

Position: the Stochastic Parrot in the Coal Mine

Model Collapse is a Threat to Low-Resource Communities

Devon Jarvis^{*12} Richard Klein¹² Benjamin Rosman²¹ Steven James¹² Stefano Sarao Mannelli³¹

Abstract

Model collapse, the degradation in performance that arises when generative models are trained on the outputs of prior models, is an increasing concern as artificially generated content proliferates. Related critiques of large language models have highlighted their tendency to reproduce frequent patterns in training data, their reliance on vast datasets, and their substantial environmental cost. Together, these factors contribute to data degradation, the reinforcement of cultural biases, and inefficient resource use. In this position paper we aim to combine these views and argue that model collapse threatens current efforts to democratize AI. By reducing training efficiency and skewing data distributions away from the tails of their support, model collapse disproportionately impacts low-resource and marginalized communities. We examine both the environmental and cultural implications of this phenomenon, situate our position within recent position papers on model collapse, and conclude with a call to action. Finally, we outline initial directions for mitigating these effects.

1. Introduction

The emergence of generative artificial intelligence (AI) techniques, such as transformers (Vaswani et al., 2017) and diffusion models (Rombach et al., 2022), has led to widespread public adoption in recent years. ChatGPT alone is estimated to generate approximately 0.1% of all words produced globally each day (Schaeffer et al., 2025). This adoption is already reshaping how people work and learn (Chen et al.,

2020; Brynjolfsson et al., 2025). While generative AI has the potential to be broadly democratising, the consequences of its large-scale use remain poorly understood (Kosmyna et al., 2025; Stankovic et al., 2025).

The low cost of content generation has led to an unprecedented growth in available data (Wei & Tyson, 2024), with recent estimates suggesting that generated content constitutes roughly 30% of the internet as of 2025 (Spennemann, 2025). As a result, successive generations of generative models are trained on increasing amounts of AI-generated data. Several recent studies show that training repeatedly on such “synthetic” data degrades generation quality across iterations (Zhu et al., 2025; Shumailov et al., 2024; Dohmatob et al., 2024a; Seddik et al., 2024). This phenomenon, known as model collapse, leads to highly repetitive or incoherent text and overly stylized, unrecognisable images (Bohacek & Farid, 2023; Dohmatob et al., 2024c; Gerstgrasser et al., 2024).

Parallel work has also raised concerns around the propensity of language models to reinforce stereotypes and biases present in their training data, and the significant computational resources required to train them (Bender et al., 2021; Samsi et al., 2023; Jegham et al., 2025). Due to the expense in collecting data from communities less prevalent on the internet (Kandpal & Raffel, 2025) and training such extremely large models, low-resource communities have less representation in emerging technologies which hinders the applicability of the technology to their contexts.

At the same time, the energy demands of large-scale computation continue to drive new infrastructure development in already energy-intensive regions (Electricity, 2024), while the environmental consequences are borne globally. Low-resource communities are particularly vulnerable to these effects, including increased exposure to climate-related disasters (Van Aalst, 2006). As a result, these communities bear a disproportionate share of the costs while receiving comparatively little of the benefit (Farnadi et al., 2024). As Strubell et al. (2019) note, marginalized communities that experience the effects of environmental change first are often the last to gain access to new technologies.

¹School of Computer Science and Applied Mathematics, University of the Witwatersrand, Johannesburg, South Africa
²Machine Intelligence and Neural Discovery Institute, University of the Witwatersrand, Johannesburg, South Africa
³Data Science and AI, Computer Science and Engineering, Chalmers University of Technology and University of Gothenburg. Correspondence to: Devon Jarvis <devon.jarvis@wits.ac.za>.

In the context of language models, these structural imbalances are compounded by the phenomenon described by Bender et al. (2021) as “Stochastic Parrots”. While large language models can produce fluent and coherent text when trained on vast datasets, their outputs remain ungrounded, lacking social context and adherence to human norms. The consequence is that the models repeat statements based on frequency rather than merit or accuracy (Orgad et al., 2025). Thus, generated text can be skewed by social constructs that are potentially harmful and factually incorrect. Even in the case of hallucinations, the errors are not arbitrary and reflect an underlying bias of the models against under-represented groups (Farnadi et al., 2024; Orgad et al., 2025). Coupled with the human proficiency at finding meaning as an interlocutor (Clark & Bangerter, 2004; Clark & Krych, 2004; Janzen & Shaffer, 2008), this creates systems that can be misleading or harmful at scale (Choi et al., 2024). Although we revisit the continued relevance of the Stochastic Parrot analogy after several years of progress in Sec.4, it remains foundational to the concerns raised in this perspective.

While model collapse, the energy consumption of Large Language Models (LLMs) and their propensity to repeat hegemonic or biased viewpoints are pressing issues in isolation, their interaction has received little attention. We begin to close this gap by arguing that **model collapse amplifies the existing harms of LLMs for low-resource languages and marginalized communities by producing models which favour hegemonic viewpoints more with each generation, remove under-represented viewpoints from the data distribution and scale poorly with the addition of real data. By poisoning the data and removing the tails of the distribution, it undermines ongoing efforts to democratize AI. Addressing model collapse should therefore be a central concern for research on low-resource settings and AI fairness.**

To substantiate this claim, we begin in Sec.2 by reviewing the recent theoretical and empirical findings on model collapse. In Sec.3 we reconsider the points raised in the seminal work by Bender et al. (2021) on Stochastic Parrots in light of the new findings on model collapse (Dohmatob et al., 2024c; Shumailov et al., 2024). We highlight how model collapse exacerbates both the environmental and social risks associated with large language models, while also introducing new risks. Sec.4 provides some alternative views which might mitigate the potential risks of model collapse on low resource settings. Sec.5 concludes with a call to action and initial recommendations for immediate directions of research to better understand or mitigate the risks of model collapse for low-resource and marginalized communities.

Throughout this work we make use of the term “low-resource community” and it is important to specify what we mean. In this case, the term is not limited to reflecting the

fact that the community has a relatively smaller proportion of their data present in datasets or on the internet, although this is part of it. Through this term we also include communities which lack sufficient compute or financial means to train large models or collect data freely, relative to the dominant communities on the internet. Furthermore, this notion of a dominant community is highly context specific and temporal. A community may be dominant in a particular dataset and not within another. Importantly, there are a number of reasons why a community’s data may occupy the tails of a data distribution and we do not consider any one reason here. We do, however point to a few beyond a lack of representation throughout this work. Our point here is not to group all of these concerns into one, but to argue that these reasons and difficulties exist disproportionately for some communities which may lead to more severe model collapse for these same communities. Thus, the term “low-resource community” is used for ease of communication and not intended to convey a uniform group. Understanding how model collapse manifests for different reasons is something we touch on in our call-to-action in Sec.5. Additionally, Tab.1 summarizes the main mechanisms of model collapse discussed throughout the paper, linking each cause to its expected effect on low-resource communities.

Conflict of Interest Disclosure: The authors have no conflict of interest to disclose.

2. Background

While the precise definition of model collapse is still evolving and often conflates distinct effects (Schaeffer et al., 2025), the literature generally groups several phenomena under this heading, ranging from a catastrophic divergence in test loss – often an artefact of unrealistic “replace” training paradigms (Schaeffer et al., 2025) – to subtler deformations of the data distribution. Despite this definitional ambiguity, the core concern generally refers to a specific failure mode: *a progressive decrease in the marginal value of additional training data (when a proportion of that data is synthetic) and a simultaneous reduction in the diversity and quality of the output distribution* (Shumailov et al., 2024).

