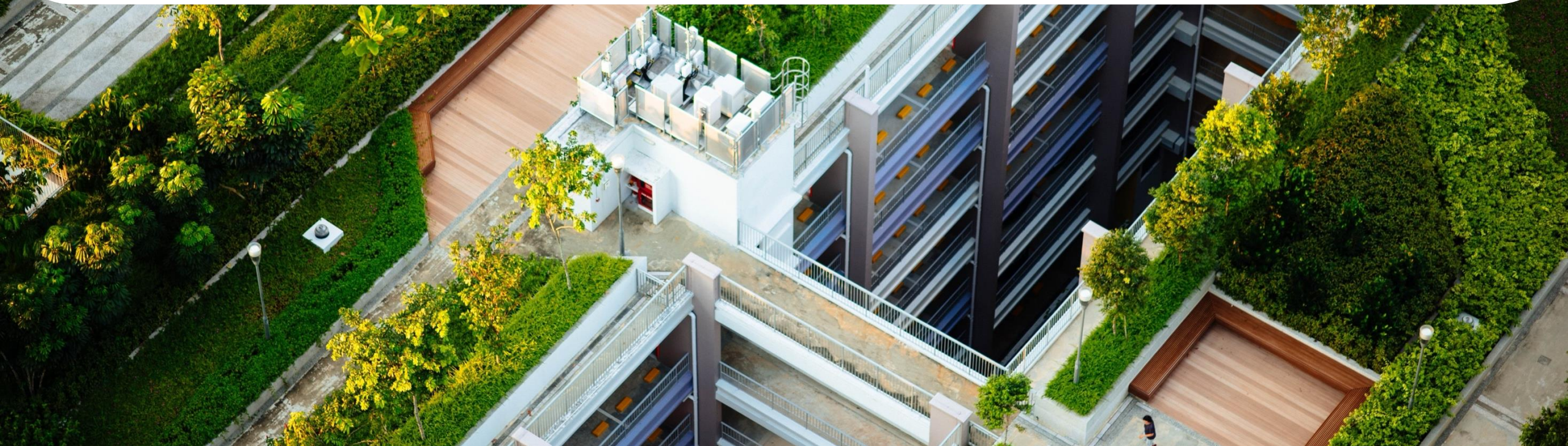
Rwanda
Centre for the
Fourth Industrial
Revolution

Bridging the Gap – Enhancing LLMs for Low Resource Languages

Dr. Jan Nehring
jnehring@c4ir.rw



Jan Nehring, Ph.D

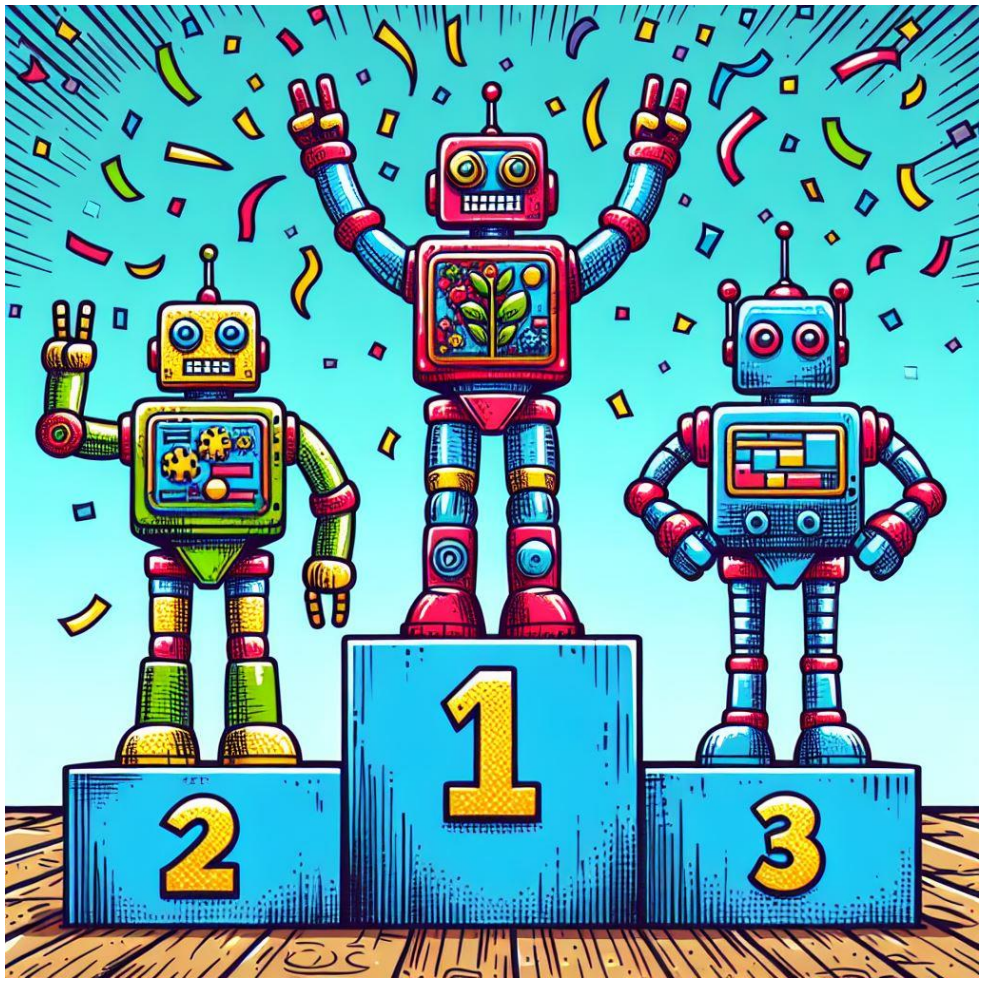
Development advisor for GIZ and C4IR Rwanda

