

The Global South has a Problem of Large Language Models and Small Corpora of Texts

Glenda Cox: glenda.cox@uct.ac.za

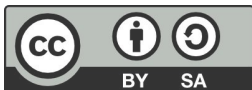
Kathryn S. Kure: kathryn@datamyna.com

Nomvuyo Mgoqi: mvuyo59@gmail.com

Jonathan A. Poritz: jonathan@poritz.net

UNCTAD classification
of economies

- Developed economies
(Global North)
- Developing economies
(Global South)



This work, except where otherwise indicated, is by the
above authors and is released under a Creative
Commons Attribution-ShareAlike 4.0 license.

[UNCTAD Classification of Economies](#)
Specialgst, CC BY-SA 4.0, via Wikimedia Commons

How LARGE is enough?

1. What do LLMs need?
2. How large is large?
3. What happens when there is a small corpus of texts?
4. Can this situation be remedied?
5. If so, how?

English to isiXhosa



Ingonyama yenza ntoni?

- ☐ a. ~~lingena~~ emanzini ingena
- ☐ b. ~~bhaha~~ iyabhabha
- ☐ c. ~~uhamba~~ iyahanba
- ☐ d. iyabaleka



Ingonyama iyabaleka

Can the deficit be addressed?



poritzj

@poritzj@mastodon.social

Recently I'm seeing folks in [#OpenEducation](#) who want to public domain-wash some specific material by passing it through the gastrointestinal tract of an LLM.

While IAmNotALawyer, I don't think this works.

And friends, keeping in mind the wonderful Audre Lorde quote that James G-G and Joy Shoemate cited in their [#OpenEd24](#) keynote

"The master's tools will never dismantle the master's house"

... do we really want to go down that path?

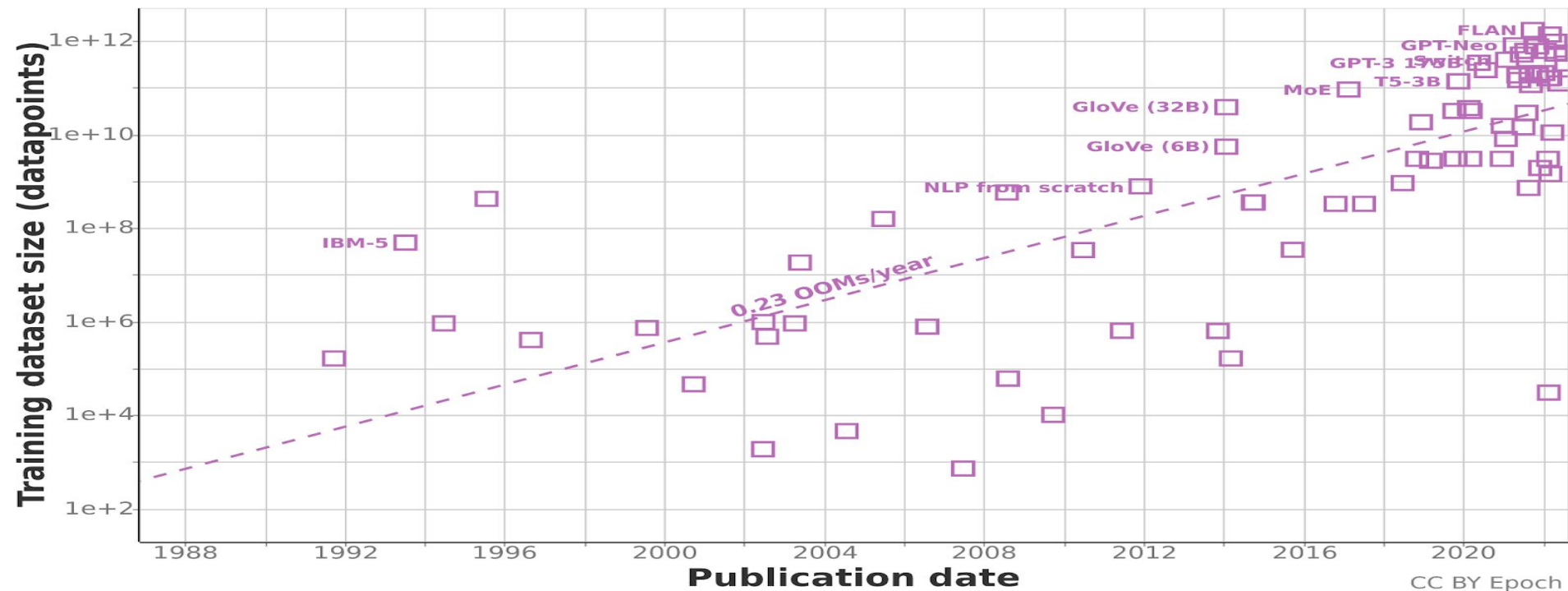
[#AI](#)