This degradation is best understood through the lens of neural scaling laws that describe the predictable power-law improvement in model performance (usually measured by cross-entropy loss) as compute, dataset size, and parameter count increase (Kaplan et al., 2020; Bahri et al., 2024). These laws provide an optimistic alternative to the performance plateaus typical of overfitting in earlier architectures (Cogswell et al., 2015; Santos & Papa, 2022). However, the introduction of synthetic data fundamentally alters this trajectory. Theoretical and empirical works indicate that at sufficient scales, even a small fraction of synthetic data (around 1%) causes models to suffer from a change in scal-

Cause / mechanism	Why it matters for low-resource communities	Likely effect	Concerns
Synthetic-data feedback loops (Whitney & Norman, 2024)	Small real-data distributions are more easily overwhelmed by generated content.	Dominant patterns are reinforced across generations, while rare linguistic and cultural forms are pushed further into the tail.	Representation loss
Tail erasure under model collapse (Wyllie et al., 2024)	Low-resource languages, dialects, and marginalized viewpoints often already occupy low-probability regions of the data distribution.	Loss of diversity, reduced expressivity, and weaker model performance on communities whose data is already scarce.	Cultural erosion
Changed scaling behaviour with synthetic data (Shumailov et al., 2024)	If more data yields less improvement, communities with limited access to real data and compute have fewer viable routes to improvement.	Higher data and compute requirements, widening the gap between high- and low-resource settings.	Unequal access
High cost of real data acquisition (Kandpal & Raffel, 2025)	Collecting, validating, and maintaining culturally representative data is expensive and often unpaid or under-resourced.	Low-resource communities may be pushed toward synthetic supplementation, increasing the risk of poisoning their own data.	Data inequity
Automated curation and synthetic-data detection (Boutadjine et al., 2025)	Synthetic detection and filtering systems may be unreliable, opaque, or biased against uncommon language use.	Authentic but rare or reclaimed language may be removed, while synthetic or hegemonic content remains.	Curation bias
Value-lock and hegemonic viewpoints (Bender et al., 2021)	LLMs tend to reproduce frequent historical patterns rather than socially evolving or locally grounded meanings.	Outdated or dominant viewpoints persist, while new terms, norms, and movements struggle to enter the model distribution.	Bias reintroduction
Multilingual transfer and language drift (Lu et al., 2020)	Low-resource languages often depend on shared multilingual representations learned from high-resource languages.	Fine-tuning or repeated generation may erode low-resource capabilities faster than high-resource ones.	Language drift
Tokenizer and morphology mismatch (Rajab, 2022)	Agglutinative languages, code-switching, and non-standard orthographies may create longer-tailed token distributions.	Model collapse may manifest differently across languages, making universal mitigation strategies unreliable.	Measurement gap

Table 1. Mechanisms by which model collapse may disproportionately affect low-resource and marginalized communities.

ing exponents – essentially hitting a “glass ceiling” where larger datasets no longer yield proportional improvements (Dohmatob et al., 2024b;a; Cui et al., 2025). This phenomenon is explained as the model overfitting to the dominant factors of variation in the original dataset which are then over-represented in the generated samples (Shumailov et al., 2024). The generated samples enter the training data for subsequent models, strengthening these dominant factors further and creating a feedback loop where the model progressively loses the ability to represent the variability and structure of real-world data (Shumailov et al., 2023; Bohacek & Farid, 2023).

This over-representation of dominant factors leads directly to a loss of variability through this feedback loop that systematically erases rare or “tail” examples (Shumailov et al., 2023; Bohacek & Farid, 2023). Consequently, the rich variance of the original real-world distribution is replaced by a smoothed, homogenized approximation. The threat of model collapse is, therefore, not necessarily that models will become gibberish, but that they will become narrow—capable of reproducing frequent patterns but increasingly blind to the long tail of human diversity (Schaeffer et al., 2025; Shumailov et al., 2024).

This process is closely related to a similar concept in cognitive science known as iterated learning (IL) which has been used to explain the refinement of language and behaviours over generations of human learners (Smith et al., 2003). In humans, the result is a highly structured (termed regular) language or set of actions that are easy to learn and can be composed to generalize systematically (Kirby et al., 2008). However, in both empirical experiments and theoretical models, expressivity can be lost without external intervention and the differences between closely related data points is forgotten (Kirby et al., 2008; 2014; Jarvis et al., 2026). Thus, to consider the inductive bias of the model to focus on dominant factors of variation as “only negative” is not grounded in the broader literature on cognitive modelling and machine learning. It is plausible that the real-world situation is more optimistic than those shown on smaller (but still naturalistic) datasets and in theoretical models. This is a point we return to in Sec.4. However, when coupled with the existing biases of LLMs, such as promoting the hegemonic viewpoints in its training dataset (Bender et al., 2021), and the already enormous costs involved in training them (Xia et al., 2024; Kandpal & Raffel, 2025), it is important to consider the potential harms which model collapse may cause for low-resource and marginalized communities.

3. Generations of Stochastic Parrots

3.1. Environmental and Financial Costs

The financial costs associated with training LLMs are often scrutinized (Strubell et al., 2019); for example, training a single BERT base model without hyperparameter tuning on GPUs was estimated to consume as much energy as a flight across America (Strubell et al., 2019), training a large transformer model with neural architecture search emits as much CO_2 emissions as the average human over 60 years (Vaswani et al., 2017; Bender et al., 2021), and training GPT-4 reportedly cost over \$100m (O’Donnell & Crownhart, 2025). However, this estimate does not consider the cost of the data used to train the models (often because it goes unpaid). Based on conservative estimated of wage rates, the costs of obtaining the training datasets are 10–1000 times larger than the costs to train the models themselves (Kandpal & Raffel, 2025). Even if the real cost is highly discounted, there is still an opportunity cost and loss of human capital associated with the use of such expensive data.

In many cases, the high cost of training an LLM is justified by its use as a foundation model (Xu et al., 2024). Hence, the cost is amortized over its entire life-cycle and uses (Ren et al., 2024). However, most research establishing the costs of training LLMs is based (necessarily) in idealized settings and with clean datasets (Zeighami et al., 2025). For example, Strubell et al. (2019) estimate that an increase in BLEU score of 0.1 from using neural architecture search for English to German translation results in an increase of \$150,000 in compute costs. Further, this cost profile shifts drastically with new modalities: generating a five-second video can consume more than 700 times the energy of generating an image (O’Donnell & Crownhart, 2025). Consequently, the compute cost of training the largest deep learning models increased roughly 300,000-fold between 2015 and 2021 (Devlin et al., 2019), a trend expected to accelerate with large-scale infrastructure projects such as the \$500 billion “Stargate” initiative (O’Donnell & Crownhart, 2025).

Using the linear model of model collapse proposed by Dohmatob et al. (2024b), achieving a tenfold reduction in test loss (from 10^{-1} to 10^{-2}) requires an order-of-magnitude increase in dataset size (from 10^3 to 10^4), consistent with established power-law scaling (Kaplan et al., 2020). Considering a dataset where 10% is synthetic, the increase in dataset size needed is *another* order of magnitude (from 10^3 to 10^5). While it is important to recognize that this is a theoretical model (including a hyperparameter to control the quality of the synthetic data), even if the model is pessimistic, model collapse can only stand to increase the amount of data and compute needed to train new models. Coupled with the point that the hidden costs of producing the training data are often overlooked (Kandpal & Raffel, 2025), this increase in cost becomes even more severe.

