

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

This is the script generated for the presentation 129 at OEGlobal 2024, 13 November 2024 presented by [Glenda Cox: glenda.cox@uct.ac.za](mailto:glenda.cox@uct.ac.za); on behalf of [Kathryn S. Kure: kathryn@datamyna.com](mailto:kathryn@datamyna.com); Nomvuyo Mgoqi: mvuyo59@gmail.com; [Jonathan A. Poritz: jonathan@poritz.net](mailto:jonathan@poritz.net) and is licensed [CC BY-SA 4.0](https://creativecommons.org/licenses/by-sa/4.0/) unless otherwise indicated.

SLIDE ONE: TITLE PAGE

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

Presented by [Glenda Cox: glenda.cox@uct.ac.za](mailto:glenda.cox@uct.ac.za) at OEGlobal 2024, session 129 on behalf of:
[Kathryn S. Kure: kathryn@datamyna.com](mailto:kathryn@datamyna.com)
 Nomvuyo Mgoqi: mvuyo59@gmail.com
[Jonathan A. Poritz: jonathan@poritz.net](mailto: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](https://creativecommons.org/licenses/by-sa/4.0/).

UNCTAD Classification of Economies
 Specialgsl, CC BY-SA 4.0, via Wikimedia Commons

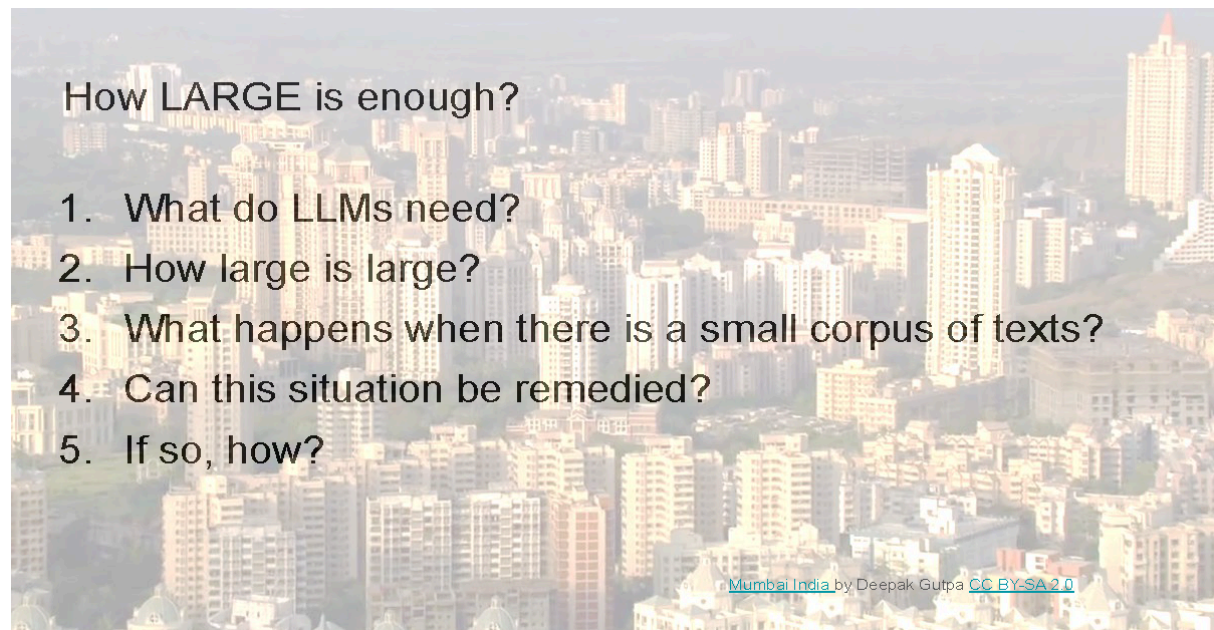
Created with mapbox

Since Open is everyone's business, and Generative AI is portrayed as a mechanism whereby to scale education for everyone everywhere, it is fundamentally problematic that large language models, which are utilised, amongst other functions, for the translation of texts, literally require extremely large corpora of texts - on both sides - to function adequately. Having an over-abundance of text on the one side cannot make up for a deficit on the other. In fact, not having sufficient data (in domains such as language and speech) for indigenous languages will have a crippling effect on the deployment of GenAI for any such language.