- Based in Kigali, Rwanda
- 10 years experience as a researcher at German Research Center for Artificial Intelligence and other universities
- PhD in language technologies / chatbots

Performance of LLMs on Low Resource Languages

In Kinyarwanda (other languages similar)

LLMs	Performance
Commercial LLMs (OpenAI GPT-4, Google Gemini, …)	high
Open Source LLMs (e.g., LLama3)	low



Why?

How to improve LLMs for Low Resource Languages?

Everyone wants their own LLM

Training new LLMs is difficult and expensive

LLama3 400B: estimated 50M USD compute and a team of 50 experts

My take: We need creative approaches how to make LLMs speak low resource languages.

Occiglot - new open source large language models for Europe released

04/11/2024 | Press release | Language & Text Understanding | Foundations of Systems AI | Speech and Language Technology | Berlin | Darmstadt

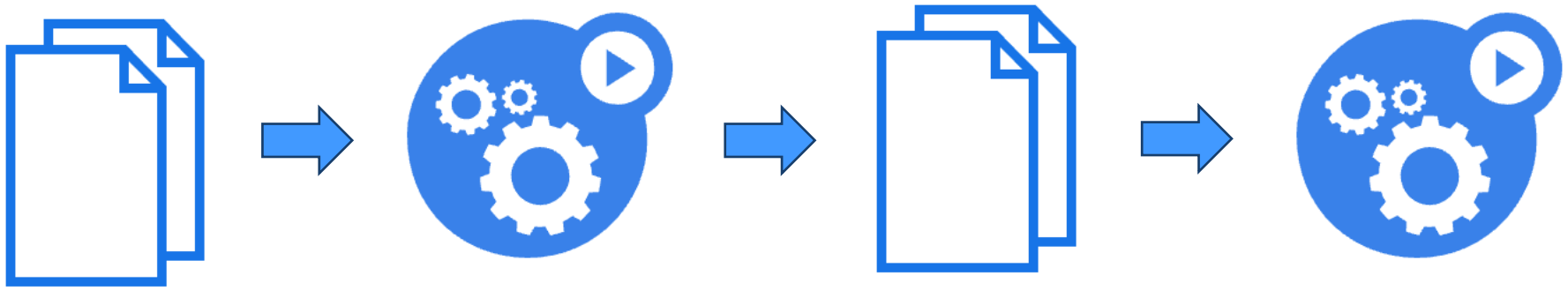
InkubaLM: A small language model for low-resource African languages

Atnafu Lambebo Tonja[1,2,*], Bonaventure F. P. Dossou[1,3,*], Jessica Ojo[1,3],
Jenalea Rajab[1,5], Fadel Thior[1], Eric Peter Wairagala[1], Anuoluwapo Aremu[1],
Pelonomi Moiloa[1], Jade Abbott[1], Vukosi Marivate[1,4], Benjamin Rosman[1,5]

WIZ.AI Unveils Southeast Asia's First Foundation Large Language Model (LLM) for Bahasa Indonesian, Amplifying Regional Representation in Mainstream AI

How to train a Large Language Model?