While the natural answer of collecting more real data may be viable in some settings, if the cost of obtaining data is even on the lower end of the estimate by Kandpal & Raffel (2025), then this will likely not be a viable solution for many populations. This leaves low-resource communities unable to obtain data without significant cost or without the use of synthetic data as a mechanism to supplement datasets. Thus, when benchmarking the environmental cost of training foundation models it is important to also consider the effect of cheap synthetic data or expense of real data on this process, or else it stands to be highly optimistic. This inequity is compounded by the fact that data centres are often clustered in regions with carbon-intensive grids, resulting in emissions 48% higher than the national average (O’Donnell & Crownhart, 2025).

While it is unrealistic to expect uniform technological advancement across different communities and sub-fields, the typical belief is that progress in one space will lend insight and lead to advancement in another. Instead, model collapse may result in fundamentally uneven access to new technology between communities. Moreover, the communities most hindered are precisely those that could benefit the most from technological advancement. Even under more optimistic scenarios, model collapse threatens to undermine ongoing efforts to reduce the environmental impact of training LLMs (Huang et al.; Iftikhar et al., 2025; Shi et al., 2025) and the injustice this causes (Strubell et al., 2019). Thus, beyond unequal access to new technology and the disproportionate impact of environmental harm on certain sub-populations, we hypothesize that model collapse and the cost of data acquisition could introduce a third disadvantage to low-resource communities: *a lower upper-bound on technological advancement* from these models.

3.2. Cultural Costs

The source of injustice from LLMs is not only in the resources required to train and use the models, however. There are also valid concerns stemming from the effect of LLMs on public perception and culture. Foundational to this section is the recognition that model collapse results in a lack of diversity of generated content and that what remains in the data is the highly probable content from the original dataset (Alemohammad et al., 2023; Bertrand et al., 2024). Thus, data which would occupy the tails of a probability distribution are forgotten or corrupted (Shumailov et al., 2024). Datasets used to train generative models are typically scraped from the internet, which has been shown to be problematic (Basta et al., 2019; Birhane & Prabhu, 2021; De Vries et al., 2019; Kurita et al., 2019). This has resulted in the models encoding biases and derogatory associations in some cases (Bao et al., 2024; Shin et al., 2024). Even when there is not overt harm, the models favour the hegemonic viewpoint presented on the internet sites where the data is

scraped from (Bender et al., 2021) – one of the foundational points leading to the stochastic parrot phrasing.

The comprehension of LLMs regarding minority groups has often been shown to be limited, absent, or stereotyped. Farnadi et al. (2024) comprehensively analyse the types of discrimination LLMs perform, identifying not only privacy issues — as LLMs tend to memorize under-represented data more — but also “hallucination gaps” where models hallucinate more facts regarding minority groups. Furthermore, they note “underspecification disparities” where models generate seemingly arbitrary text when addressing topics related to marginalized groups (Farnadi et al., 2024). Consequently, LLMs flatten complex concepts (Gallegos et al., 2024), often aligning towards Western and English-speaking values (Rao et al., 2023). Together, these factors produce an outcome that amplifies bias and removes diversity. This effect is further amplified by standard compression and acceleration techniques — including knowledge distillation, quantization, pruning, and caching — which have been shown to further increase bias (Silva et al., 2021; Ahn et al., 2022; Kirsten et al., 2025). Cases of biased data generation are also not limited to LLMs. For example, it has been shown that stable diffusion has a propensity to perpetuate stereotypes regarding race and skin tone which has a homogenising effect on less frequently represented racial and ethnic groups (Wilson et al., 2025).

Data curation has been proposed as a potential solution to this problem (Birhane & Prabhu, 2021; Jo & Gebru, 2020), but there are a number of ways that model collapse can undermine the curation process. If mitigating model collapse becomes a part of data curation, then it will be necessary to detect synthetic data for its removal. However, this is becoming an extremely difficult task as generative models become more convincing to human perception (Boutadjine et al., 2025). Thus, curation will rely more on automated detection systems. Although there has been some success in automating data mixing and curation through methods like instruction-tuning (Jiang et al., 2024; Raj et al., 2024; Liu et al., 2025; Thudi et al.) these approaches are still potentially fallible, uninterpretable and require large quantities of real or synthetic data to train (Chaka, 2023; Elkhataat et al., 2023; Ghiurău & Popescu, 2024). This could stand to worsen the of curation to remove data from marginalized groups or those who express fair but less frequent views. For example, the Colossal Clean Crawled Corpus (Raffel et al., 2020) was cleaned by discarding any sample with one of 400 bad words. While this is effective at removing some of the worst known biases and slurs from data, it also reverses any attempts to reclaim works by marginalised groups (Bender et al., 2021). This effect of LLMs to re-establish outdated connotations for words is termed “value-lock” (Bender et al., 2021) and disturbs the role of language in shaping social norms and culture. As Twyman et al. (2017) note, a central

aspect of recent social movements revolves around using online forums to document events, promote coverage of new events and marginalized perspectives, and change the perspectives on existing knowledge in light of new information.

Value-lock was already a concern for first-generation LLMs based only on generated content being more likely to represent previously dominant (and hence frequently represented in the training data) viewpoints. The likely outcome of model collapse is that these dominant viewpoints will become even more over-represented in the dataset as subsequent generations of models are trained (Whitney & Norman, 2024). Equally, the less frequently encountered language which is deliberately used to change social perception will be forgotten over the generations. Wyllie et al. (2024) contextualize this within the rapid increase of LLM-generated text, defining the “Fairness Feedback Loop.” They argue that model collapse is not merely a quality issue; it is an erasure engine that systematically removes under-represented communities from the digital reality defined by AI models. All of this stands to reinforce existing regimes of power and undermine attempts to promote social change.

Importantly, the looming harm from model collapse is not limited to a shift of the data distributions towards previous views, as shown by recent cases in high-resource settings. For example, Wang et al. (2025) show that GPT-2 favours right-leaning media when successively fine-tuned on a balanced media dataset from the US such that it undergoes model collapse. *Thus, even in ideal settings model collapse has the propensity to inject bias where there seemingly is none, or based on underlying biases which are not immediately obvious to the expert and public raters of their dataset.*

3.3. Transfer Learning

Recent work on low-resource languages has increasingly relied on transfer learning from multilingual models (Lai et al., 2023; Devlin et al., 2019). The rationale is that combining data from multiple languages—particularly high-resource ones—provides sufficient data to learn a semantically rich embedding space. The low-resource language can then leverage this space, enabling the model to represent meaning more effectively despite limited direct data. However, fine-tuning in this manner introduces several issues, including language drift, where the embedding space is progressively forgotten as the model is trained on a more narrowly defined task (Lee et al., 2019; Lu et al., 2020). In other words, while transfer learning provides benefits, the model can lose capabilities originally acquired during pretraining on a rich embedding space.