[#Alethics](#)

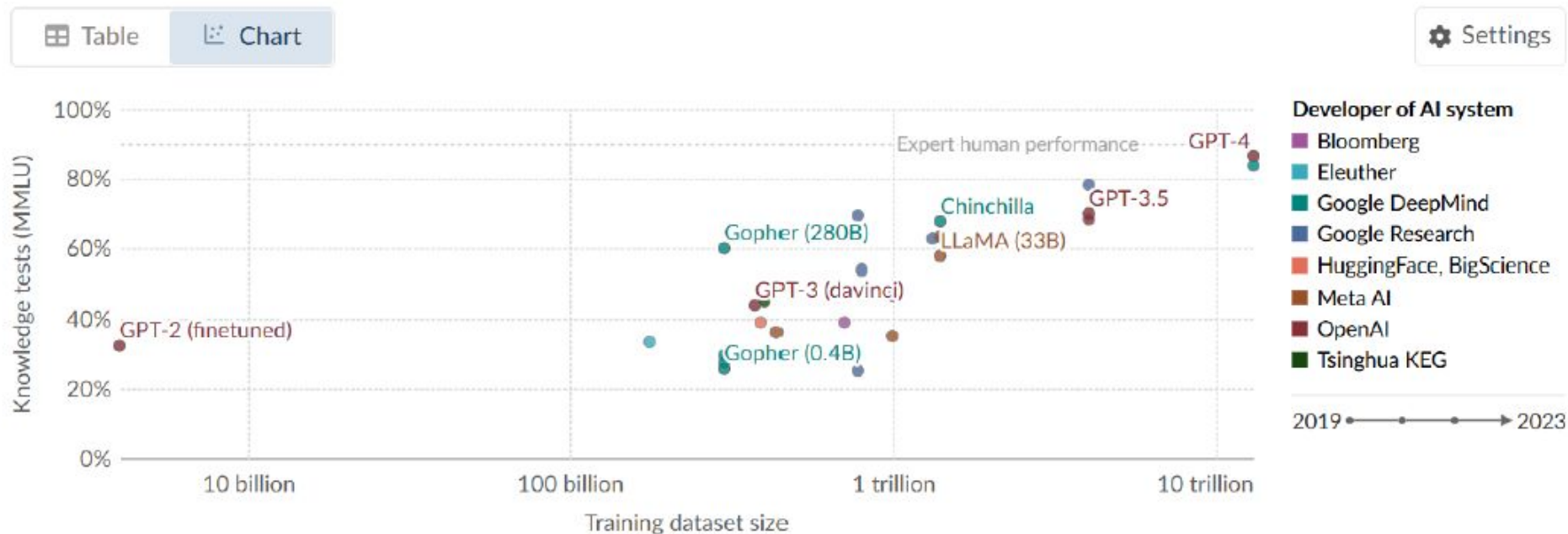
[#open](#)