The relative paucity of texts in indigenous languages generally, not to mention the fact that many of these texts may not have been digitised and loaded online leads to another, more overarching concerns relating to using OER globally and aligning curricula internationally. especially if these initiatives are driven by those

countries of the Global North, as defined by the United Nations Trade and Development, which countries are depicted in blue on the map. Not only would this ensure increased dominance of already dominant languages and world-views, which could be read as a potential re-colonisation, but this would negatively impact on local indigenous languages, cultures and ways of knowing, and on the study and dissemination of indigenous knowledge systems.

SLIDE TWO: Large Language Models



But first of all, what is required by a Large Language Model? We all know that more data is better data, but how large a dataset is required? The key differentiator of current Generative AIs from prior statistical models, or even earlier versions of themselves, simply lies in how much data is used to train the models, a product both of the enormous datasets available from the internet and the data structures (such as spreadsheets or binary trees) inside the programs, which help to manipulate the stored data. Apart from this, there is nothing particularly new, philosophically or structurally, about the powerful statistical model underlying Generative AIs.

But how large are these datasets?

“Large language models (LLMs) have been growing over the years, and the model size reflects this. It is stated that GPT-4, which is 10x the size of its predecessor GPT-3, which in turn was a 15 000X increase over the previous five years, was trained on roughly 13 trillion tokens, which is roughly 10 trillion words (KDNuggets, 2023). To fit the estimated 1.8 trillion “GPT4 parameters in a giant Excel Sheet, it will need to be the size of 30,000 Football Fields or 180 square kms (more than the size of Mumbai City!)” (KDNuggets, 2023) pictured.

What happens when there is a small amount of data on the one side is that mechanical translations are inadequate to the task, especially as complexity increases. To illustrate this, we will use as an example a translation from English into isiXhosa using ChatGPT4 undertaken for a kindergarten book, which was undertaken in a project by STEAM Foundation NPC in South Africa, which project was supported by SABIER and funded by Creative Commons and Siemens Stiftung (McGuire, 2023). The project consisted of piloting a MoodleBox (the Moodle Learning Management System ported to a Raspberry Pi) at Khulisani Child Development Academy in Langa, Western Cape. The principal, Ms. Nomvuyo Mgoqi engaged with it as a possible solution for under-resourced schools with low connectivity and intermittent electricity. The reality is that GenAI produced errors in translation even for this elementary education reader consisting of very few words.

SLIDE THREE: EXAMPLE

English to isiXhosa



Ingonyama yenza ntoni?

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



Ingonyama iyabaleka

Images from [Literacy in the Mother Tongue in Africa CC BY 4.0](#)

Pratham Books Storyweaver produced this OER which consists mostly of two word sentences, such as “Lion Runs”. Despite the simplicity of the text, the machine translation still produced an error in translating one specific word. While the model did predict the more frequently used noun for “cloth”, it was not the correct term to use within that specific context. This indicates that there is an insufficiency in the current word/text bank for isiXhosa to rectify the wrong contextual usage.

Secondly, when translating comprehension assessment questions into isiXhosa, errors were produced because isiXhosa is an agglutinative language with a total of fifteen morphological noun classes, and not only prefixes but also suffixes attached to a root word differ depending on the noun class, all of which have to concord within a sentence. It is therefore evident that the statistical modelling involved does not have a large enough dataset at its disposal for accuracy with regard to the noun classes and their associated suffixes and prefixes.

Although this paper anticipated reflecting upon how 'language shapes - or is shaped by - culture, the time allocated is insufficient to address it adequately, but it is noted as a site of further inquiry.

SLIDE FOUR: CAN THE DEFICIT BE ADDRESSED?

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 though 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

We are aware that numerous organisations such as the Wikimedia Foundation are strongly committed to supporting and enabling multiple languages on the platform, though the catch always lies in recruiting sufficient numbers of unpaid volunteers for this important work. While there remains the possibility of creating a special corpus of texts, strong arguments are also being made that there should be a drive towards digitising and uploading far more texts in the various languages, assuming such texts are available.