How model collapse manifests in this setting where only data at the edge of the distribution (the low-resource language) is being generated is an open question. If the model remains multi-lingual, then model collapse will likely result

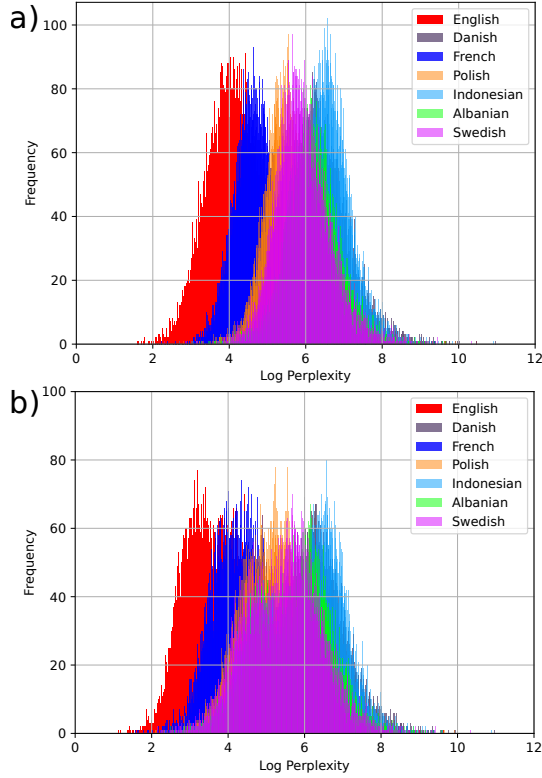


Figure 1. Perplexity of multiple languages using the Latin alphabet (potentially with some added characters) calculated using a pretrained GPT-2 (Radford et al., 2019). We present this result to support our position only.

a) **Original Distribution:** Note how the lower-resource languages occupy a distribution closer to the tails (at a higher perplexity) than the more high-resource languages such as English. Each language distribution is calculated using 15000 input sentences (agentlans, 2025) and show log-perplexity for visibility.

b) **Collapsing Distribution:** Note how the low-resource languages shift the most when 5000 samples are generated for each language and added to the distribution with replacement. Swedish and Polish are affected worse than the other languages, while all distributions shift left and begin to overlap – highlighting the homogenising effect of collapse.

in the erosion of low-resource languages from the data over time, and the model will progressively deteriorate faster on these languages. Figure 1 demonstrates this by showing the distribution of log-perplexity values GPT-2 (Radford et al., 2019) attributes to 16000 sentences from various languages using the Latin alphabet (agentlans, 2025). Clearly, high-resource languages such as English and French have a lower initial perplexity compared to Indonesian and Albanian. By our current understanding of model collapse, these languages at the tails of the combined distribution should be forgotten first, which is supported by the initial shift in distributions when just 2000 generated samples are added. While this is a preliminary, it is still a stark result demonstrating more need for work focusing on the effect of model collapse on low-resource languages.

Multi-lingual models are also often told which language to

generate text for (or at least prompted in a particular language) and so it is unlikely to collapse in the same manner as if it was generating text from its intrinsic distribution of language. This deliberate mechanism introduced by the human user is a means by which model collapse could be mitigated. However, prompting and guiding the generation in this manner reintroduces a mechanism for bias (Yang et al., 2023; Hida et al., 2025; Chisca et al., 2024). These and many more questions, such as the effect of model collapse on languages where code-switching is common (Woolford, 1983) (and itself a lingering issue for many low-resource languages (Yoo et al., 2025a;b; Mohamed et al., 2025)) remain open and are important to fully grasp the effect model collapse could have on marginalized and low-resource communities.

This leads to our final consideration: the deliberate use of generative models to produce content for low-resource settings and marginalized communities. One reason this is needed is to ensure that this data remains represented in the data distributions of multi-lingual models to some degree. In essence, having synthetic data for your community enter into the broader distribution is likely better than leaving your community’s data to slide further into the tails. Implicitly, there is a race between communities to keep pace with the rate of data generation of the others. Real data is certainly the best content to produce, but this favours communities with resources, as noted in Sec.3.1. The net result is that marginalized communities may be pushed to poison their data merely to ensure their data remains salient in combined datasets. The other reason marginalized communities may still need to use generative models, and likely the larger reason, is to keep pace with the growing productivity internationally which can result from new technologies. Thus, these communities will need to adopt generative models to remain economically active within a globalized society with blurred cultural boundaries.

The big question is then: how quickly will content be generated in these settings with AI, compared to human content creation? In the best case, content generation will follow the same distribution of general content on the internet. In this case, all communities will have roughly the same percentage of synthetic content in their training data at a given time and will face a similar degradation from model collapse. Since the plateau from model collapse gets worse as more data is added (Shumailov et al., 2024), this may even make model collapse initially worse for high-resource settings. However, it seems plausible that the technology will provide a proportionally larger effect on low-resource settings as the cost of generating content becomes significantly cheaper. Essentially, synthetic content generation is democratised far quicker than the internet as a whole. In this case the rate of generated content production will more closely follow the population of a country, while the rate of real content generation will align to resource access. Thus, in absolute terms,

the amount of content generated by low-resource communities will be less but the proportion of synthetic data in these languages will be greater. *This will result in low-resource communities encountering severe model collapse far quicker than high-resource communities and the erosion of these cultures, before we have the methods to reverse the collapse.*

4. Alternative Views

There are a number of alternative views to be considered, ranging from the problem being real but mitigated, to model collapse (or at least the bottleneck imposed by generations of trainers) being beneficial. Considering the former, the extent of model collapse is still being debated and some work does show that models can stabilize after a certain number of generations (Kazdan et al., 2025). Similarly, cases where extreme collapse are shown tend to be on rather small datasets and measured using metrics like the perplexity of individual samples (Shumailov et al., 2023). While these metrics and experiments match qualitative evaluations of the generated content, how to accurately measure or characterize model collapse remains a valuable question. Due to this difficulty, isolating its effects is also difficult. A subtle but important alternative view may consider whether model collapse is really accelerating or introducing bias in the data distributions or just a symptom of a growing gap between communities based on access to resources. Thus, it is plausible that with sufficient real data, models will not be overwhelmed by synthetic data and that the data distributions may stabilize to content which is no more biased than it otherwise would be in the first generation (Alemohammad et al., 2023; Kazdan et al., 2025).

The comparison of model collapse to iterated learning supports the latter view that model collapse may even be beneficial. As noted in Sec.2, the generational transmission of language and the regularity this affords is a benefit for human language use and cognition (Kirby et al., 2008). One could argue that model collapse will result in the loss of invalid information (which includes biases) due to the fact that it is less grounded within the environment.¹ As a result, the language which will be most represented will be grounded language which expresses truths about nature. In a sense, there are many ways to be wrong and so no one way should dominate the data distribution.

It is important to note that the iterate learning algorithm can also result in a loss of expressivity when run with humans and some artificial agents (Kirby et al., 2008; 2014; Jarvis et al., 2026). What is required is a mechanism for the reintroduction of words and forgotten information and some creativity from the speakers. It is conceivable that with such

¹We certainly run the risk of oversimplification here, as the degree to which biases are ingrained into the environment once they exist is a subtle point beyond the scope of this discussion.

a mechanism, LLMs will contribute to the natural evolution of language without causing collapse. This process could even be partially achieved in the near-term through prompting and curriculum learning (Jiang et al., 2024), the inclusion of richer multi-modal context windows, and frameworks such as algorithmic reparation (Davis et al., 2021; Wyllie et al., 2024). Once again, this will require more data and deliberate use by different communities to fight model collapse. Moreover, the worse the collapse, the more difficult it will be to reintroduce sufficient language to stabilise the data distribution. If any communities have a smaller margin for error and should be seeking such interventions, it is the low-resource and marginalized communities.

