WikiQA-IS: Assisted Benchmark Generation and Automated Evaluation of Icelandic Cultural Knowledge in LLMs

Þórunn Arnardóttir
Miðeind
thorunn@mideind.is

Elías Bjartur Einarsson
Miðeind
elias@mideind.is

Garðar Ingvarsson Juto
Miðeind
gardar@mideind.is

Þorvaldur Páll Helgason
Miðeind
thorvaldur@mideind.is

Hafsteinn Einarsson
University of Iceland
hafsteinne@hi.is

Abstract

This paper presents WikiQA-IS, a novel question-answering dataset focusing on Icelandic culture and history, along with an automated pipeline for dataset generation and evaluation. Leveraging GPT-4 to create questions and answers based on Icelandic Wikipedia articles and news sources, we produced a high-quality corpus of 2,000 question-answer pairs. We introduce an automatic evaluation method using GPT-4o as a judge, which shows strong agreement with human evaluations. Our benchmark reveals varying performances across different language models, with closed-source models generally outperforming open-weights alternatives. This work contributes a resource for evaluating language models' knowledge of Icelandic culture and offers a replicable framework for creating similar datasets in other cultural contexts.

1 Introduction

Recent advancements in natural language processing (NLP) have led to significant improvements in question-answering systems, particularly through large language models (LLMs) (Brown, 2020). While these models show impressive capabilities, they can generate incorrect or fabricated information, a phenomenon known as hallucination (Bender et al., 2021; Huang et al., 2024). This makes it crucial to systematically measure how much factual knowledge these models actually possess about specific domains, such as individual cultures or topics. Current evaluation methods often lack domain-specific benchmarks, making it difficult to assess models' true understanding of particular cultural contexts. This paper presents an automated approach to generate and evaluate questions and answers, using Icelandic culture and history as a case study.

Icelandic, despite its small speaker base, has a rich literary and historical heritage. However, creating comprehensive QA datasets for such domains is resource-intensive if done manually. While prior work on Icelandic QA datasets has focused on language and reading comprehension (Snæbjarnarson and Einarsson, 2022b; Skarphedinsson et al., 2023; Snæbjarnarson and Einarsson, 2022a; Geirsson, 2013; De Bruyn et al., 2021), there remains a need for a dataset testing knowledge of culture and history in an open-ended fashion.

Our research introduces a method leveraging an LLM to automate the generation of high-quality questions and answers based on Icelandic Wikipedia articles, inspired by previous work extracting knowledge from Wikipedia (Yang et al., 2015; Auer et al., 2007) and work on automatic QA dataset creation (Lewis et al., 2019). This approach addresses the challenge of creating large-scale datasets for low-resource languages and extends the application of language models to cultural and historical knowledge evaluation.

The main contribution of this paper is the WikiQA-IS corpus[1] along with the pipeline used to generate the corpus and the automatic evaluation approach[2]. This research not only contributes to create benchmarks focusing on Icelandic cultural knowledge but also offers a replicable framework adaptable to other languages and cultural contexts.

[1] Dataset released under a CC BY license: https://repository.clarin.is/repository/xmlui/handle/20.500.12537/347

[2] Code: https://github.com/icelandic-lt/AutomaticQAPipeline and https://github.com/mideind/lm-evaluation-harness/blob/add-icelandic-evals/lm_eval/tasks/icelandic_qa/icelandic_wiki_qa.yaml

Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL 2025), pages 64–73
March 2, 2025 ©2025 Association for Computational Linguistics

2 Methods

2.1 Dataset Preparation

The questions in this work are based on the Icelandic Wikipedia and on the news from RÚV in the Icelandic Gigaword corpus (Steingrímsson et al., 2018). From Wikipedia, the 41,569 articles that contained at least 250 characters were used, and from the RÚV news, because the data is extensive, only a portion of articles that contained at least 500 characters were used. For each page used in a given source, we kept track of the "url", "title" and "text" as fields in a JSONL file. The text field serves as the basis for question generation.

2.2 Question Generation Pipeline

2.2.1 Document to Request Conversion

We first convert the documents into requests suitable for the GPT model. This process involves creating a JSON object for each document, which includes a system prompt and a user prompt, both in Icelandic. The system prompt is: *Þú ert vandvirk aðstoðarmanneskja* which translates to *You are a meticulous assistant.*

The complete prompt structure pairs document text with an instruction component that guides the model in generating questions and performing dual evaluations: it must score both the quality and relevance of each generated question and assess the document's connection to Icelandic culture and history, using a scale from 0 to 1 for both metrics. These scores enable automatic filtering of questions and documents that would likely be rejected by human annotators. The instruction component underwent several rounds of refinement until it reliably produced high-quality questions from the input texts, and is provided below in both Icelandic and English.

```
Semdu almenna spurningu upp úr
 ↪  þessu skjali og svaraðu
 ↪  henni ef skjalið fjallar að
 ↪  einhverju leyti um íslenska
 ↪  menningu og/eða íslenska
 ↪  sögu.
Spurningin á að vera um innihald
 ↪  skjalsins, ekki skjalið
 ↪  sjálft. Ekki vísa í skjalið
 ↪  í spurningunni.
Hafðu svarið eins hnitmiðað og
 ↪  hægt er.
Ef spurning og/eða svar vísar
 ↪  til tíma þarf sá tími eða
 ↪  ártal að vera tekið fram í
 ↪  bæði spurningu og svari.
Spurning og/eða svar má ekki
 ↪  vísa til hluta sem eru
 ↪  núverandi, heldur þarf
 ↪  tímasetning að vera til
 ↪  staðar.
Skilaðu niðurstöðunni á
 ↪  eftirfarandi json sniði:

{"question": [question],
 ↪  "answer": [answer], "id":
 ↪  [doc["url"] OR
 ↪  doc["xml_id"]],
 ↪  "question_score": [score
 ↪  0.0-1.0], "document_score":
 ↪  [score 0.0-1.0], "source":
 ↪  [doc["source"]]}

Spurningin á að vera almenn og
 ↪  tengjast íslenskri menningu
 ↪  og/eða íslenskri sögu.
 ↪  "question_score" á að meta
 ↪  hversu mikið spurning
 ↪  tengist íslenskri menningu
 ↪  og/eða íslenskri sögu og
 ↪  hversu góð og almenn hún er
 ↪  en "document_score" á að
 ↪  meta hversu gott skjalið er
 ↪  og hversu mikið það tengist
 ↪  íslenskri menningu og/eða
 ↪  íslenskri sögu.
Ef skjalið er stutt, slæmt eða
 ↪  ekki er hægt að skapa
 ↪  spurningu upp úr skjalinu,
 ↪  skilaðu þá sama json sniði
 ↪  með engu innihaldi fyrir
 ↪  "question" og "answer".
Ef skjalið fjallar ekki um
 ↪  íslenska menningu eða
 ↪  íslenska sögu, skilaðu þá
 ↪  sama json sniði með engu
 ↪  innihaldi fyrir "question"
 ↪  og "answer".
```

An English translation of the prompt is given below.

```
Generate a general question from
 ↪  this document and answer it
 ↪  if the document relates in
 ↪  any way to Icelandic culture
 ↪  and/or Icelandic history.
The question should be about the
 ↪  content of the document, not
 ↪  the document itself. Don't
 ↪  reference the document in
 ↪  the question.
Keep the answer as concise as
 ↪  possible.
If the question and/or answer
 ↪  refers to time, that time or
 ↪  year must be specified in
 ↪  both question and answer.
Question and/or answer must not
 ↪  refer to current things,
 ↪  rather a timestamp must be
 ↪  present.
Return the result in the
 ↪  following json format:

{"question": [question],
 ↪  "answer": [answer], "id":
 ↪  [doc["url"] OR
 ↪  doc["xml_id"]],
 ↪  "question_score": [score
 ↪  0.0-1.0], "document_score":
 ↪  [score 0.0-1.0], "source":
 ↪  [doc["source"]]}

The question should be general
 ↪  and relate to Icelandic
 ↪  culture and/or Icelandic
 ↪  history. "question_score"
 ↪  should evaluate how much the
 ↪  question relates to
 ↪  Icelandic culture and/or
 ↪  Icelandic history and how
 ↪  good and general it is,
 ↪  while "document_score"
 ↪  should evaluate how good the
 ↪  document is and how much it
 ↪  relates to Icelandic culture
 ↪  and/or Icelandic history.
If the document is short, poor,
 ↪  or it's not possible to
 ↪  create a question from the
 ↪  document, then return the
 ↪  same json format with no
 ↪  content for "question" and
 ↪  "answer".
If the document does not discuss
 ↪  Icelandic culture or
 ↪  Icelandic history, then
 ↪  return the same json format
 ↪  with no content for
 ↪  "question" and "answer".
```

Note that if the document is inadequate or unrelated to Icelandic culture/history, an empty response should be returned in the same JSON format.

2.2.2 API Calls to GPT

We make API calls to OpenAI's `gpt-4-turbo` model using the prepared requests. The model generates questions, answers, and scores based on the input documents.

2.2.3 Filtering Generated Questions

The generated questions and answers are filtered based on the scores provided by the LLM. We selected only questions that had both a document score of at least 0.7 and a question score of at least 0.7. Note that these thresholds were chosen based on intuition after inspecting the documents and questions. We discarded 29,450 questions created from the 41,569 Wikipedia articles through this approach, meaning that 29% of the automatically created questions were deemed adequate. For the RÚV news data, 5,350 questions, created from 6,672 articles, were discarded, which means that 20% of questions were adequate. The difference in adequacy can be explained by the fact that the question and document had to relate to Icelandic culture and/or history. The question-answer pairs that were not discarded were then eligible for manual question-answer pair review (see below), but note that not all pairs were manually reviewed.

2.2.4 Spelling and Grammar Correction

While `gpt-4-turbo` demonstrates strong comprehension of Icelandic, its generative capabilities in the language exhibit some limitations. The model produces generally intelligible output, but frequently requires grammatical corrections, particularly in terms of nominal inflection, which is a crucial feature of Icelandic morphology[3].

We use a Byte-Level Neural Error Correction Model for Icelandic to correct spelling and grammar in the generated questions and answers (Ingólfsdóttir et al., 2023). During this process, 26.49% of questions were corrected and 41.99% of answers.

2.2.5 Dataset Format

The dataset is available in different formats compatible with BIG-bench (Srivastava et al., 2022), OpenAI-evals and the Language Model Evaluation Harness (Gao et al., 2023).

2.3 Manual Question-Answer Pair Review

Question-answer pairs generated with the pipeline were reviewed by a single human annotator, a native speaker of Icelandic with a B.A. degree in general linguistics. Due to time restraints, only a portion of the generated question-answer pairs were manually reviewed. All pairs are, however, published as part of the dataset. In this process, the annotator had access to the context used to generate the question-answer pair. The annotator was instructed to work based on the following annotation guidelines and to discard or improve questions and answers if they did not meet some of these points. As a result, the majority of question-answer pairs were manually corrected so that they met the points in the guidelines.

- Questions and answers must be in Icelandic.
- Questions and answers must relate to Icelandic culture and/or history.
- A question can only include one question, and the answer must answer that question unambiguously and contain no information beyond that.
- A question and answer cannot include any spelling or grammar errors, and the text must be natural.

2.4 Automatic Evaluation

To evaluate the performance of language models on our dataset, we employed an automated evaluation process using `gpt-4o-2024-08-06` as a judge model. This process involves presenting the model under evaluation with a question, collecting its generated answer, and then providing the question, generated answer, and correct answer to the judge model for assessment. The judge model evaluates the correctness and relevance of the generated answer, providing a rating of "poor" (0 points), "fair" (0.5 points), or "excellent" (1 point). The instructions for the LLM are given below:

```
Please act as an impartial judge
 ↪  and evaluate the quality of
 ↪  the response provided by an
 ↪  AI assistant to the user
 ↪  question displayed below.
 ↪  Your evaluation should
 ↪  consider correctness. You
 ↪  will be given the question
 ↪  which was asked, a correct
 ↪  reference answer, and the
 ↪  assistant's answer. Begin
 ↪  your evaluation by briefly
 ↪  comparing the assistant's
 ↪  answer with the correct
 ↪  answer. Identify any
 ↪  mistakes. Be as objective as
 ↪  possible. Additional
 ↪  information beyond the
 ↪  reference answer's content
 ↪  should not be considered. If
 ↪  the assistant's answer is
 ↪  not in Icelandic but the
 ↪  reference answer is, you
 ↪  should rate the answer
 ↪  poorly. After providing your
 ↪  short explanation, you must
 ↪  rate the assistant's answer
 ↪  using the following scale:
 ↪  [[poor]]: Incorrect,
 ↪  off-topic or in a different
 ↪  language; [[fair]]:
 ↪  Partially aligns with the
 ↪  reference answer with some
 ↪  inaccuracies or irrelevant
 ↪  information; [[excellent]]:
 ↪  Accurate and relevant,
 ↪  matching the reference
 ↪  answer in content and
 ↪  language.
```

[3] The current ranking of models on the Icelandic inflection benchmark is shown on the Icelandic LLM leaderboard

2.5 Manual Evaluation

To validate the performance of the automatic evaluation process, three annotators also perform manual evaluation. In the manual evaluation phase, a human annotator receives the question and compares the generated answer to the reference answer. The human annotator is tasked with providing a rating of "poor" (0 points), "fair" (0.5 points), or "excellent" (1 point) and receives the same instructions as the LLM. We compute the agreement between the annotators and the LLM as a judge using Cohen's Kappa (Cohen, 1960).

2.6 Question Classification

We used `gpt-4o-2024-08-06` to classify the questions into five classes we considered to be representative of the majority of questions in the dataset. The prompt is given below and we used structured output so the model could only respond with one of the five given categories.

```
Categorize the question (written
 ↪  in Icelandic) 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 as 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'.
```

An annotator was tasked with evaluating whether the categorization was correct or not. They received instructions stating how the questions were categorized, along with the prompt, and were asked to judge whether the categorization was correct or not for 200 questions chosen uniformly at random.

3 Results

3.1 Dataset Generation and Curation

Our data generation and curation process produced a dataset of high-quality question-answer pairs focusing on Icelandic culture and history. The automatically generated pairs were reviewed to ensure their quality and relevance.

For the Wikipedia-based dataset, we examined 2,116 question-answer pairs, ultimately including 1,900 in the final set. This high retention rate of 89.8% demonstrates the effectiveness of our automated generation process. In contrast, the IGC-RÚV (Icelandic Gigaword Corpus – RÚV) dataset yielded a lower retention rate. Out of 274 reviewed pairs, only 100 met our inclusion criteria, resulting in a 36.5% retention rate. The observed difference can be attributed to the distinct focus of each corpus: while the RÚV corpus consists primarily of contemporary news content, the Wikipedia corpus contains a higher proportion of articles dedicated to Icelandic culture and history.

It is worth noting that most retained pairs required some level of correction. These ranged from minor spelling adjustments missed by our automatic correction tool to more substantial revisions of questions and answers based on the source documents. This manual refinement process was crucial in ensuring the dataset's overall quality, naturalness and accuracy. For the resulting dataset, the questions varied in length ranging from 15 to 210 characters and the answers varied from 2 to 233 characters. The distributions of question and answer lengths are shown in Figure 1.

3.2 Evaluation of Automatic Evaluation

To assess the reliability of our automatic evaluation method, we conducted a human evaluation study. Our automatic evaluation uses GPT-4o as a judge to evaluate responses from other LLMs, categorizing them as "Excellent", "Fair", or "Poor". To validate this approach, we sampled 100 responses each from `gpt-4o-2024-08-06` and `claude-3-5-sonnet-20240620`, that were then evaluated manually as described in Section 2.5. Tables 1 and 2 present the confusion matrices for GPT-4o and Claude 3.5 Sonnet, respectively.

The results demonstrate high agreement between our automatic evaluation method and human judgments. For GPT-4o judging GPT-4o responses, we observed an a Cohen's kappa score with human annotators ranging from 0.81 to 0.91. The evaluation of GPT-4o judging Claude 3.5 Sonnet responses showed slightly lower agreement but still strong agreement with Cohen's kappa ranging from 0.75 to 0.82. These results suggest that our judge based on GPT-4o provides a robust and effi-

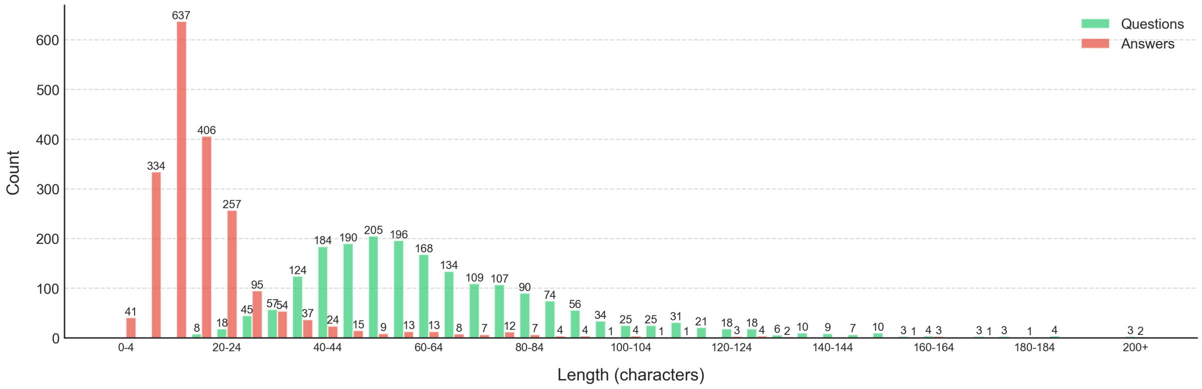

Figure 1: Distribution of question and answer lengths.

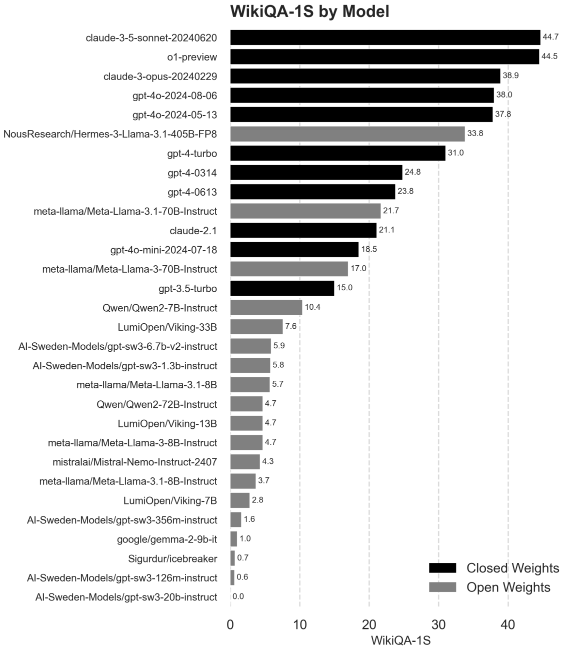

Figure 2: Performance comparison of various models on the WikiQA-IS dataset. The plot illustrates the accuracy of different models, with black bars representing closed weight models and gray bars representing open weight models.

cient means of evaluating LLM responses, closely aligning with human judgments.

In an effort to reveal systemic biases in the evaluation, we manually inspected the few examples where human and GPT-4o annotations differed. We see that in almost all cases where a human rated an answer higher than GPT-4o, the answer was either partially or fully correct, but contained some additional information which the LLM judge penalized it for more severely than the human. We also notice an opposite trend where in half of the cases where GPT-4o scored an answer as "fair" but a human as "poor", the answer was factually correct but required more domain knowledge to verify than the human could be expected to infer from the reference answer. This suggests that GPT-4o might be biased towards rewarding answers that align with its own factual knowledge, instead of comparing the answer against the reference answer in isolation. The LLM judge, however, never awarded an answer with the "excellent" score if it did not semantically match the reference answer, even when it was factually correct, indicating that this slight bias has limited impact.

3.3 LLM Performance

Figure 2 presents the performance of various language models on the WikiQA-IS benchmark. The results demonstrate a clear performance hierarchy among the evaluated models. The top-performing models are predominantly large, closed-source language models developed by major AI research companies. `Claude-3.5-sonnet-20240620` and `o1-preview` achieve the highest scores of 44.7 and 44.5, respectively, closely followed by `claude-3-opus-20240229` with 38.9. The GPT-4o variants also perform well, scoring 38.0 and 37.8, respectively.

Among the open-weights models, Llama 3.1 (405B) (Dubey et al., 2024) stands out with a score of 33.8, demonstrating competitive performance with some of the closed-weights models. This suggests that well-trained open-weights models can approach the capabilities of proprietary models in specialized tasks like answering questions about Icelandic culture and history.

There is a noticeable performance gap between the top-tier models and the rest of the field. Models such as GPT-4 variants show moderate performance, with scores ranging from 23.8 to 31.0. The performance then drops significantly for smaller models and earlier versions, with scores falling below 20 for models like Llama 3.1 (70B) and `claude-2.1`.

Open-weights models generally perform less well than their closed-source counterparts, with most scoring below 10 on the WikiQA-IS benchmark. However, there is significant variation among open-weights models, with some (like the top Llama models) performing much better than others. We specifically chose to include models from AI-Sweden (Ekgren et al., 2022) as they were amongst the only models trained specifically for Nordic languages at the time of the evaluation.

	Human Rating		
Judge Rating	Poor	Fair	Excellent
Poor	117	3	0
Fair	3	39	21
Excellent	0	0	117

Table 1: Agreement between three human annotators and GPT-4o judge for responses generated by GPT-4o.

	Human Rating		
Judge Rating	Poor	Fair	Excellent
Poor	79	5	0
Fair	7	15	12
Excellent	1	1	80

Table 2: Agreement between two human annotators and GPT-4o judge for responses generated by Claude 3.5 Sonnet

3.4 Question Difficulty Analysis

The analysis of model performance across questions revealed substantial variation in question difficulty (Figure 3). Most notably, 761 questions (roughly 38% of the dataset) received no "Excellent" rated responses from any of the 30 models tested, indicating that these questions were particularly challenging. The distribution of high-quality responses shows a rapid decline, with progressively fewer questions receiving multiple "Excellent" rated responses. Only a small subset of questions were answered excellently by more than 7 models, suggesting that most models tested struggle with consistent high-quality performance on questions related to Icelandic culture and history.

The gap between questions receiving "Excellent" rated responses and those receiving either "Fair" or "Excellent" rated responses remains relatively constant across the distribution, indicating that for most questions, several models typically provided "Fair" rather than "Excellent" responses. This pattern suggests that while models often capture some relevant information, they frequently include unnecessary details or minor inaccuracies in their responses. The rapid decline in both distributions also highlights that achieving a majority consensus among models on correct answers is rare, pointing to the continuing challenges in providing factful responses in this domain. 761 questions received no "Excellent" ratings, while 488 questions garnered neither "Excellent" nor "Fair" ratings. These findings indicate that a substantial portion of our dataset consists of questions that pose significant challenges for LLMs. To further investigate the nature of these challenging questions, we employed an LLM to systematically categorize each question into one of five types (object, people, place, time, and other). An annotator manually reviewed 200 questions to estimate the performance of this categorization and they were judged to be appropriate in 95.5% of cases. The confusion occurred where the category "other" should have been used instead of "object".