...and 1 more

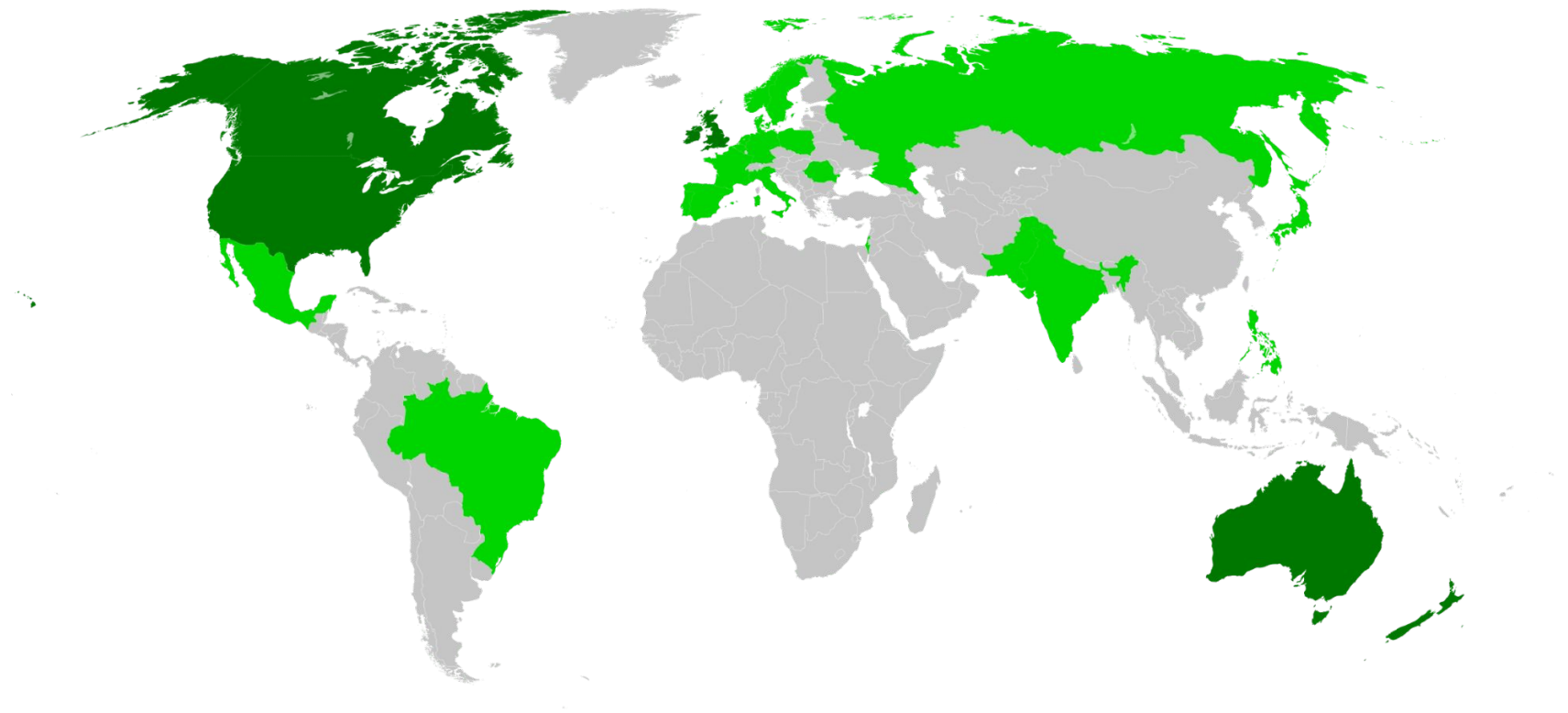
Trends in training dataset sizes



Performance on knowledge tests vs. dataset size



Inclusion, Diversity, Equity, and Access



Equitable

VC's actual use case for AI is treating workers badly

Castor and Gerard

Original:



Generated:





SUSTAINABLE DEVELOPMENT GOALS





Thank you!

References in the notes

Glenda Cox: glenda.cox@uct.ac.za
Kathryn S. Kure: kathryn@datamyna.com
Nomvuyo Mgoqi: mvuyo59@gmail.com
Jonathan A. Poritz: jonathan@poritz.net