Schaeffer et al. (2025) presents one important prior position paper which aimed to provide some needed precision in the discussion of model collapse. The primary point of Schaeffer et al. (2025) is that the literature on model collapse both conflates multiple definitions and over-simplifies the phenomenon. One of the primary critiques is that within a given paper, the definition of model collapse changes and distinct phenomena are used to characterize its occurrence. We fall into the same trap here: Sec.3.1 relies on a notion of model collapse that causes a change in neural scaling laws (Dohmatob et al., 2024c, and Definition 5 in Schaeffer et al. (2025)), while Sec.3.2 focuses mainly on the loss of data at the tails of the original distribution (Shumailov et al., 2023; Wyllie et al., 2024; Shumailov et al., 2024, and Definition 7 in Schaeffer et al. (2025)). It is not our aim here to argue for a particular definition, but rather to show that regardless of the definition used it is plausible that model collapse will disproportionately affect low-resource communities.

From the three “types” of model collapse definitions identified by Schaeffer et al. (2025), we have mainly considered the consequences of two types. Specifically, we show that type 2, which considers the deformation of the data distribution, will likely first erase marginalized community data and promote the hegemonic viewpoints more so than even prior works considered (Bender et al., 2021). Secondly, we considered type 3, which is the decreasing value from additional data, and reconsidered the already alarming trend of marginalized communities disproportionately suffering the costs of environmental damage while benefiting the least from the new technologies (Strubell et al., 2019). We extend this with the recent point from Kandpal & Raffel (2025) claiming that data acquisition is the most expensive part of a training pipeline, and highlight that strategies revolving around collecting more data to mitigate model collapse will only be feasible for communities able to “pay” this data cost.

The type 1 definition of Schaeffer et al. (2025) is the progressive inability of the model to fit the original data distribution due to the introduction of synthetic data over generations. This type of definition is certainly important, particularly

when lending theoretical understanding. However, this effect will also hit low-resource communities first as these data distributions will be more easily overwhelmed by synthetic data. Secondly, this type of definition provides a slightly more long-term perspective, while we are arguing for the immediate consideration of model collapse by low-resource communities. It is also worth noting that the intermediate generations are where the cultural harms noted in Sec.3.2 will arise, long before the generated content becomes incomprehensible and while it is still capable enough to influence public perception. Long before the models are unable to fit the original data distribution, we would argue that immense harm would have already been caused to low-resource communities through the disproportionate erasure of their data and the environmental damage caused by training sufficiently many models to reach this point.

Thus, while Schaeffer et al. (2025) argue for a more discerning and muted rhetoric around model collapse, we are actually in agreement with their position, as they note that real tail data will be lost and that scaling laws may change with the introduction of synthetic data. They even note that “loss of diversity is a real issue, with disproportionate harms often-times born by subgroups”—pre-empting our position in this work. We may even contextualize our position as a push towards a new axis of concern around model collapse: that model collapse will happen for different communities’ data at different rates and have uneven impact. For example, Schaeffer et al. (2025) note that humanity may be able to train trillions of models before we notice the onset of model collapse. This conclusion was drawn based on a premise of there being a wealth of data available. This is certainly not the case for many communities and languages. Consequently, while the discourse around model collapse may be over-cautious and pessimistic, it is certainly accurate for some low-resource communities, regardless of which definition of model collapse is ultimately favoured.

A final alternative view may consider whether the Stochastic Parrot analogy is still true, five years after the original work (Bender et al., 2021). While it is certainly true that immense progress has been made in recent years on grounded language understanding and generation (Radford et al., 2021; Rombach et al., 2022; Alayrac et al., 2022), particularly through the addition of multiple modalities, the statistical basis for the analogy remains true. LLMs base their predictions for the next token on statistical patterns which they have previously observed, which makes them susceptible to frequency and sample biases just like any other cognitive agent (Geva et al., 2025). The question then becomes: to what extent does the additional modalities allow multimodal models to ground their language generations in reality rather than the statistics of the language they encounter? This is beyond the scope of this discussion, however it appears to be of paramount importance for future

work. For now, we retain the Stochastic Parrot analogy as sufficiently relevant to provide the conceptual foundation for this work, particularly for low-resource languages and multilingual models, where clear sample biases and dominant patterns in the language data may not reflect broader societal realities (Bender et al., 2021).

5. Conclusions and Call to Action

In this work we have aimed to contextualize the current discussions of model collapse (Shumailov et al., 2024; Dohmatob et al., 2024a;b; Bohacek & Farid, 2023) around prior discussions on the biases and injustices perpetuated by generative models and their training (Basta et al., 2019; Bender et al., 2021). While many of the points raised here are not novel in isolation, it is the combination of perspectives we aim to contribute, as summarised in Tab.1. While some works argue that model collapse may be less severe than the current literature would lead one to believe (and importantly caution against excessive rhetoric) (Schaeffer et al., 2025), we would argue that model collapse has a higher likelihood of reaching some communities which are already marginalized. *Our position is that the degree of concern and concern should be based on the distribution, properties and quantity of a community’s data.* Thus, whether one believes that model collapse is an imminent threat to our online cultures and progress in machine learning, it appears that low-resource and marginalized communities will be affected by it first and most severely. Consequently these communities should have a more prominent role in the academic discourse on the topic.

Yet we find it troubling that this has yet to occur, with most experiments remaining on theoretical model (Shumailov et al., 2024), highly controlled datasets (Wyllie et al., 2024) or high-resource settings (Wei & Tyson, 2024; Wang et al., 2025). While these highly impactful works have set the foundation for subsequent studies on model collapse, it cannot be assumed that preserving cultural sensitivity and enhancing representation of smaller communities aligns with the goals of the broader community. It is the responsibility of low-resource and marginalised communities to be proactive in maintaining the representation of their data and cultures as far as possible. Yet, we also note that the majority of people will form part of a minority group in some aspect. While axes such as race, gender and ethnicity rightly receive significant focus there are a number of different factors influencing a person’s identity within their community. The homogenising effect of model collapse will affect all of these axes. Consequently, even if it is left to members within under-represented groups to understand and mitigate model collapse, model collapse still stands to be of wide-reaching concern for a large number of people. Thus, we conclude with a call to action: that the low-resource NLP and ML

communities should take the forefront in understanding and mitigating model collapse. To this end, we provide some suggestions for open questions and first steps.

The open questions we highlight revolve primarily around the differences many low-resource languages present compared to some dominant languages such as English. Characterising the consequences of these differences on model collapse appears to be of utmost importance. For example, model collapse may manifest very differently on languages which are agglutinative (such as Turkish, Korean and Swahili) (Durrant, 2013). Agglutinative languages tend to form new words by the structured composition of other morphemes. In contrast to English, which is an analytic language (with some fusional elements) (Li, 2018) and has a few highly reusable words within the vocabulary, agglutinative languages have a large vocabulary and a very long tailed distribution of word use. While LLMs typically use sophisticated tokenizers which are able to extract the morphemes from agglutination (Kudo & Richardson, 2018), this still highlights the potential for differences between languages. Even the choice of tokenizers, which have been shown to materially impact model performance on low-resource languages (Rajab, 2022), could play a key role in mitigating model collapse for these languages.

