alexandrainst/foqa

Funded by the European Union

Thank you!

TrustLLM

UNIVERSITY OF ICELAND

ALEXANDRA INSTITUTE

Funded by the European Union

References

- 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.
- 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.

Funded by the European Union