Its all about data and compute



Collect a large corpus of data in the respective languages

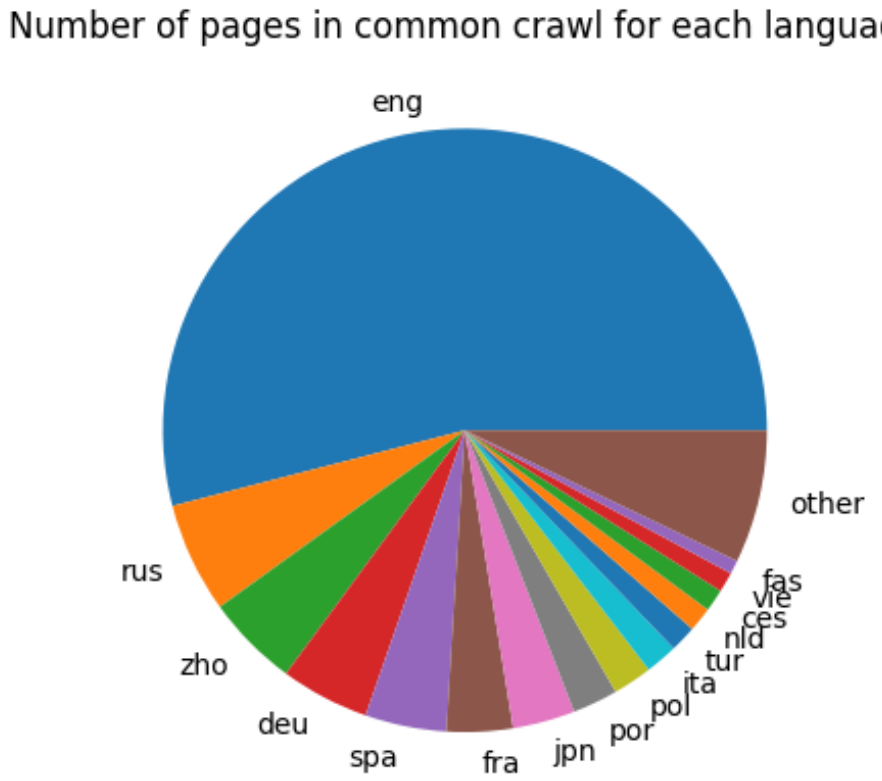
Train the language model

Create a dataset with instructions

Train the model to follow instructions

Common Crawl is the main data source for nearly all open source LLMs

Number of pages in CommonCrawl for each Language



Kinyarwanda accounts for only 0.000886% of the documents in the CommonCrawl.

Peak Data

/ “We’ve achieved peak data and there’ll be no more,” OpenAI’s former chief scientist told a crowd of AI researchers.

By Kylie Robison, a senior AI reporter working with The Verge's policy and tech teams. She previously worked at Fortune Magazine and Business Insider.

Dec 14, 2024, 2:34 AM GMT+2

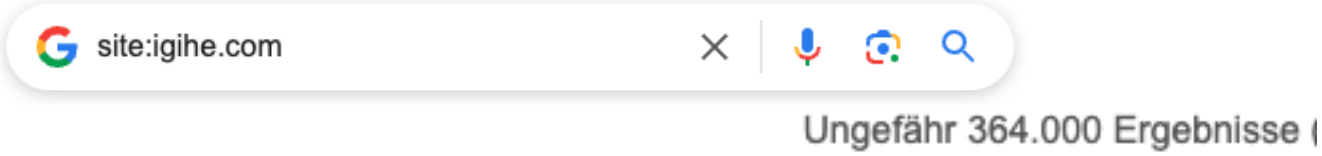
13 Comments (13 New)

- Ilya Sutskever, 12/2024
- They have crawled the whole internet, they do not know where to find more training data for LLMs.