Another example of a unique property of low-resource languages is their propensity for code-switching (Olaleye et al., 2025), which occurs when words from the vocabulary of one language are used inter-changeable in another. Code-switching, in general, currently poses a technological problem for low-resource languages (Winata et al., 2023; Amol et al., 2025) and invasive words will shift the data distribution. If invasive words push other words to the tails, then parts of the original language will be forgotten. Otherwise, the invasive words will occupy the tails and be removed. In both cases the redundancy acts to increase the tail of the distribution and will impact how model collapse occurs.

These examples highlight an important point: there are many reasons why a community’s data may occupy the tails of the data distribution and be at risk of model collapse. While quantity of data is a clear contributor and saliently connected to marginalisation of communities, historical discrepancies in how technology is developed can also influence the data distribution even if a huge amount of data is acquired. The ability to influence the early design of foundational technology is also highly correlated with resource access and marginalisation. Since the underlying technologies behave differently between cultures and demographics, the established shift in data distribution from model collapse (Shumailov et al., 2024; Wyllie et al., 2024) will manifest as different tangible impacts on various cultures in the real world.

Turning now towards our recommendations, we can group these into three categories: 1) data, 2) detection, 3) decreas-

ing. Regarding data, we agree with prior work on AI fairness that data curation remains an extremely valuable tool for avoiding bias and harm in LLMs (Rajab et al., 2025). The formation of datasets which have been sequestered and have guaranteed real data appears to be an imperative step (Schaeffer et al., 2025). As Schaeffer et al. (2025) note, many of these datasets already exist from the work conducted by frontier AI labs and were created naturally through the development of the early LLMs and foundation models. This remains a solution only for the languages which have already seen significant development and are likely high-resource languages. Clearly, the existing range of well-curated datasets for thousands of low-resource languages (Agić & Vulić, 2019; Siminyu & Freshia, 2020; Rajab et al., 2025) would be insufficient to preserve the authenticity of the cultures they represent if model collapse were to become a larger issue. Thus, it should be the goal of policymakers, governments and academic institutions to construct such datasets which are guaranteed to contain real data and are representative of culture. How to systematically approach such a task remains a key open question.

The second category revolves around the clear need to both detect synthetic data within datasets, so that curation can be performed more accurately, and to be able to detect model collapse when it happens. In this case, we echo the important point raised by Schaeffer et al. (2025) that clear definitions are needed and that ideally support tractable measurements for model collapse or correlates of it. How feasible it is to detect synthetic data remains to be seen, and given the challenges already shown (Doughman et al., 2025), this should not be relied upon in isolation.

Finally, strategies to decrease model collapse would aim to fix the distributional shift which results from training on synthetic data. A number of strategies may be feasible and would follow from the first two steps once we understand the problem and have data which is usable. Some places to start may be in self-supervised learning or reinforcement learning to promote robust features (Hendrycks et al., 2019) and out-of-distribution generalisation (Feng et al., 2024).

In conclusion, this work adds to the growing discussion and line of position papers (Farnadi et al., 2024; Kandpal & Raffel, 2025; Schaeffer et al., 2025) related to model collapse. While uncertainty remains around the severity and prevalence of model collapse, we argue that this will differ between communities and disproportionately affect low-resource and marginalised communities. Thus, at least some of the most vulnerable communities will face the worst case scenario of model collapse. Consequently, our position is that significantly more effort from the low-resource ML and AI fairness communities is *needed immediately* to understand how to measure and combat model collapse.

Acknowledgments

SSM was supported by the Wallenberg AI, Autonomous Systems, and Software Program (WASP) and by OpenAI’s Alignment Team, awarded by the UK AI Security Institute, through the Alignment Project. B.R. is a Canadian Institute for Advanced Research Fellow. Computations were performed using High Performance Computing infrastructure provided by the Mathematical Sciences Support unit at the University of the Witwatersrand.

References

- agentlans. High quality multilingual sentences, 2025. URL <https://huggingface.co/datasets/agentlans/high-quality-multilingual-sentences>.
- Agić, Ž. and Vulić, I. Jw300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3204–3210, 2019.
- Ahn, J., Lee, H., Kim, J., and Oh, A. Why knowledge distillation amplifies gender bias and how to mitigate from the perspective of distilbert. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pp. 266–272, 2022.
- Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.
- Alemohammad, S., Casco-Rodriguez, J., Luzi, L., Humayun, A. I., Babaei, H., LeJeune, D., Siahkoobi, A., and Baraniuk, R. Self-consuming generative models go mad. In *The Twelfth International Conference on Learning Representations*, 2023.
- Amol, C., Wanzare, L., and Obuhuma, J. Modelling misinformation in swahili-english code-switched texts. *International Journal of Information Technology and Computer Science*, 17(1):67–80, 2025.
- Bahri, Y., Dyer, E., Kaplan, J., Lee, J., and Sharma, U. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27):e2311878121, 2024.
- Bao, K., Zhang, J., Zhang, Y., Huo, X., Chen, C., and Feng, F. Decoding matters: Addressing amplification bias and homogeneity issue for llm-based recommendation. *CoRR*, 2024.
- Basta, C., Costa-Jussà, M. R., and Casas, N. Evaluating the underlying gender bias in contextualized word embeddings. In *Proceedings of the first workshop on gender bias in natural language processing*, pp. 33–39, 2019.
- Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623, 2021.
- Bertrand, Q., Bose, J., Duplessis, A., Jiralspong, M., and Gidel, G. On the stability of iterative retraining of generative models on their own data. In *The Twelfth International Conference on Learning Representations*, 2024.
- Birhane, A. and Prabhu, V. U. Large image datasets: A pyrrhic win for computer vision? In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1536–1546. IEEE, 2021.
- Bohacek, M. and Farid, H. Nepotistically trained generative image models collapse. In *ICLR 2025 Workshop on Navigating and Addressing Data Problems for Foundation Models*, 2023.
- Boutadjine, A., Harrag, F., and Shaalan, K. Human vs. machine: A comparative study on the detection of ai-generated content. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 24(2):1–26, 2025.
- Brynjolfsson, E., Li, D., and Raymond, L. Generative ai at work. *The Quarterly Journal of Economics*, 140(2): 889–942, 2025.
- Chaka, C. Detecting ai content in responses generated by chatgpt, youchat, and chatsonic: The case of five ai content detection tools. *Journal of Applied Learning and Teaching*, 6(2):94–104, 2023.
- Chen, L., Chen, P., and Lin, Z. Artificial intelligence in education: A review. *IEEE access*, 8:75264–75278, 2020.
- Chisca, A.-V., Rad, A.-C., and Lemnaru, C. Prompting fairness: Learning prompts for debiasing large language models. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pp. 52–62, 2024.
- Choi, A., Akter, S. S., Singh, J., and Anastasopoulos, A. The llm effect: Are humans truly using llms, or are they being influenced by them instead? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 22032–22054, 2024.
- Clark, H. H. and Bangerter, A. Changing ideas about reference. In *Experimental pragmatics*, pp. 25–49. Springer, 2004.

- Clark, H. H. and Krych, M. A. Speaking while monitoring addressees for understanding. *Journal of memory and language*, 50(1):62–81, 2004.
- Cogswell, M., Ahmed, F., Girshick, R., Zitnick, L., and Batra, D. Reducing overfitting in deep networks by decorrelating representations. *arXiv preprint arXiv:1511.06068*, 2015.
- Cui, H., Pehlevan, C., and Lu, Y. M. A solvable model of learning generative diffusion: theory and insights. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.
- Davis, J. L., Williams, A., and Yang, M. W. Algorithmic reparation. *Big Data & Society*, 8(2): 20539517211044808, 2021.
- De Vries, T., Misra, I., Wang, C., and Van der Maaten, L. Does object recognition work for everyone? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 52–59, 2019.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.
- Dohmatob, E., Feng, Y., and Kempe, J. Model collapse demystified: The case of regression. *Advances in Neural Information Processing Systems*, 37:46979–47013, 2024a.
- Dohmatob, E., Feng, Y., Subramonian, A., and Kempe, J. Strong model collapse. *arXiv preprint arXiv:2410.04840*, 2024b.
- Dohmatob, E., Feng, Y., Yang, P., Charton, F., and Kempe, J. A tale of tails: Model collapse as a change of scaling laws. In *International Conference on Machine Learning*, pp. 11165–11197. PMLR, 2024c.
- Doughman, J., Afzal, O. M., Toyin, H. O., Shehata, S., Nakov, P., and Talat, Z. Exploring the limitations of detecting machine-generated text. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 4274–4281, 2025.
- Durrant, P. Formulaicity in an agglutinating language: The case of turkish. 2013.
- Electricity, I. Analysis and forecast to 2026. *International Energy Agency: Paris, France*, 2024.
- Elkhatat, A. M., Elsaid, K., and Almeer, S. Evaluating the efficacy of ai content detection tools in differentiating between human and ai-generated text. *International Journal for Educational Integrity*, 19(1):1–16, 2023.
- Farnadi, G., Havaei, M., and Rostamzadeh, N. Position: Cracking the code of cascading disparity towards marginalized communities. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL <https://openreview.net/forum?id=XDz9leJ9iK>.
- Feng, Y., Dohmatob, E., Yang, P., Charton, F., and Kempe, J. Beyond model collapse: Scaling up with synthesized data requires reinforcement. In *ICML 2024 workshop on theoretical foundations of foundation models*, 2024.
- Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., and Ahmed, N. K. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179, 2024.
- Gerstgrasser, M., Schaeffer, R., Dey, A., Rafailov, R., Korbak, T., Sleight, H., Agrawal, R., Hughes, J., Pai, D. B., Gromov, A., et al. Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data. In *First Conference on Language Modeling*, 2024.
- Geva, T., Goldstein, A., Lary, E., and Levy, C. Do llms exhibit human-like cognitive biases? a large-scale systematic evaluation. *A Large-Scale Systematic Evaluation (September 17, 2025)*, 2025.
- Ghiurău, D. and Popescu, D. E. Distinguishing reality from ai: approaches for detecting synthetic content. *Computers*, 14(1):1, 2024.
- Hendrycks, D., Mazeika, M., Kadavath, S., and Song, D. Using self-supervised learning can improve model robustness and uncertainty. *Advances in neural information processing systems*, 32, 2019.
- Hida, R., Kaneko, M., and Okazaki, N. Social bias evaluation for large language models requires prompt variations. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 14507–14530, 2025.
- Huang, K., Yin, H., Huang, H., and Gao, W. Towards green ai in fine-tuning large language models via adaptive back-propagation. In *The Twelfth International Conference on Learning Representations*.
- Iftikhar, S., Alsamhi, S. H., and Davy, S. Enhancing sustainability in llm training: Leveraging federated learning and parameter-efficient fine-tuning. *IEEE Transactions on Sustainable Computing*, 10(6):1158–1172, 2025.
- Janzen, T. and Shaffer, B. Intersubjectivity in interpreted interactions: The interpreter’s role in co-constructing meaning. In *The shared mind: Perspectives on intersubjectivity*, pp. 333–355. John Benjamins Publishing Company, 2008.

- Jarvis, D., Klein, R., Rosman, B., and Saxe, A. M. Compositionality and systematicity emerge from iterated learning in deep linear networks. *Proceedings of the National Academy of Sciences*, 123(19):e2509739123, 2026.
- Jegham, N., Abdelatti, M., Koh, C. Y., Elmoubarki, L., and Hendawi, A. How hungry is ai? benchmarking energy, water, and carbon footprint of llm inference. *arXiv preprint arXiv:2505.09598*, 2025.
- Jiang, Z., Sun, Z., Shi, W., Rodriguez, P., Zhou, C., Neubig, G., Lin, X., Yih, W.-t., and Iyer, S. Instruction-tuned language models are better knowledge learners. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5421–5434, 2024.
- Jo, E. S. and Gebru, T. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 306–316, 2020.
- Kandpal, N. and Raffel, C. Position: The most expensive part of an llm* should* be its training data. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025.
- Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.
- Kazdan, J., Schaeffer, R., Dey, A., Gerstgrasser, M., Rafailov, R., Donoho, D. L., and Koyejo, S. Collapse or thrive: Perils and promises of synthetic data in a self-generating world. In *International Conference on Machine Learning*, pp. 29469–29494. PMLR, 2025.
- Kirby, S., Cornish, H., and Smith, K. Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, 105(31):10681–10686, 2008.
- Kirby, S., Griffiths, T., and Smith, K. Iterated learning and the evolution of language. *Current opinion in neurobiology*, 28:108–114, 2014.
- Kirsten, E., Habernal, I., Nanda, V., and Zafar, M. B. The impact of inference acceleration on bias of llms. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1834–1853, 2025.
- Kosmyrna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X.-H., Beresnitzky, A. V., Braunstein, I., and Maes, P. Your brain on chatgpt: Accumulation of cognitive debt when using an ai assistant for essay writing task. *arXiv preprint arXiv:2506.08872*, 4, 2025.
- Kudo, T. and Richardson, J. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, 2018.
- Kurita, K., Vyas, N., Pareek, A., Black, A. W., and Tsvetkov, Y. Measuring bias in contextualized word representations. In *Proceedings of the first workshop on gender bias in natural language processing*, pp. 166–172, 2019.
- Lai, V. D., Ngo, N., Veyseh, A. P. B., Man, H., Dernoncourt, F., Bui, T., and Nguyen, T. H. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the association for computational linguistics: EMNLP 2023*, pp. 13171–13189, 2023.
- Lee, J., Cho, K., and Kiela, D. Countering language drift via visual grounding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4385–4395, 2019.
- Li, X. The transition from comprehensive to analytical characteristics of english language. *Theory and Practice in Language Studies*, 8(9):1241–1245, 2018.
- Liu, Q., Zheng, X., Muennighoff, N., Zeng, G., Dou, L., Pang, T., Jiang, J., and Lin, M. Regmix: Data mixture as regression for language model pre-training. In *International Conference on Learning Representations*, volume 2025, pp. 38305–38339, 2025.
- Lu, Y., Singhal, S., Strub, F., Courville, A., and Pietquin, O. Countering language drift with seeded iterated learning. In *International Conference on Machine Learning*, pp. 6437–6447. PMLR, 2020.
- Mohamed, A., Zhang, Y., Vazirgiannis, M., and Shang, G. Lost in the mix: Evaluating llm understanding of code-switched text. *arXiv preprint arXiv:2506.14012*, 2025.
- O’Donnell, J. and Crownhart, C. We did the math on ai’s energy footprint. here’s the story you haven’t heard. *MIT Technology Review*, May 2025. URL <https://www.technologyreview.com/2025/05/20/1116327/ai-energy-usage-climate-footprint-big-tech/>.
- Olaleye, K., Oncevay, A., Sibue, M., Zondi, N., Terblanche, M., Mapikitle, S., Lastrucci, R., Smiley, C., and Marivate,

- V. Afrocs-xs: Creating a compact, high-quality, human-validated code-switched dataset for african languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 33391–33410, 2025.
- Orgad, H., Toker, M., Gekhman, Z., Reichart, R., Szpektor, I., Kotek, H., and Belinkov, Y. Llm know more than they show: On the intrinsic representation of llm hallucinations. In *International Conference on Learning Representations*, volume 2025, pp. 66880–66913, 2025.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21 (140):1–67, 2020.
- Raj, C., Mukherjee, A., Caliskan, A., Anastasopoulos, A., and Zhu, Z. Breaking bias, building bridges: Evaluation and mitigation of social biases in llms via contact hypothesis. In *Proceedings of the AAIL/ACM Conference on AI, Ethics, and Society*, volume 7, pp. 1180–1189, 2024.
- Rajab, J. Effect of tokenisation strategies for low-resourced southern african languages. In *3rd Workshop on African Natural Language Processing*, 2022.
- Rajab, J., Aremu, A., Chimoto, E. A., Dunbar, D., Morrissey, G., Thior, F., Potgieter, L., Ojo, J., Tonja, A. L., Nekoto, W. N., et al. The esethu framework: Reimagining sustainable dataset governance and curation for low-resource languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30763–30776, 2025.
- Rao, A. S., Khandelwal, A., Tanmay, K., Agarwal, U., and Choudhury, M. Ethical reasoning over moral alignment: A case and framework for in-context ethical policies in llms. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 13370–13388, 2023.
- Ren, S., Tomlinson, B., Black, R. W., and Torrance, A. W. Reconciling the contrasting narratives on the environmental impact of large language models. *Scientific Reports*, 14(1):26310, 2024.
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.
- Samsi, S., Zhao, D., McDonald, J., Li, B., Michaleas, A., Jones, M., Bergeron, W., Kepner, J., Tiwari, D., and Gadepally, V. From words to watts: Benchmarking the energy costs of large language model inference. In *2023 IEEE High Performance Extreme Computing Conference (HPEC)*, pp. 1–9. IEEE, 2023.
- Santos, C. F. G. D. and Papa, J. P. Avoiding overfitting: A survey on regularization methods for convolutional neural networks. *ACM Computing Surveys (Csur)*, 54(10s):1–25, 2022.
- Schaeffer, R., Kazdan, J., Arulandu, A. C., and Koyejo, S. Position: Model collapse does not mean what you think. *arXiv preprint arXiv:2503.03150*, 2025.
- Seddik, M. E. A., Chen, S.-W., Hayou, S., Youssef, P., and DEBBAH, M. A. How bad is training on synthetic data? a statistical analysis of language model collapse. In *First Conference on Language Modeling*, 2024.
- Shi, J., Yang, Z., and Lo, D. Efficient and green large language models for software engineering: Literature review, vision, and the road ahead. *ACM Transactions on Software Engineering and Methodology*, 34(5):1–22, 2025.
- Shin, J., Song, H., Lee, H., Jeong, S., and Park, J. C. Ask llms directly, “what shapes your bias?”: Measuring social bias in large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 16122–16143, 2024.
- Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., and Anderson, R. J. The curse of recursion: Training on generated data makes models forget. *CoRR*, 2023.
- Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., and Gal, Y. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024.
- Silva, A., Tambwekar, P., and Gombolay, M. Towards a comprehensive understanding and accurate evaluation of societal biases in pre-trained transformers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2383–2389, 2021.
- Siminyu, K. and Freshia, S. Ai4d-african language dataset challenge. In *Proceedings of the Fourth Widening Natural Language Processing Workshop*, pp. 68–77, 2020.

- Smith, K., Kirby, S., and Brighton, H. Iterated learning: A framework for the emergence of language. *Artificial life*, 9(4):371–386, 2003.
- Spennemann, D. H. Delving into: the quantification of ai-generated content on the internet (synthetic data). *arXiv preprint arXiv:2504.08755*, 2025.
- Stankovic, M., Hirche, E., Kollatzsch, S., and Doetsch, J. N. Comment on: Your brain on chatgpt: Accumulation of cognitive debt when using an ai assistant for essay writing tasks. *arXiv preprint arXiv:2601.00856*, 2025.
- Strubell, E., Ganesh, A., and McCallum, A. Energy and policy considerations for deep learning in nlp. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 3645–3650, 2019.
- Thudi, A., Rovers, E., Ruan, Y., Thrush, T., and Maddison, C. J. Mixmin: Finding data mixtures via convex minimization. In *Forty-second International Conference on Machine Learning*.
- Twyman, M., Keegan, B. C., and Shaw, A. Black lives matter in wikipedia: Collective memory and collaboration around online social movements. In *Proceedings of the 2017 acm conference on computer supported cooperative work and social computing*, pp. 1400–1412, 2017.
- Van Aalst, M. K. The impacts of climate change on the risk of natural disasters. *Disasters*, 30(1):5–18, 2006.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- Wang, Z., Wu, Z., Zhang, Y., Guan, X., Jain, N., Lu, Q., Gupta, S., and Koshiyama, A. Bias amplification: Large language models as increasingly biased media. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 115–132, 2025.
- Wei, Y. and Tyson, G. Understanding the impact of ai-generated content on social media: The pixiv case. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 6813–6822, 2024.
- Whitney, C. D. and Norman, J. Real risks of fake data: Synthetic data, diversity-washing and consent circumvention. In *Proceedings of the 2024 ACM conference on fairness, accountability, and transparency*, pp. 1733–1744, 2024.
- Wilson, K., Ghosh, S., and Caliskan, A. Bias amplification in stable diffusion’s representation of stigma through skin tones and their homogeneity. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pp. 2705–2717, 2025.
- Winata, G. I., Aji, A. F., Yong, Z.-X., and Solorio, T. The decades progress on code-switching research in nlp: A systematic survey on trends and challenges. *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 2936–2978, 2023.
- Woolford, E. Bilingual code-switching and syntactic theory. *Linguistic inquiry*, 14(3):520–536, 1983.
- Wyllie, S., Shumailov, I., and Papernot, N. Fairness feedback loops: training on synthetic data amplifies bias. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2113–2147, 2024.
- Xia, Y., Kim, J., Chen, Y., Ye, H., Kundu, S., Hao, C. C., and Talati, N. Understanding the performance and estimating the cost of llm fine-tuning. In *2024 IEEE International Symposium on Workload Characterization (IISWC)*, pp. 210–223. IEEE, 2024.
- Xu, M., Yin, W., Cai, D., Yi, R., Xu, D., Wang, Q., Wu, B., Zhao, Y., Yang, C., Wang, S., et al. A survey of resource-efficient llm and multimodal foundation models. *CoRR*, 2024.
- Yang, K., Yu, C., Fung, Y. R., Li, M., and Ji, H. Adept: A debiasing prompt framework. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 10780–10788, 2023.
- Yoo, H., Park, C., Yun, S., Oh, A., and Lee, H. Code-switching curriculum learning for multilingual transfer in llms. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 7816–7836, 2025a.
- Yoo, H., