The Icelandic Approach:

Preserving and Revitalizing Linguistic and Cultural Diversity in AI

Ideation and rationale for a **Project** and **Playbook** supporting less-resourced languages and cultures in AI, inspired by lessons learned from Iceland's collaboration with OpenAI

September 2024

ALMANNARÓMUR

Government of Iceland
Ministry of Culture and Business Affairs

MIÐEIND

The Ministry of Culture and Business Affairs
Sölvhólsgata 7 — 101 Reykjavík
+354 545 9800 | mvf@mvf.is
In Collaboration with:
Almannarómur, The Icelandic Centre for Language Technology
Miðeind

Reykjavík, September 2024

For further information please contact:
mvf@mvf.is,
almannaromur@almannaromur.is
or mideind@mideind.is

https://mvf.is
https://almannaromur.is/en
https://mideind.is/en

1. The Goal: Preserving and Revitalizing Diversity in AI

Artificial intelligence (AI) holds immense potential for enriching people's lives and boost productivity. However, much of the technology—especially as manifested in large language models (LLMs)—has so far been heavily **focused on English** and a few other major languages of the world, reflecting the amount of readily available training data. This language bias introduces a corresponding cultural bias, raising the alarm about AI potentially pulling our future world further towards a **monoculture**, to the detriment of cultural and linguistic diversity.

If AI is to deliver benefits for all of humanity, it should support, and even **strengthen our cultural heritage**, be proficient in a **vast array of languages**, and possess reliable knowledge of a wide range of **community backgrounds and histories**.

Coordinated and determined action is needed to address this issue. We propose an **open, international project** to define the problem, collect and establish best practices, promote standards, develop benchmarks, facilitate data collection and archiving, and support research in the field of multicultural and multilingual AI. Such a project should involve stakeholders from AI companies, research and academia, governments and community representatives, and international institutions such as UNESCO.

Through its deliverables (described below), and through outreach efforts, the project should aspire to become a **go-to, central partner to companies and language groups alike**, to improve representation of long tail languages and cultures in AI models and applications. It should also strive to become a trusted forum and a **resource for long tail language and cultural communities** seeking to contribute local data and knowledge, for use in an open and fair manner to facilitate improved support.

Among the inspirations for this paper is the **digital journey of Iceland**, a country that has been making focused efforts to protect and strengthen its language—spoken by 350,000 people—in the face of change and challenges brought about by AI. Iceland's initiatives facilitated a fruitful collaboration with OpenAI to support Icelandic in the GPT models, demonstrating what can be achieved when a nation takes decisive action to protect its linguistic future.

Iceland's experience can serve as a foundation, forming the basis of a **playbook**, or blueprint, that other communities can adopt and adapt to help safeguard and strengthen their own heritage.

2. Making It Happen: The Deliverables

The project should aim to deliver some or all of the following:

- A set of guidelines and **best practices for sourcing and licensing** data for use in AI models, considering copyright, stakeholder concerns, and other applicable constraints.

- A shared, open repository of monolingual and **instruction tuning datasets** in various long tail languages, in standard formats, along with a set of documented best practices for collecting and filtering such data.

- A standardized, regularly updated, and versioned **suite of benchmarks**, with an accompanying **leaderboard**, to measure AI model performance across a spectrum of tasks that test proficiency in long tail languages and knowledge of diverse cultures and histories.

- A suite of benchmarks to measure **biases and toxicity in model output**, for a wide range of languages, cultures and communities. These biases may range from, e.g., grammatical preference for certain genders, to culture-specific misnomers and slurs.

- Initiatives to **sponsor and support research** in the field of diverse language proficiency and cultural knowledge in AI, through a grant program and in collaboration with universities and research institutions globally.

3. An Appeal for Action

This project presents a strategic initiative to address the danger of an “AI divide” between the haves and the have-nots in terms of linguistic and cultural representation. Our proposal is that major stakeholders and players in the AI space come together to lead a global, open effort to ensure that AI technologies evolve as inclusive tools that respect and promote cultural and linguistic diversity.

By endorsing and actively supporting this effort, stakeholders can help bridge gaps in AI representation, safeguard cultural heritage, and contribute to making the benefits of AI accessible to all communities.