At this stage it is important to note that there are already concerns raised to the level of legal action as to whether the training datasets that have been utilised only included openly licensed materials, or if some of these violated copyright, not helped by the lack of transparency with regards to what was datasets have been used to train the various LLMs.

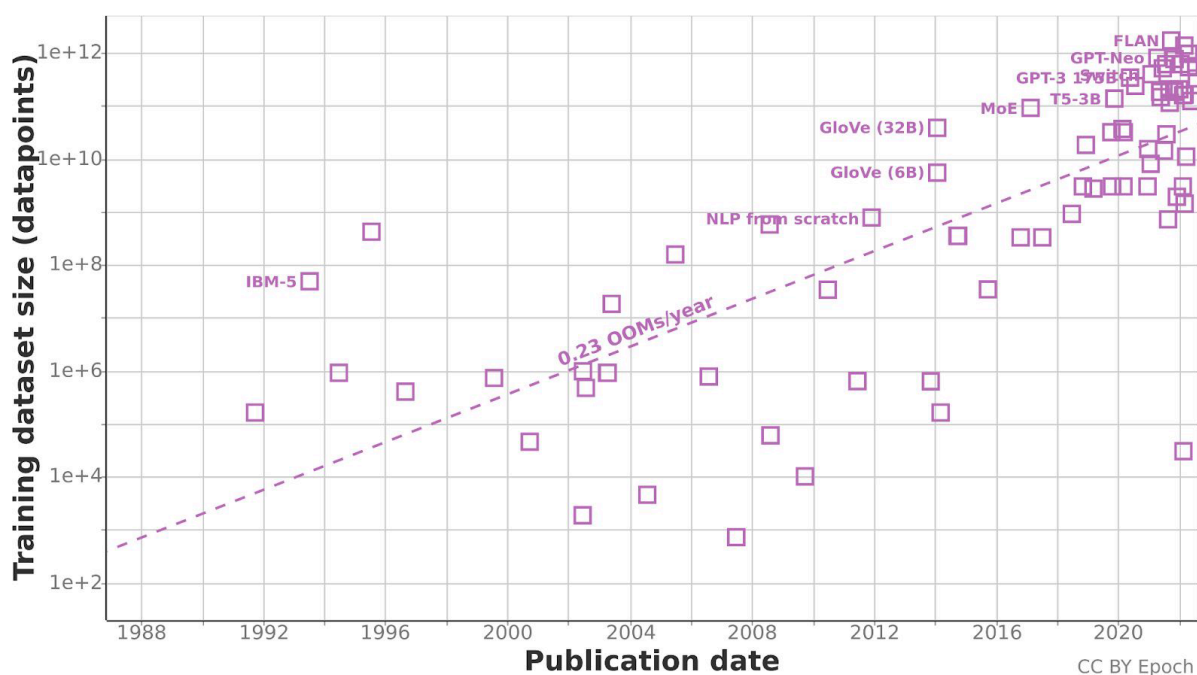
In relation to the question of possible copyright misuse, ben_ui notes drily that, “Grand theft autocomplete is my new favourite term for LLMs”. While on the topic of uploading texts, Jonathan Poritz notes it is highly problematic to attempt to “public domain-wash some specific copyrighted material by passing it through the gastrointestinal tract of an LLM” (Poritz), for as the wonderful Audre Lorde quote that James G-G and Joy Shoemate cited in their #OpenEd24 keynote says, ‘The master's tools will never dismantle the master's house’ ... do we really want to go down that path?” (Poritz, Mastodon, 2024).

The bottom-line is, practitioners from the Global South realistically fear a widening of the divide as a result of the fact that many local, indigenous languages only have a small corpus of texts online and even with everyone’s best efforts, this is not likely to change dramatically quickly, would require significant investments in time and money and may not even be achievable.

The key question is, what dataset size would be needed for such languages to have sufficient data to support translations or even their own GenAI?

SLIDE FIVE: Training dataset sizes

Trends in training dataset sizes



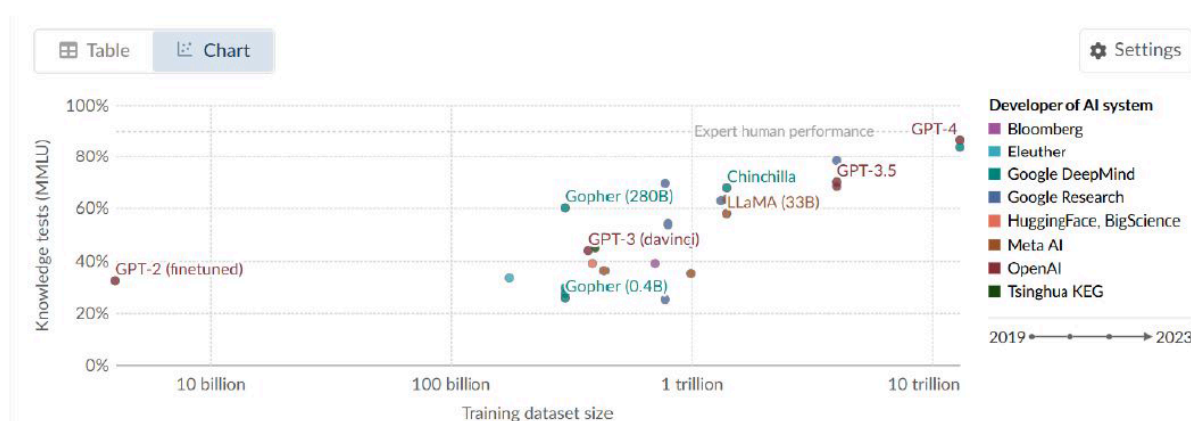
This graph represents training data set sizes of notable Machine Learning models, over time. It is important to note that while it may look roughly linear the y axis is logarithmic, meaning that it represents roughly exponential growth.

GenAI, based on its training data, puts together tokens in a way that is of a relatively high probability. If there are small amounts of training data, the model is far less effective. It does not reason, or answer a question. It is about feeding some (pseudo) random numbers into the statistical model to generate new sequences with high probability by that model. So the question then arises, not as to how much data is required for the model to work sufficiently well?

The answer is: literally gargantuan amounts.

SLIDE SIX: Performance on knowledge tests vs. dataset size

Performance on knowledge tests vs. dataset size



[Epoch \(2023\)](#) – with minor processing by Our World in Data, [CC BY 4.0](#)

This graph plots performance on knowledge tests against dataset size. It is important to note that in this graph the x axis is logarithmic, so the actual growth is logarithmic, meaning it is quite slow (log grows more slowly than any polynomial, for example; that means one needs 10 times the training data set size to get a single tick mark higher on the y axis - which is supposedly "performance", although of course that is a very problematic thing to measure!)

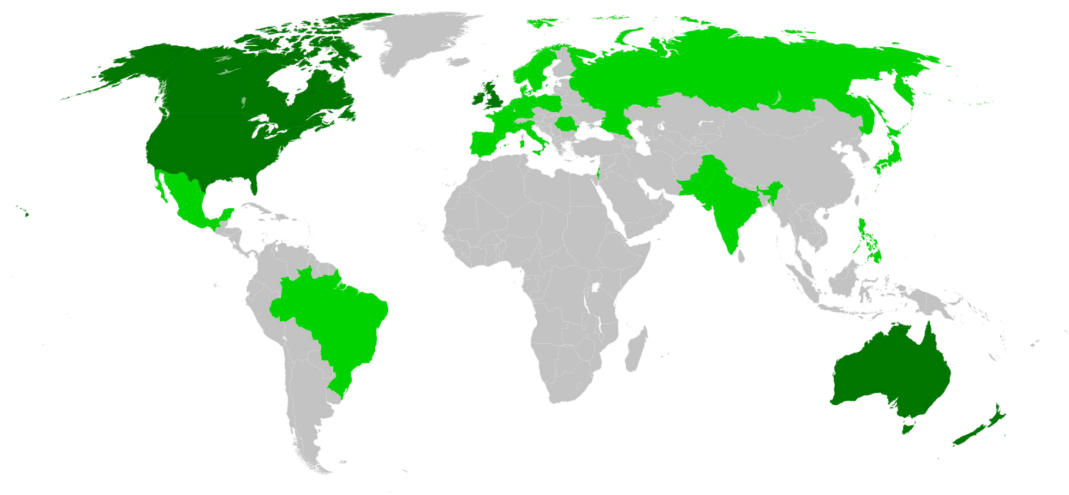
If we look at ChatGPT-2, its performance is less than 40%, which is pitiful. Hence, the beguiling results we are seeing from the current iteration of GenAI (even if they are highly problematic when a real human, particularly an expert, looks at them) only become compelling when the training datasets are vast. Languages in the Global South that do not have these huge datasets lying around that are

publicly available (even if their use as training material may be suspect from an IP point of view) are never going to be able to support beguiling outputs!

In other words, small corpora for minority languages and cultures not well represented on the scrapable internet means genAI in those languages and for those cultures will be terrible. In a world committed to scaling all the things, models trained on small corpora can't do anything really worthwhile, while those trained on large corpora are more convincingly wrong, especially when AI models are trained on recursively generated data as “indiscriminate use of model-generated content in training causes irreversible defects in the resulting models, in which tails of the original content distribution disappear” ([Shumailiv, I, et al., 2024](#)).

SLIDE SEVEN Inclusion, Diversity, Equity, and Access

Inclusion, Diversity, Equity, and Access



[Top 25 countries from which Wikipedians contribute to English Wikipedia](#) by 16@r CCBY 3.0

We are not only concerned with language and translation, but on how the use of Generative AI can impact on issues of inclusion, diversity, equity, and access.

The primary datasets used are predominantly derived from the Global North, as a direct result of many factors including connectivity and the price of bandwidth and hence are inadvertently lacking in representativeness and diversity and focus predominantly on those languages for which there is a lot of data available. As a result, the output data is skewed.

Wikipedia is not only “one of the world’s largest and most comprehensive knowledge and educational resources” but its importance has grown with the rise of GenAI as almost every LLM, including ChatGPT, “relies on Wikipedia data as

their primary source for training” (Alikan, 2024). The Wikimedia Foundation itself notes that, despite its global reach, it “lacks the participation and perspectives of about half the world’s population” as “only 13% of active editors on Wikimedia projects identify as women” (Alikan, 2024). Hence, some commentators have wryly, but not inaccurately, designated the outputs of GenAI as “mansplaining as a service”. Apart from gender, Wikipedia editors are also far more likely to emanate from the Global North.

With regard to imagery, “AI image generators like Stable Diffusion and DALL-E amplify bias in gender and race, despite efforts to detoxify the data fueling these results” (Washington Post, 2023).

In South Africa and many other low to middle income countries, “the digital divide means many people still lack reliable, inexpensive, high- speed internet connectivity”(Timcke & Rens, 2024). “While the global technology discourse is now dominated by artificial intelligence (AI), this focus often overshadows the critical importance of robust digital public infrastructure. Regardless of whether one deems AI’s potential to solve complex problems to be illusionary, without a digital public infrastructure foundation almost all of the benefits of AI will be out of reach” (Timcke & Rens, 2024).

As a result, the acquisition of skills necessary for the contemporary workplace is hindered, and the digital divide “threatens to worsen existing inequalities, creating new forms of social and economic stratification based on digital literacy and access”(Timcke & Rens, 2024).

Finally, allegations of copyright abuse are not going away, and in a dataset scrutinised by the Washington Post 200 million instances of the © symbol were found, indicative of the usage of copyrighted works (Schaul, K. inter alia, 2024). This can have a chilling effect on creators, whether textbook authors, photographers, artists and others, leading them to be less likely to share openly.

SLIDE EIGHT: EQUITABLE

Equitable

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

Castor and Gerard



[Extracting Training Data from Diffusion Models](#) by Carlini, Hayes, Nasr, Jagielski, Sehwag, Tramèr, Balle, Ippolito, and Wallace

This slide and notes taken from another presentation by Kure and Poritz graphically depicts that “GenAI Image Fabricators are not artists, but are arguably being used to “normalise appropriation of art” from artists (Jiang et al) since often what is produced is very like, if not identical to what has been created previously (Carlini inter alia) and if the original is copyrighted, this would render this digital forgery” (Kure, K.S. & Poritz, 2023).

“Copyright law is currently “unequipped to tackle many of the types of harms posed by these systems to content creators” (AIES, ’23, p.364) and of huge concern is the practice of “data laundering” wherein non-profits such as academic institutions gather images under the US exemptions for ‘fair use’, ostensibly for educational and research purposes, but which receive funding from for-profit entities which in turn train their commercial products on these data-sets, thereby deriving down-stream benefit for their donation and ensuring the data has the appearance of something that would fall under the “Fair Use” provision in US Copyright Law. This is not equitable, no matter how you try to parse it” (Kure & Poritz, 2023).

SLIDE 9: GenAI and the SDGs

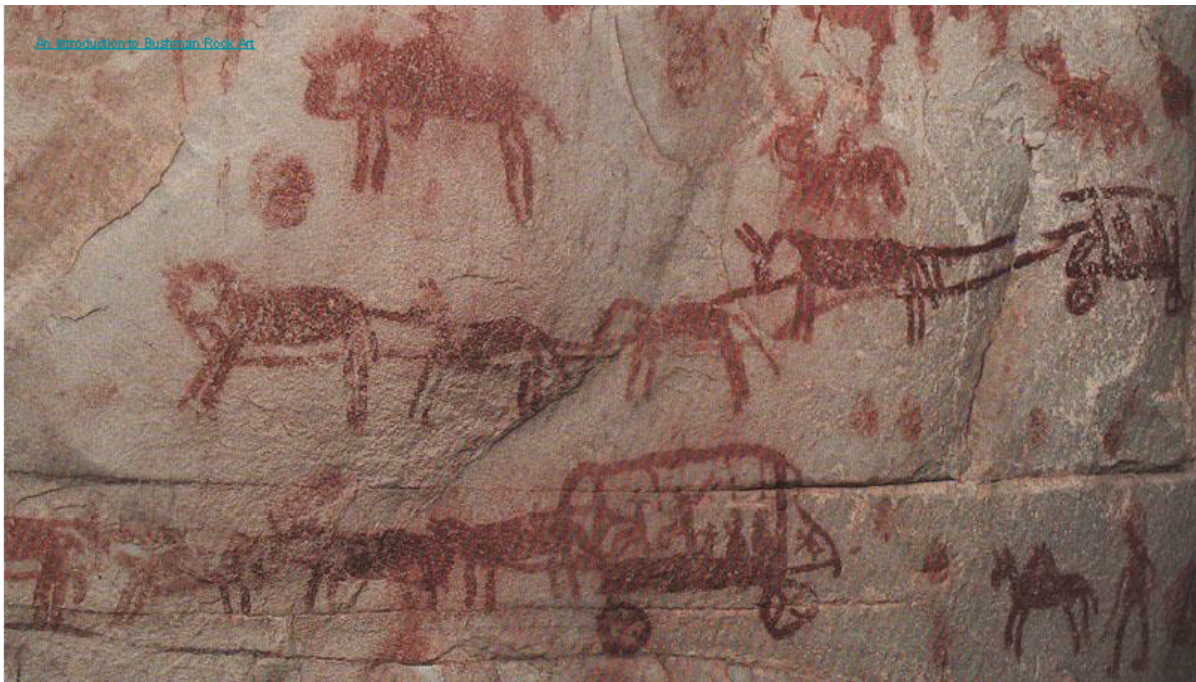


[Sustainable Development Goals](#)

Sustainability issues must also be addressed and GenAI must not only be evaluated in relation to SDG4: Quality Education, but also on 6: Clean Water as well as 7: Affordable and Clean Energy as GenAI consumes massive, and increasing amounts of fossil fuels together with copious amounts of water required to cool the vast warehouses full of data servers. One estimate is that a query in ChatGPT4 uses over 500ml of water (AP News) and enough power to fully charge a cell phone. Google alone uses up to 2% of the world's energy, more than 100 countries, and it is only one of the Big Five. We are concerned about SDGs 10 in terms of increasing inequality and 12 in terms of responsible consumption and production.

Furthermore, recent research indicates that while the power of the model has grown and grown with the size of the training datasets, these power curves are starting to level off (Marcus, 2024) and this has implications in terms of sustainability since the energy uses are increasing dramatically without the concomitant increase in performance.

SLIDE 10: WEIRD about not being WEIRD



It is always tempting to take what has already been produced in terms of OER, curate it and adapt it to local context, or at least language. It can be argued that LLMs simply scale that approach. But even if translations are perfected, the massive corpora of texts in dominant languages also capture particular worldviews and ways of conceptualising and understanding knowledge. Henrich (2020) argues those raised in Western, Educated, Industrialised, Rich and Democratic or WEIRD countries differ from others, particularly along the dimension of individual versus collective responsibility and WEIRD individuals are notably more analytical, individualistic, and impersonal - which worldview is then reflected in the types of texts created and what is deemed evidence. If we prioritise what is already created and attempt to transform all digitised content by means of a particular framework, how can this allow the voiceless to find their voice, especially if those voices find expression in myths and folklore and strong oral traditions which may not even be recorded, or if recorded, not yet transcribed into written evidence (Darold, 2018) and even if transcribed may not be sufficiently robust to be elevated to the level of evidence required by a rigorous and stringent textbook editor for instance?

A cursory analysis of some of the information in Volume 1 of World History, encompassing the time period up until 1500 and published by OpenStax, immediately throws up some concerns in their section on southern Africa. Just a single sentence will suffice to demonstrate how its understanding of the First Peoples of Southern Africa is both dated, and limited. To quote:

“Hunting and gathering peoples of Africa have historically included the ... San people of the Kalahari Desert.” (World History, Vol 1, p. 377)

The term San is derogatory at its root, as it is an exonym or non-native term “meaning “foragers” (Wikipedia) and used dismissively for those considered too poor to have cattle. It is increasingly rejected by those few survivors of genocide and enslavement who do not consider themselves as a category at all but conceptualise themselves more along the lines of what we think of as clan or tribes, speaking three distinct languages in very different geographical areas. Henrich contends that WEIRD people use analytic thinking which “places people or objects into distinct categories and assigns them properties to account for their behaviour” (Siliezar, J. 2020) in contradistinction to holistic thinkers who “focus on relationships, context, and interaction” (Siliezar, J. 2020). Although some do live contemporaneously in some parts of the Kalahari, during ancient historical times they roamed freely throughout southern Africa, leaving behind many thousands of rock paintings.

This painting, for instance, found in the Cederberg in the Western Cape, depicts Boer settlers with their ox-wagon caravans, proving the coexistence of both in that area at that time. The colonial appropriation of land, and formalisation of farming with fences meant the hunting groups became considered pests and were systematically literally hunted to death by legitimised hunting parties granted a licence to hunt, though they also were enslaved as farm labour. The Kalahari became a tenuous place of final refuge, since it was uninhabitable by others, until the discovery of diamonds at which stage there was further dislocation, marginalisation and dispossession.

This is an example where context is critical, and the most elementary, albeit appalling facts available to someone living in southern Africa appear not to be common knowledge in the United States.

In conclusion, those in Open Education cannot look to GenAI to solve the issues of how best to consider under-represented views and voices so as to enable inclusion, diversity, equity, and access in open education policies and strategies. In short, Large Language Models are reliant upon what is there, they cannot solve for what is not there.

SLIDE ELEVEN: REFERENCES

Thank you!

References in the notes

Kathryn S. Kure: kathryn@datamyna.com
Nomvuyo Mgoqi: mvuyo59@gmail.com
Jonathan A. Poritz: jonathan@poritz.net

[The Forgotten 'Bushman' Genocide - The Namibian](#) 8 June 2021

Arya, N. [GPT-4 Details Have Been Leaked! - KDnuggets](#) 19 July 2023

[An Introduction to Bushman Rock Art](#) Visionary Art Gallery.

Lebedeva, I. & Moklyak, O. [ChatGPT vs GPT-4 vs GPT-3: Comparison chart for businesses](#). Grenice.

George, A. [Visualizing the size of Large Language Models](#) | by Anil George | [Medium](#)

Alikhan, A. [Wikipedia Needs More Women – Wikimedia Foundation](#) 8 March 2024

[Artificial intelligence: Performance on knowledge tests vs. dataset size - Our World in Data](#)

Villalobos, P. & Ho, A (2022), "Trends in Training Dataset Sizes". Published online at epochai.org. Retrieved from: [Trends in Training Dataset Sizes | Epoch AI](#) [online resource]

Tiku, N, Schaul, K. & Yu Chen, S. (2023) [AI generated images are biased, showing the world through stereotypes - Washington Post](#)

[AI models collapse when trained on recursively generated data | Nature](#) 2024

Timcke, S. & Rens, A. [Upgrading digital public infrastructure is less glamorous but more important](#) Business Live Opinion Piece 14 October 2024

O'Brien M., & Fingerhut, H. [Artificial intelligence technology behind ChatGPT was built in Iowa — with a lot of water | AP News](#) 9 September 2023

Insight: [Reading: Xhosa Noun Classes](#) 16 October 2018

Peleg, Y. [GPT-4's details are leaked](#) Internet Archive.

Schaul, K. Yu Chen, S. & Tiku, N. (2024) [See the websites that make AI bots like ChatGPT sound so smart - Washington Post](#) The Washington Post.

[The Impact of Contact and Colonization on Indigenous Worldviews, Rock Art, and the History of Southern Africa : "The Disconnect" | Current Anthropology: Vol 63, No S25](#)

Zhou, L, Schellaert, W. Martínez-Plumed, F., Moros-Daval, Y. Ferri, C. & , Hernández-Orallo, J. [Larger and more instructable language models become less reliable | Nature](#) 634, pages 61–68 (2024),

[World History Volume 1](#) OpenStax.

Marcus, G. [CONFIRMED: LLMs have indeed reached a point of diminishing returns](#) Substack. 10 November 2024

Altun, M. [\(PDF\) What Global Education Should Focus on](#) International Journal of Social Sciences September 2017

Marcus, G. Substack [Evidence that LLMs are reaching a point of diminishing returns - and what that might mean](#) 9 April 2024.

Selizier Juan. How the West became WEIRD. [Joseph Henrich explores WEIRD societies — Harvard Gazette](#) 16 September 2020

Darold [Wikimedia to the Rescue? How Wikipedia's Crowdsourcing Model Could Catalyze the Field of Oral History](#). Columbia Blog, 22 December 2018.

McGuire, D., Kure, K.S., Mgoqi, Amobil, P. [Literacy in the Mother Tongue in Africa](#), Creative Commons Global Summit, 5 October 2023, pre-recorded video.

Kure, K.S. and Poritz, J. [Will Generative AI DDOS The Commons?](#) The Creative Commons Global Summit 2023, pre-recorded video. (24 September 2023). Presentation and notes here: [Will Generative AI DDOS The Commons?](#)

Carlini, N., Hayes, J. Nasr, M., Jagielski, M. Sehway, V., Tramèr, F., Balle, B., Ippolito, D. Wallace, E. [Extracting Training Data from Diffusion Models](#) the Proceedings of the 32nd USENIX Security Symposium. August 9–11, 2023 • Anaheim, CA, USA

Bryce, R. [How Google Powers Its 'Monopoly' With Enough Electricity For Entire Countries](#) Forbes. 21 October 2020.

[Hardware and energy cost to train notable AI systems](#) Our World in Data.

Pablo Villalobos, P., Ho, A., Sevilla, J. Besiroglu, T., Heim, L. & Hobbhahn, M. [Will We Run Out of Data to Train Large Language Models? | Epoch AI](#) 6 June 2024.