Experiment

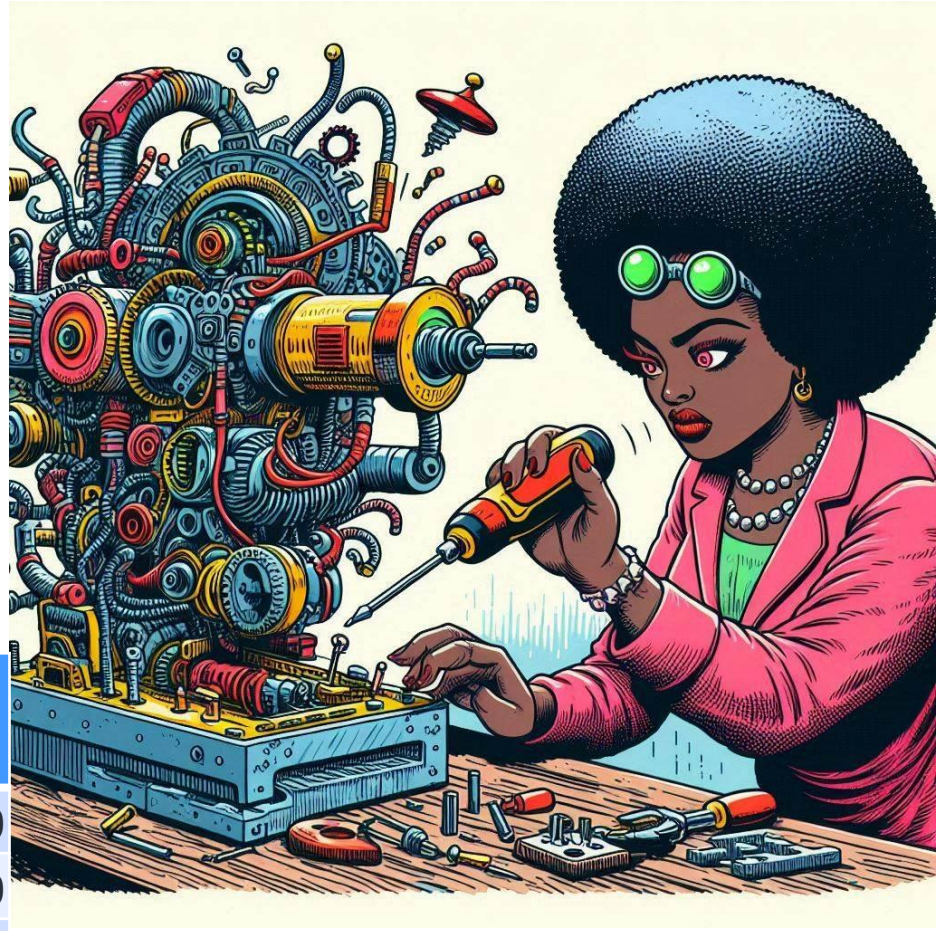
What is the coverage of low resource languages in Common Crawl?

- Get a list of pages from Common Crawl Index
- Count number of pages from each domain
- For the same domains, I performed a Google search



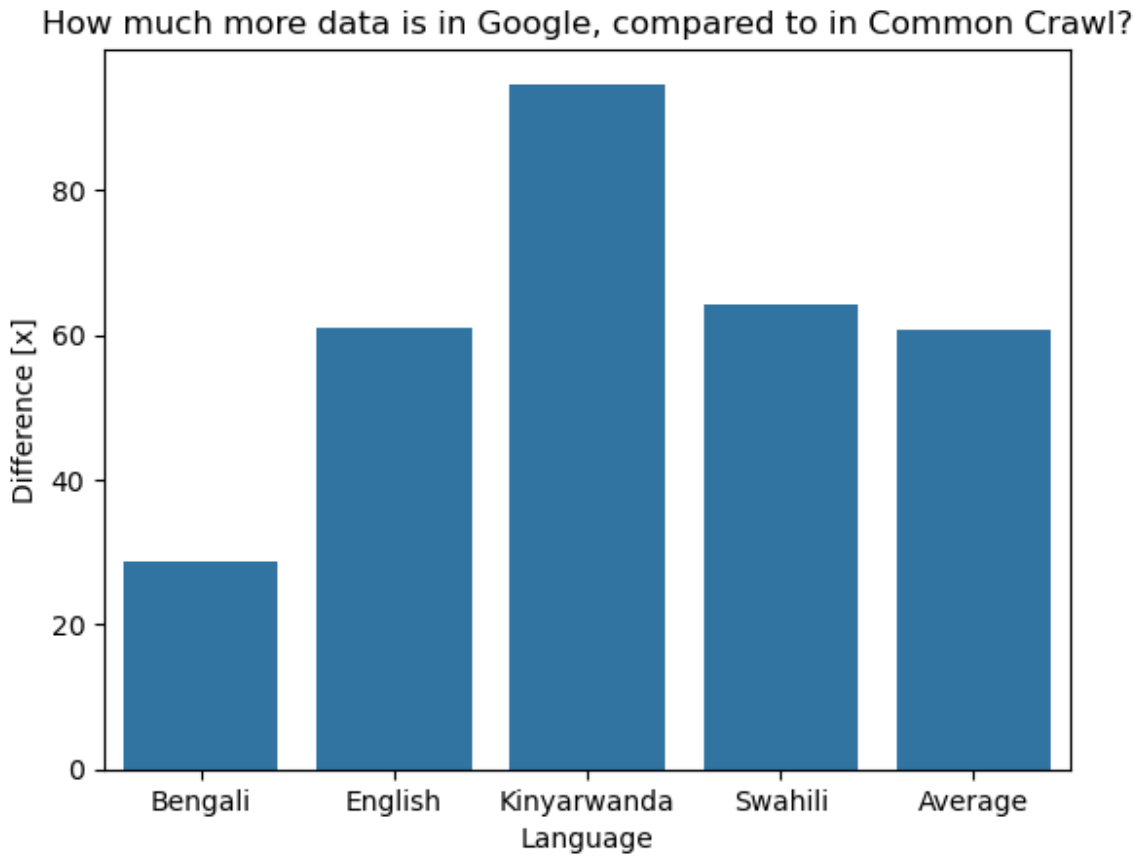
Ungefähr 364.000 Ergebnisse (

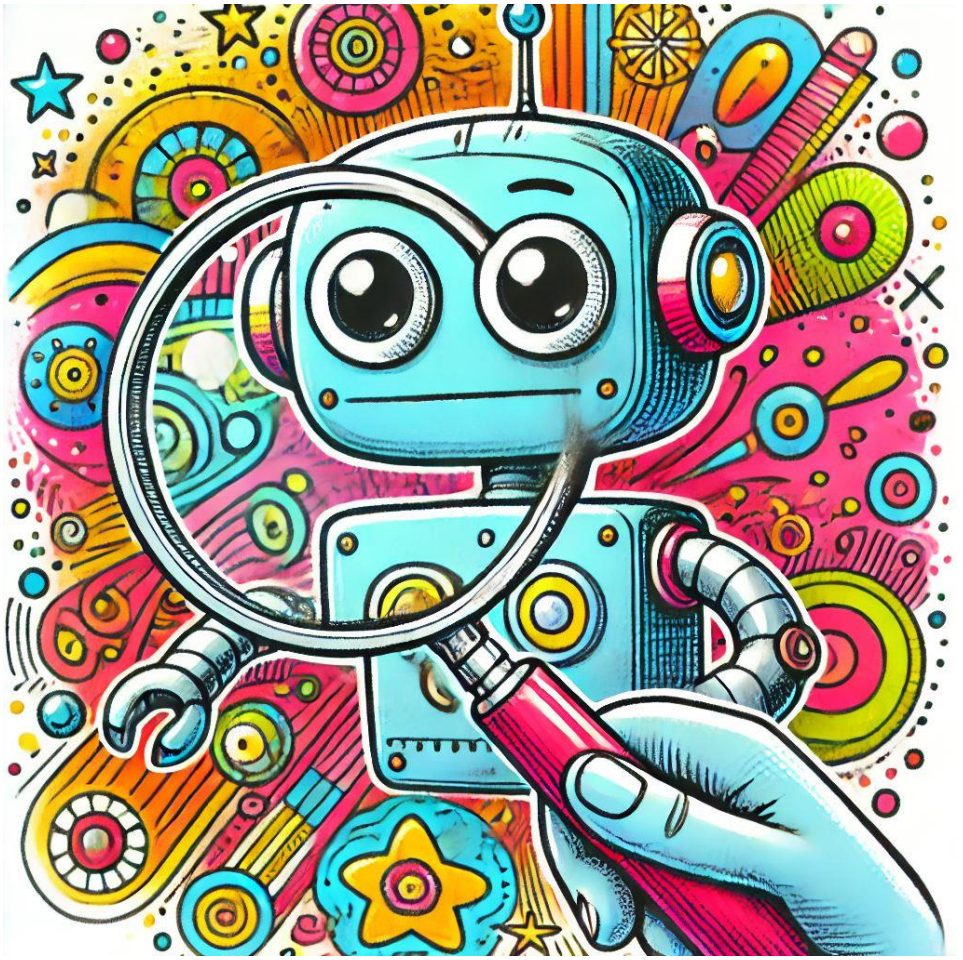
Domain	Language	Number of pages in Common Crawl	Number of pages in Google
igihe.com	Kinyarwanda	500	361.000
banglanews24.com	Bengali	22.635	957.000
radiotadio.co.tz	Swahili	990	16000
56 domains in total			



Analysis

Google contains about 60x more data than the Common Crawl.

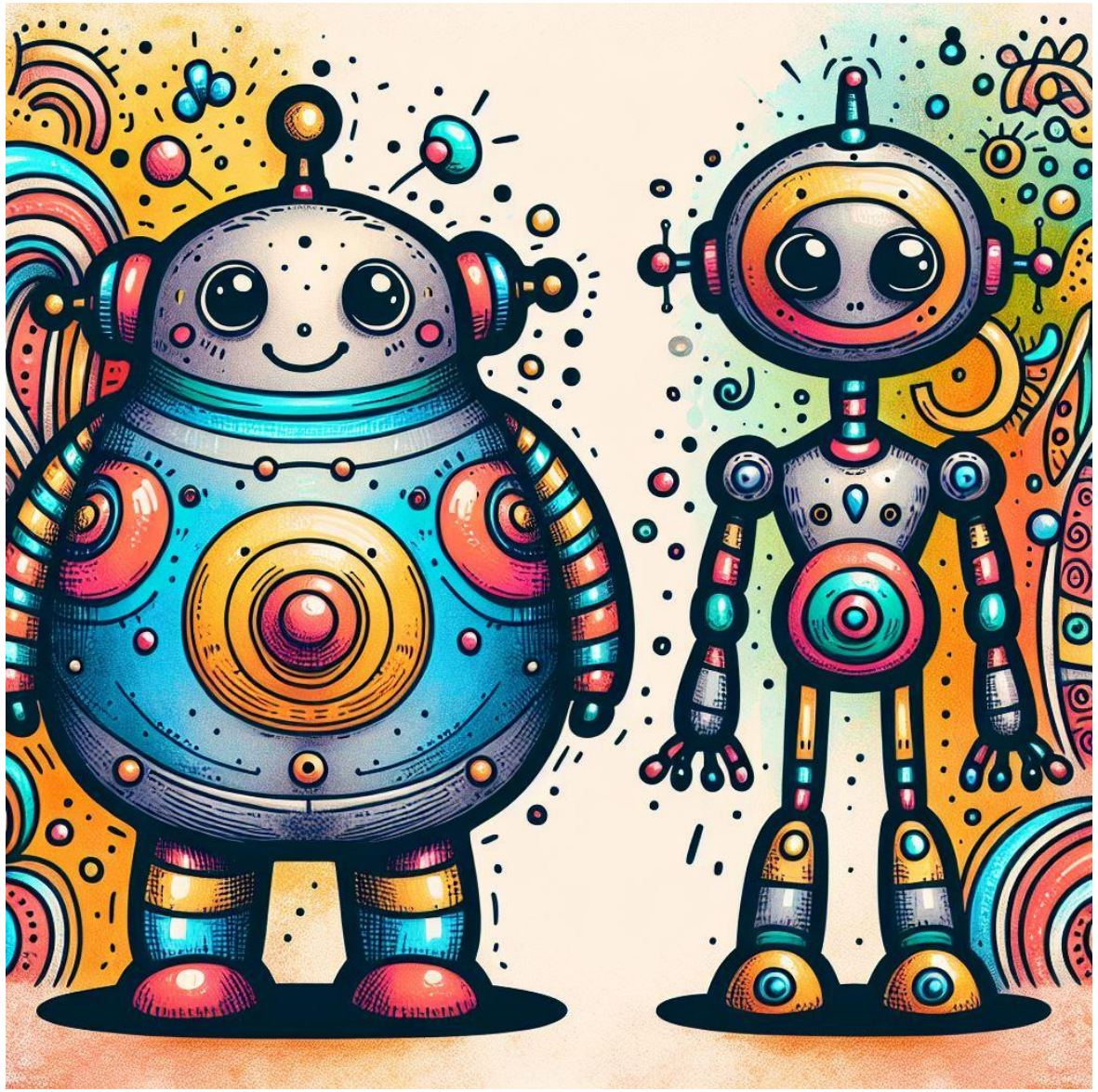




My Hypotheses

1) I find it plausible that the performance difference in commercial and open source LLMs stems solely from the extended training data.

2) If good data for low resource language existed, Open Source LLMs (Meta Llama, Mistral, …) would be trained on this data and speak Low Resource languages automatically.



Novel data crawling method

How to crawl certain languages only from the web

- It is very challenging to crawl the whole internet
- Novel method: Crawl only certain languages
  1. Get all URLs in certain language from Common Crawl
  2. Crawl all URLs within these domains
  3. Use language detection to stop crawling for undesired languages

Preliminary results

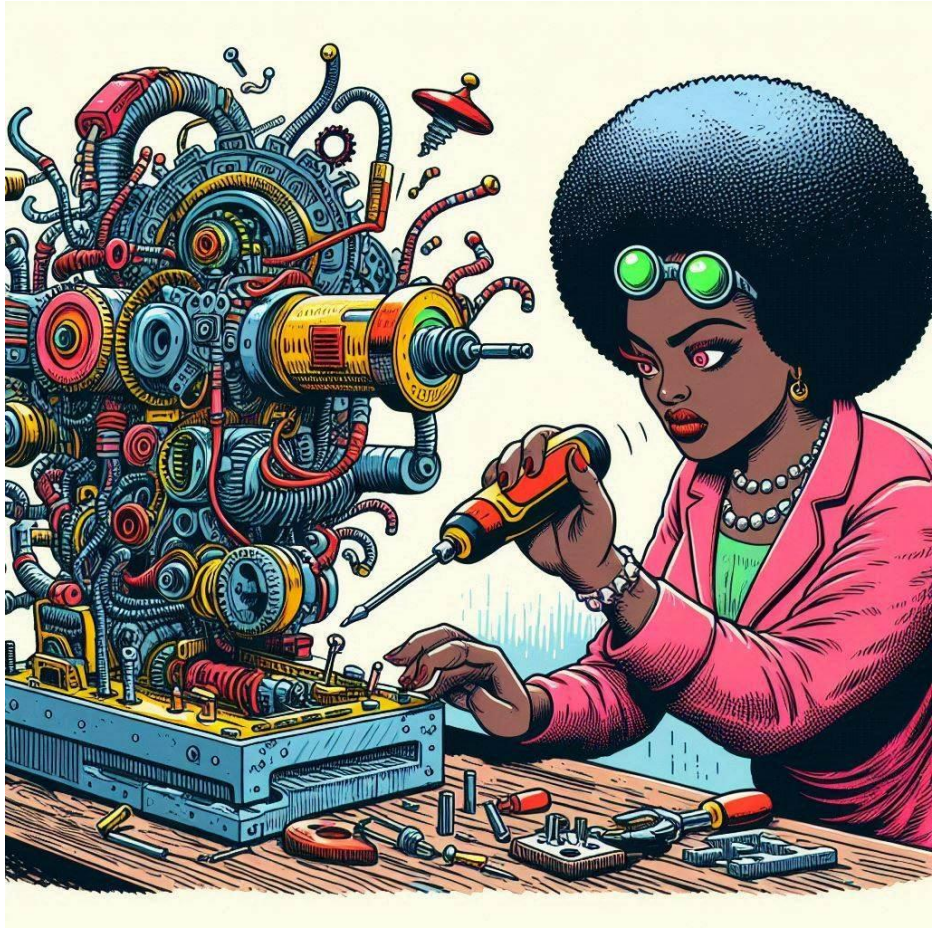
We crawled a corpus of 35 mio words of Kinyarwanda words in ~6 weeks.

Source	Number of pages
Latest common crawl	58,557
All common crawls	577,411
Crawled by our method	6,150,378

Next steps

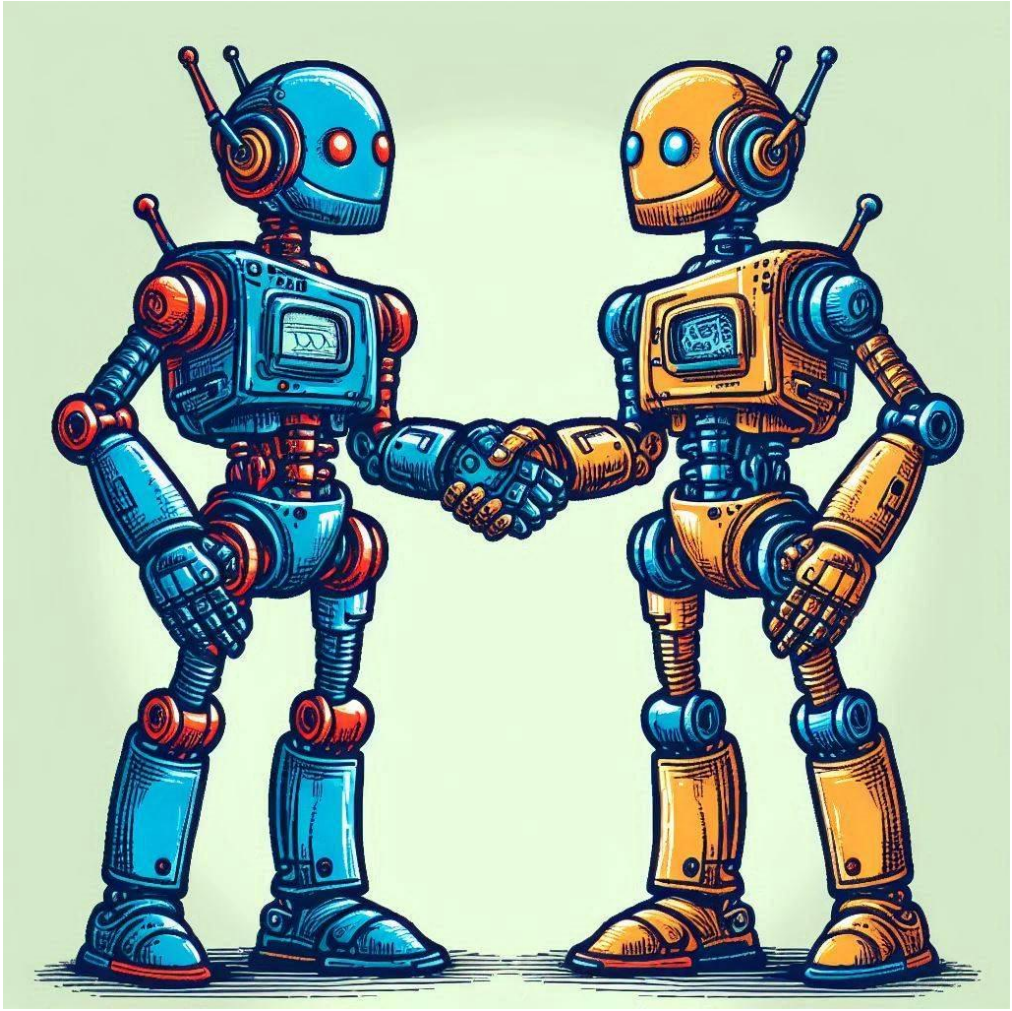
How to create a large corpus of unlabeled text with many low resource languages?

- Validate and debug Kinyarwanda crawl
- Scale to other languages (as automatically as possible)
  - African languages
  - Other Low Resource Languages
- Improve language detection
- Add more parallelization, add more compute power
- Prioritize languages: Which low resource language is more important than another?



Summary

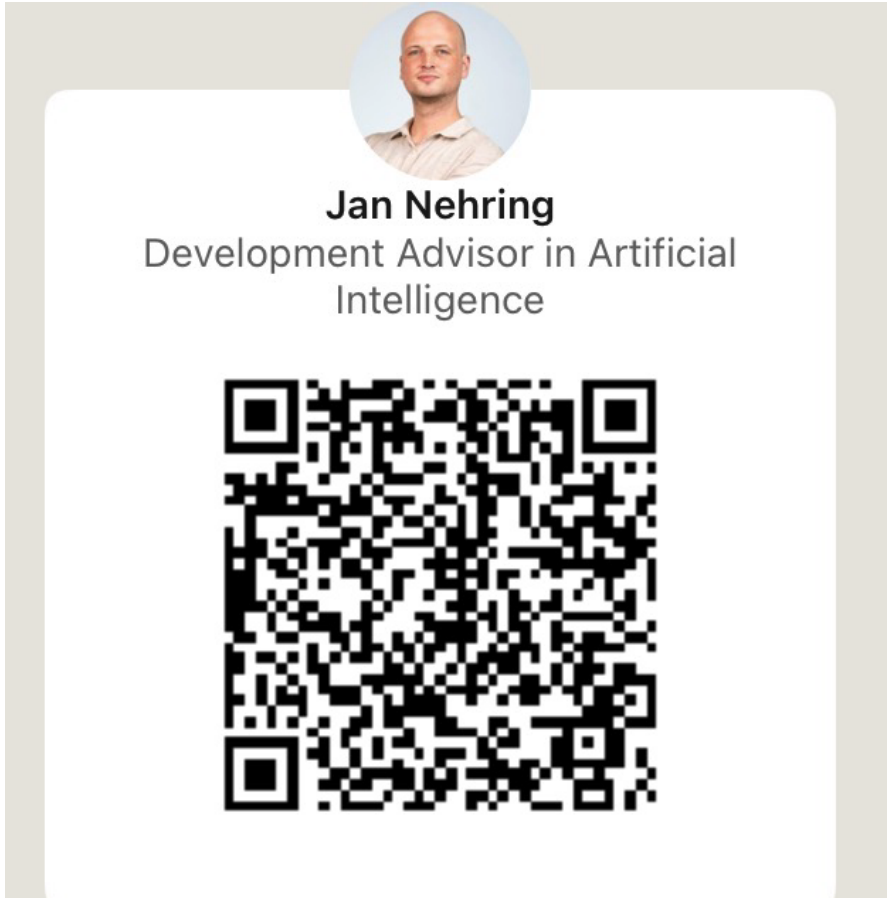
- There is much more data out there than we thought
- We can crawl this data relatively easy
- It would make sense to crawl this for many low resource languages
- It might be possible that big tech includes additional data in their LLM training so that we get high quality LLMs for many low resource languages for free.



Lets stay in touch

This is an open source, community effort – I am looking for contributors

jnehring@c4ir.rw



Linked In