4. A Case in Point: Iceland’s Digital Journey

COMMITMENT TO LANGUAGE PRESERVATION

Iceland, an island nation in the North Atlantic with a population of around 400,000, is home to a rich literary tradition rooted in the Icelandic language. With globalization and the increasing dominance of larger languages like English, the Icelandic language has come under threat.[1] Children are increasingly immersed in English through digital media and games, the number of non-Icelandic speakers is rising, tourism has surged, and new interactive technologies—primarily available in English—have become integral to everyday life.

Recognizing that the loss of their language would equate to the loss of a vital part of their identity, Icelanders have united in a mission to preserve their linguistic and cultural heritage in the digital world. Over the last two decades, the Icelandic government has made substantial investments in developing digital language resources and technologies. This effort has been supported by academia, private industry, and the public through crowdsourcing initiatives.

MILESTONES ON THE PATH TO DIGITAL RESILIENCE

In 2014, recognizing the aforementioned challenges, the Icelandic Parliament *Alþingi* unanimously passed a decisive resolution to address the status of Icelandic in the digital age. This solidified the nation’s commitment to its linguistic heritage and marked the beginning of a focused, nationwide effort to safeguard the language for future generations.

By 2018, the government had crafted a comprehensive action plan - the *Language Technology Programme for Icelandic 2019-2023* (LT Programme)[2], investing significantly in the development of language resources and key software tools. A collaborative partnership was established among Icelandic universities, public institutions, and private companies to implement the plan, with government funding alongside matching contributions from the private sector.

Resources developed as a part of the LT Programme included a large monolingual corpus comprising a variety of text genres, parallel corpora for machine translation, voice recordings and transcripts for speech applications, reference implementations of spell and grammar checking, and other vital technologies. All of this was made open source under permissive licenses, allowing anyone wishing to incorporate Icelandic support in their products to use the data.

The datasets developed under the umbrella of the Icelandic Language Technology Programme proved durable and played a key role in giving Icelandic a relative advantage when it comes to inclusion in LLMs.

[1] John Henley. 2018. *Icelandic language battles threat of digital extinction. The Guardian*, UK.

[2] Anna Nikulásdóttir, Jón Guðnason, Anton Karl Ingason, Hrafn Loftsson, Eiríkur Rögnvaldsson, Einar Freyr Sigurðsson, and Steinþór Steingrímsson. 2020. *Language Technology Programme for Icelandic 2019-2023*. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3414–3422, Marseille, France. European Language Resources Association.

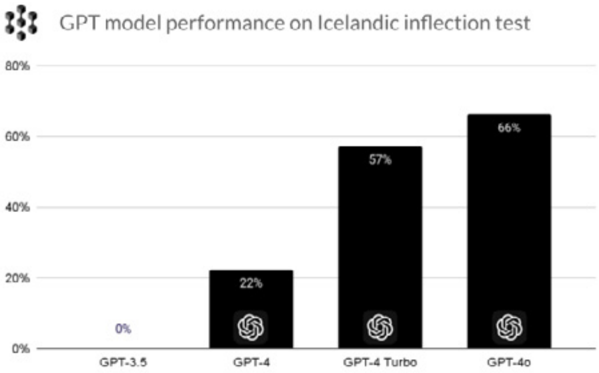
ICELAND & OPENAI: THE PROJECT STORY

Having laid the necessary groundwork, the Icelandic government recognized the next crucial step: delivering the benefits of language technology to the general public by facilitating its incorporation into products that people and companies use in daily life. Efforts were made to promote and share these resources with major tech players worldwide, particularly those behind the most popular and commonly used operating systems, apps and other software.

In May 2022, the President of Iceland, along with the Minister of Culture and Business Affairs, led a delegation on a visit to the United States to meet with some of the world's leading tech companies. The goal was to showcase the technologies and data sets developed for Icelandic, demonstrate how easily Icelandic could be integrated into these companies' solutions, and offer support for such integrations.

The most successful of those meetings marked the beginning of a collaboration between Icelandic software company Miðeind[3] and OpenAI[4] on training the GPT models, starting with GPT-4, in Icelandic, resulting in a pivotal demonstration of what needs to be done to integrate a small language into the world's leading LLMs.

The first collaboration involved training GPT-4 with ideal Icelandic prompts and completions in a process called reinforcement learning from human feedback, or RLHF. The aim was to estimate the effort needed for an LLM to understand and generate text in a language spoken by a small population, such as Icelandic. This would help in planning and guiding work for other lesser-spoken languages, showing that with focused effort, even languages with relatively few speakers can thrive in the digital age.



To date, over 4 billion words of quality Icelandic text data have been made available to OpenAI, accompanied by benchmarks for performance evaluation. Iceland's high quality monolingual data was used in GPT-4 Turbo to facilitate targeted improvement, and the same high-quality data has been used in all subsequent models.

This has helped make Icelandic a valuable test case for OpenAI, as the company has developed and experimented with methods for improving their models' performance on long tail languages.

[3] See https://mideind.is

[4] See OpenAI's case study: https://openai.com/index/government-of-iceland/

5. Thinking Globally: A Project and a Playbook

What follows is a first attempt at synthesizing the Icelandic approach as an inspiration and a baseline for a project and a playbook, to address similar requirements and concerns in other linguistic and cultural communities globally.

This paper posits that a unified effort is needed to preserve and revitalize the world's cultural diversity in AI and other digital technologies. We propose that this effort include standardizing methods for collecting, structuring, and evaluating data for training and benchmarking AI language models. This should be done in open collaboration and consultation with technology companies to maximize the use and utility of this data. Research into methods to promote linguistic and cultural inclusion should also be encouraged and supported.

It remains to be seen whether all languages, regardless of size or language family, can be reasonably and adequately supported by leading global LLMs. Any effort to follow our playbook should start with a realistic evaluation of where a language stands with respect to the linguistic, community, and monetary resources available to make an impact. However, collecting corpora—even smaller ones—is always an important undertaking from a language preservation perspective and can serve as preparation for a future where transfer learning, synthetic data generation, and other methods for advancing low-resource languages in AI will be further along.

LOCALIZED AI BENEFITS SOCIETY

AI models that are proficient in a local language can help the community flourish in various ways—in daily life, in private enterprise, and in government. We already have evidence of this in Iceland. Through technologies such as cross-language chatbots, machine translation, and text summarization, services and applications that would otherwise only be available in English (or other major languages), can be delivered in Icelandic—and, vice versa, services that have only been available in Icelandic can be made more readily available for second language speakers, who make up 18% of Iceland's population, and tourists. For enterprises, choosing to implement internal and external business processes in Icelandic should not result in disadvantage relative to competitors who "cop out"[5] by going English-only.

An important potential application of multimodal AI models that support small languages is to facilitate real-time translation of video content and games, especially for children. Starting with automatic subtitling and evolving towards high-fidelity real-time voice translation with pitch emulation, such functionality will help children learn their native language in their formative years, aiding language acquisition.

AI can also help people with disabilities access services in their native language, with features such as image recognition and description, text simplification, speech-to-text, text-to-speech, and spelling correction.

[5] Or "defect", in terms of the game theory of the Prisoner's Dilemma.

These and other opportunities to strengthen language and culture through innovative technological solutions are the driving force behind our effort to create a blueprint for the inclusion of long tail languages in AI models.

HIGH-LEVEL SUCCESS FACTORS: LESSONS LEARNED IN ICELAND

We believe that the relative success in integrating Icelandic into the world's most powerful LLMs can and should be replicated for other languages that have historically been left behind in the development and deployment of language technology solutions. A number of factors have contributed to Iceland's progress in this area.

FUNDING ARRANGEMENTS

Goal: To support the development of necessary language resources.

For many long tail languages and communities, the local market is too small to support impactful development of language resources and technologies as part of revenue-generating applications and services. It is therefore necessary to secure other sources of funding and adopt a more centralized approach. The Government of Iceland recognized this and was ready to fund its LT Programme(s) with a long-term view, based on requirements analyses and work estimates from experts in the LT field.

Designated core subprojects of the LT Programme were funded 100%, and a matched-contribution fund was established for product development that incorporated Icelandic language technology, where companies could apply to receive a reimbursement of 50% of eligible development cost.

FREE OPEN SOURCE SOFTWARE / PERMISSIVE LICENSING

Goal: To facilitate the broadest possible support for the Icelandic language in tech products.

The Icelandic government made the key decision that all deliverables of the LT Programme would be published under open, permissive licenses.[6] The licenses allow the deliverables to be integrated into third-party commercial and non-commercial applications and services, free of charge.

CROWDSOURCING

Goal: To collect quality data in the most cost-efficient way possible, and to give the public a say in how models are aligned.

Public involvement and support have been important throughout the LT Programme. A crowdsourcing initiative was launched, with the President of Iceland as patron, encouraging the public to donate their speech to the Programme via a website. This initiative was a great success, raising awareness and resulting in thousands of hours of voice samples that made it easier to develop speech applications in Icelandic.

[6] These include CC-BY, Apache and MIT.

In addition to voice samples, preparations for a crowdsourcing campaign to help identify and mitigate biases and toxicity in Icelandic text are underway. This is an important initiative as it allows the public to weigh in on what constitutes toxic speech, various kinds of bias, and other types of discourse that language models should be trained to avoid perpetuating. Such decisions should ideally not be left up to tech companies alone.

COOPERATION WITH WRITERS' UNION

Goal: To ensure that diverse text corpora are available under permissive licenses, and to address stakeholder concerns regarding the use of text under copyright.

As a limited amount of Icelandic text is available digitally, and a small fraction of that is literary text, it was important to find ways to incorporate as large a proportion of such text into open training corpora as possible. However, authors (and publishers) are understandably skeptical about releasing their texts into the open digital domain. Their concerns are both about loss of sales due to copying and copyright infringement, and about the potential ability of LLMs to generate text that imitates their style.

Within the LT Programme, a dialogue was set up with the Writers' Union of Iceland and a compromise was agreed upon: Literary text (i.e. book manuscripts) contributed by authors would be broken into fragments of approximately 500 words each, and these fragments would be included in corpora in random order. Also, author names would by default be excluded from training data, preventing language models from associating texts and styles with individual authors.

This solution enables language models to be trained on a significantly larger quantity of high-quality literary Icelandic text, but prevents the recreation of entire works or substantial portions thereof, as well as the generation of text that imitates the style of particular named authors.

QUALITY, SUSTAINABILITY AND MAINTENANCE

Goal: To future-proof data and applications in a rapidly changing technical environment.

High quality, comprehensive, and accessible resources enable effective development and deployment of digital tools that support the Icelandic language. The selection of highly compatible, open, and future-proof standards—for example for data formats and programming languages—was important to ensure the sustainability of the LT Programme. Testing, validation, and continuous updating of language resources are also necessary to meet evolving technological and linguistic needs.

[7] See the crowdsourcing website Comment Analysis https://www.xn--ummlagreining-5fb.is/ (in Icelandic)

NARROWING IT DOWN: SPECIFIC STEPS ON THE ROAD TO INCLUSIVE LLMS

What follows are some concrete suggestions for how language communities might consider investing their money and efforts in order to achieve AI readiness. Each suggestion is paired with an appeal to model developers for their input, guidance, and, in some cases, increased consideration for language diversity when designing and building their models.

ESTABLISHING A THOROUGH DATA CLEANING PIPELINE FOR WEB-CRAWLED DATA

For many languages, the largest available repository of digital text data is the World Wide Web. However, a large portion of that data is of inferior quality and can hurt model performance if used indiscriminately in training. Developers of large language models have various tools at their disposal to clean and filter data but, for optimal results, knowledge of the particulars of each language is needed. We therefore recommend that language communities build their own web-crawled corpora that have been thoroughly cleaned and filtered using state-of-the-art methods.[8]

Considerations for model developers:

Tools and best practices for **automatic quality measurement and filtering** of text, adaptable to various long tail languages should be documented and made available. We note that for smaller languages, web-crawled corpora such as CommonCrawl tend to be of relatively lower quality than for major languages. Aggressive filtering is often required to extract the most usable training corpora.

OTHER DATA SOURCES

LLMs are mostly trained on monolingual corpora, often sourced en masse from the web. Creating a higher quality, curated corpus of digital monolingual data of various genres, such as the Icelandic Gigaword Corpus,[9] is very valuable and a good use of resources. Furthermore, investing in good Optical Character Recognition (OCR) technology for a language will enable the digitization of data that can be used for training models.

AI models are increasingly becoming multi-modal, meaning they can learn from other types of data besides text, such as audio and video data.

Considerations for model developers:

A list of (minimum) requirements and desirable attributes for the type, format, mix and amount of data required to meaningfully contribute to the training of multilingual and eventually multimodal AI models.

DATA LICENSING

A crucial component of any (open) data collection effort is ensuring that deliverables are published under permissive licenses, such as CC BY, that allow for them to be used for training AI models, both for commercial and non-commercial purposes.

[8] See A Warm Start and a Clean Crawled Corpus - A Recipe for Good Language Models
to learn about Iceland's approach.

[9] See https://igc.arnastofnun.is/

Considerations for model developers:
Cooperation on establishing a set of guidelines and **best practices for sourcing and licensing** data for use in AI models, considering copyright and other applicable constraints.

BENCHMARKS FOR LANGUAGE PERFORMANCE

Establishing benchmarks for model capabilities is essential for evaluating and improving their performance. These benchmarks provide standardized metrics to assess how well the LLMs understand and generate a particular language. The Icelandic benchmarks developed by Miðeind thus far have given guidance to OpenAI development teams and confirmed gradual improvement with each update of the GPT models. They have also provided valuable insights into the amounts and types of data needed to enhance the performance of a large language model in a small language like Icelandic.

If a sufficient machine translation tool already exists for a particular language, some benchmarks can be generated via machine translation, usually from existing benchmarks for English (e.g. ARC[10], Belebele[11], and MMLU[12]). Depending on the quality of the tool and the nature of the benchmarks, human quality control and post-editing may be required. Other benchmarks (e.g. WinoGrande[13]) can be localized to fit a particular language and culture, with human input. Translation and localization of existing benchmarks can be especially useful when comparing performance on similar tasks across languages.

- Some benchmarks will be language-specific, emphasizing the particular aspects of a language and culture, and what sets it apart from others.
- Even though certain benchmarks are language-specific it is still useful to seek inspiration from the structure and content of other language-specific test sets.
- Benchmarks need to be challenging to be helpful in LLM development. The purpose of benchmarks is to identify areas for improvement and set aspirational goals, not to confirm that an LLM has already mastered the subject at hand.
- Leaderboards are useful for aggregating benchmarking data in one place and helping foster comparison and competition.[14]

[10] Abstraction and Reasoning Corpus by François Chollet, see https://paperswithcode.com/sota/common-sense-reasoning-on-arc-challenge

[11] Belebele is a multilingual reading comprehension dataset, see https://arxiv.org/abs/2308.16884

[12] Massive Multitask Language Understanding, see https://arxiv.org/abs/2009.03300

[13] “An Adversarial Winograd Schema Challenge at Scale”, see https://winogrande.allenai.org/

[14] See e.g. Miðeind’s Icelandic LLM leaderboard: https://huggingface.co/spaces/mideind/icelandic-llm-leaderboard

Considerations for model developers:

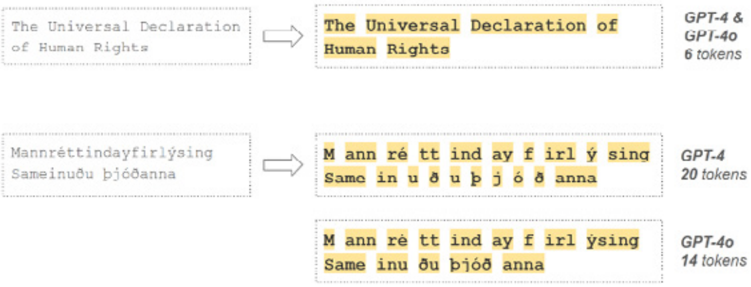
The ability to measure the current state of the art in terms of lesser-resourced language understanding and generation is a critical success factor. Standardized benchmarking methodologies and tools in this area, and compliant benchmarks and tests, need to be researched and developed.

Ideally, such a benchmark suite would become an expected and standard part of the reporting of LLM performance, joining or being integrated into established references such as MMLU, HellaSwag[15], TruthfulQA[16], etc.

Furthermore, a prescribed standard way to calculate an **aggregated performance score** from the benchmark suite, taking into account both the average performance across languages and the variance between languages, would be of great value. To obtain a high score, a model would need to perform uniformly well across languages.

TOKENIZATION

Text that is input into and output from language models is typically tokenized for processing. This means that it is segmented into fragments or subwords (such as “frag”, “ment”, “sub”, “word”), and each distinct fragment is numbered. The fragment vocabulary is determined by statistical methods before the model is trained and remains fixed thereafter. A typical LLM vocabulary contains around 50,000-100,000 unique fragments.


The Universal Declaration of Human Rights
The Universal Declaration of Human Rights
GPT-4 & GPT-4o 6 tokens
Mannréttindayfirlýsing Sameinuðu þjóðanna
M ann ré tt ind ay f irl ý sing Same in u ð u þ j ó ð anna
GPT-4 20 tokens
M ann ré tt ind ay f irl ýsing Same inu ðu þjóð anna
GPT-4o 14 tokens

Tokenization efficiency for Icelandic has improved significantly in newer GPT models but still lags behind English.

Vocabularies have tended to be heavily biased towards English text, meaning that fragments fairly common in English (and using the English alphabet) are likely to get their own vocabulary entry, while even common fragments from languages like Japanese or Hindi would not make the cut as distinct vocabulary entries, let alone fragments from less common languages. This makes input and output sequences longer for these languages. Since dominant AI model architectures typically require an amount of calculation that is proportional to the square of sequence length, the inference cost for these languages increases significantly as a result. This has an impact both on model performance and costs of usage for underrepresented languages.

[15] HellaSwag tests natural language inference by common sense reasoning. See https://arxiv.org/abs/1905.07830

[16] TruthfulQA measures how models mimic human falsehoods. See https://arxiv.org/abs/1905.07830

Considerations for model developers:

Miðeind's experiments with OpenAI's models for Icelandic highlighted some key opportunities for improvements, such as better support for rare alphabetic characters and in tokenization efficiency. The resulting back-and-forth with OpenAI's developers helped improve multi-language support in the subsequent GPT model releases, especially in GPT-4o, where a new, larger and more multilingual vocabulary was adopted. Further research and testing are needed to identify optimum token vocabularies that support multilingual processing without compromising too much on English capabilities.

BIASES

The work done so far on developing multilingual LLMs has focused heavily on linguistic capabilities, i.e. training models to both understand and produce grammatically correct output in a larger portion of long tail languages. However, going forward, the scope needs to be broadened to ensure that knowledge of local culture and history is embedded in AI technologies, and that language- and culture-specific biases and toxicity are also mitigated. This requirement applies not only to text and voice modalities, but to images and video as well.

Language-related bias can, e.g., involve grammatical indicators that are not present in English. As an example, in Icelandic, adjectives are grammatically gendered—masculine, neutral or feminine—and LLMs have been shown to be more likely to assign feminine gender to certain adjectives associated with negative personality traits.[17]

As for cultural bias, AI models should possess knowledge of diverse cultural and historical facts and norms, and be able to base their responses on this knowledge without hallucination. This may often be conditioned on the language of the interaction[18] but that may not be appropriate in all cases since significantly different cultures can share a common language.

Language communities can assist in this task by ensuring that the datasets they build reflect cultural norms and history as much as possible. Furthermore, comprehensive benchmarks that test for the presence of bias, as well as general knowledge of a culture's history,[19] need to be constructed.

[17] Sólmundsdóttir, A., Guðmundsdóttir, D., Stefánsdóttir, L. B., & Ingason, A. K. (2022). *Mean Machine Translations: On Gender Bias in Icelandic Machine Translations*. In N. Calzolari, F. Bechet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), 2022 Language Resources and Evaluation Conference, LREC 2022 (pp. 3113-3121).

[18] As an example, the query "Who discovered America?" might in Icelandic ("Hver fann Ameríku?") be answered with "Leifur Eiríksson", while in English the answer is probably "Christopher Columbus" (and "Cristóbal Colón" in Spanish).

[19] For an example, see Miðeind's Icelandic Wiki QA benchmark: https://huggingface.co/datasets/mideind/icelandic_wiki_qa

Considerations for model developers:

Cultural norms regarding acceptable speech, toxicity, and biases are obviously quite diverse across and even within communities, and models need to be flexibly adapted to this. Furthermore, it should be verified that general “red teaming” safety constraints regarding harmful output (instructions for building bombs, etc.) are applied in lesser-resourced languages as well as in English.

Methodologies for creating effective benchmarks to measure biases and toxicity across languages and cultures need to be developed in collaboration with the diverse communities, shared openly, and subsequently maintained.

6. Conclusion: A Call to Action for a More Diverse AI

The preservation and revitalization of linguistic and cultural diversity in AI is not just a technological challenge—it's a moral imperative. As we face a new era shaped by artificial intelligence, we must ensure that this transformative technology serves all of humanity, not just a dominant majority. The lessons learned from Iceland's digital journey, combined with the proposed international project and playbook, offer a roadmap for communities worldwide to safeguard their heritage in the digital age. By fostering collaboration between AI developers, language communities, and policymakers, we can create a future where AI enhances rather than diminishes global cultural diversity. **The time to act is now**. Together, we can shape an AI-powered world that celebrates diversity, preserves languages, and ensures that all voices are represented in the digital landscape of tomorrow.