Table 3 presents a comparative distribution of these question types, contrasting the overall dataset with the subset of questions that no LLM could answer correctly. We observe that among the most difficult questions for LLMs, nearly half (48.05%) pertain to people or individuals, a marked increase from the 34.74% in the overall dataset. This disparity reflects the hallucination tendency of LLMs (Kalai and Vempala, 2024) since the names in the questions and the facts asked about rarely appear in the pretraining data.

4 Discussion

Our study demonstrates the effectiveness of leveraging LLMs for creating specialized question-answering datasets. The significant difference in retention rates between Wikipedia-based questions (89.8%) and news articles (36.5%) underscores the importance of source material selection

Question Set	People	Time	Object	Place	Other
All Questions	660 (34.7%)	576 (30.3%)	310 (16.3%)	229 (12.1%)	125 (6.6%)
Difficult Questions	234 (48.0%)	136 (27.9%)	53 (10.9%)	55 (11.3%)	10 (2.0%)

Table 3: Distribution of question types.

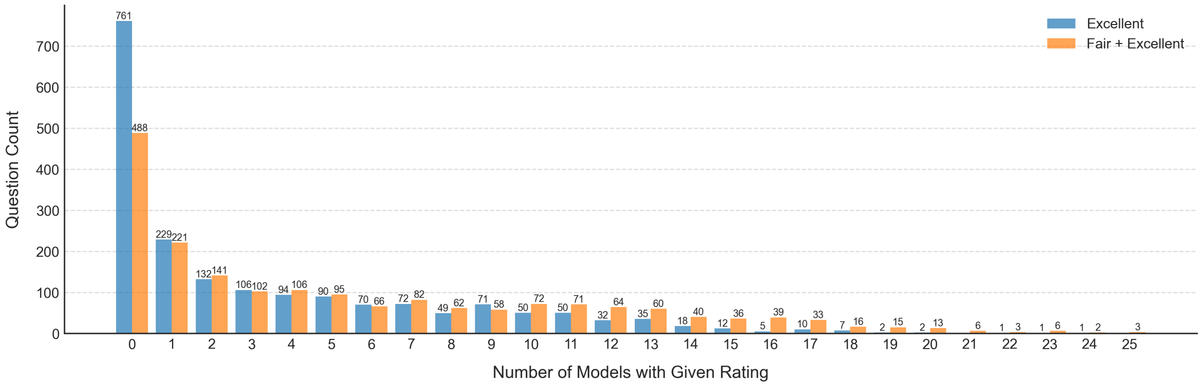

Figure 3: Distribution of question difficulty based on large language model performance. The histogram shows how many questions (y-axis) received a specific number of high-quality responses (x-axis). The blue bars represent questions receiving "Excellent" rated responses from LLMs, while the orange bars show questions receiving either "Fair" or "Excellent" rated responses. 488 questions received zero combined "Fair" or "Excellent" rated responses, indicating these questions were particularly challenging.

in QA dataset creation.

The performance analysis reveals a clear hierarchy among models, with closed-source models generally outperforming open-weights alternatives. This gap highlights ongoing challenges in democratizing advanced language understanding capabilities for specialized domains. Our automatic evaluation method shows promise for efficient, large-scale assessment, though it may be influenced by the judge model's capabilities and biases.

Future work could explore expanding source materials to reduce potential biases, and develop more comprehensive categorization of questions to uncover specific areas of model strength and weakness. While our method provides valuable insights into models' cultural knowledge, it represents just one facet of measuring world knowledge, and complementary approaches could offer a more holistic assessment of cultural understanding.

Ethics Statement

Experiments were conducted via OpenAI's API services, Anthropic's API services and on a local machine with eight A100 GPUs. While the exact computational infrastructure is not publicly disclosed, we estimate the carbon footprint based on the assumption that computation was performed in Microsoft Azure datacenters in Western Europe, with an estimated grid carbon intensity of 0.57 $kgCO_2eq/kWh$. Given OpenAI's non-disclosure of infrastructure details, we estimate that the experiments consumed in the order of 10 GPU hours, presumably on NVIDIA A100 PCIe 40/80GB GPUs with a Thermal Design Power of 250W.

The total estimated emissions for 10 GPU hours amount to 1.4 $kgCO_2eq$. For context, these emissions are equivalent to driving approximately 5.7 kilometers in a conventional internal combustion engine vehicle. We also 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[4].

We similarly estimate conservatively that answer generation and evaluation of other models is at most 20 GPU hours amounting to at most 2.8 $kgCO_2eq$.

Estimations were conducted using the Machine-Learning Impact calculator presented in (Lacoste et al., 2019).

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

Limitations

While our approach demonstrates promising results in creating and evaluating culturally-specific QA datasets, several limitations should be acknowledged. First, our reliance on Wikipedia and RÚV news articles as source material may introduce coverage biases. These sources, while authoritative, may not fully represent the breadth of Icelandic cultural knowledge, particularly oral traditions, contemporary cultural developments, or specialized academic research not covered in these venues.

The use of GPT-4 Turbo for question generation, while efficient, may introduce systematic biases in question formulation and potentially limit the diversity of question types. Although our manual review process helps mitigate these issues, it may not completely eliminate them. Using GPT-4 Turbo also introduces limitations on using the generated dataset based on OpenAI's terms of use,[5] particularly the clause on using output to develop models that compete with OpenAI. The generated dataset is published under a CC BY license but as its intended use is for benchmarking, we do not consider its publication to violate the terms of use.

Our automated evaluation method, despite showing strong correlation with human judgments, relies on large language models as judges, which may perpetuate certain biases or limitations inherent to these systems. The nature of our scoring system (poor/fair/excellent) may not fully capture nuanced differences in answer quality, particularly for questions about cultural interpretations or historical perspectives where multiple valid viewpoints might exist.

Finally, while our dataset size of 2,000 questions is substantial for a language with limited resources like Icelandic, it may not be comprehensive enough to fully evaluate an LLM's knowledge of Icelandic culture and history. The current version of the dataset also lacks explicit categorization of different aspects of cultural knowledge (e.g., literature, folklore, social customs), which could provide more granular insights into model performance across different cultural domains.

Acknowledgments

Þórunn and Elías were supported by the Icelandic Language Technology Programme and Garðar was supported by the European Commission under grant agreement no. 101135671. We thank the reviewers for their constructive feedback that helped us improve the manuscript.

References

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *international semantic web conference*, pages 722–735. Springer.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Maxime De Bruyn, Ehsan Lotfi, Jeska Buhmann, and Walter Daelemans. 2021. MFAQ: a multilingual FAQ dataset. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 1–13, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 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.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation.

Ólafur Páll Geirsson. 2013. Iceqa: Developing a question answering system for icelandic.

[5] `https://openai.com/policies/terms-of-use/`

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2024. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.* Just Accepted.

Svanhvít Lilja Ingólfsdóttir, Petur Ragnarsson, Haukur Jónsson, Haukur Simonarson, Vilhjalmur Thorsteinsson, and Vésteinn Snæbjarnarson. 2023. Byte-level grammatical error correction using synthetic and curated corpora. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7299–7316, Toronto, Canada. Association for Computational Linguistics.

Adam Tauman Kalai and Santosh S Vempala. 2024. Calibrated language models must hallucinate. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pages 160–171.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700.*

Patrick Lewis, Ludovic Denoyer, and Sebastian Riedel. 2019. Unsupervised question answering by cloze translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4896–4910, Florence, Italy. 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.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615.*

Steinþór Steingrímsson, Sigrún Helgadóttir, Eiríkur Rögnvaldsson, Starkaður Barkarson, and Jón Guðnason. 2018. Risamálheild: A very large Icelandic text corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, Lisbon, Portugal. Association for Computational Linguistics.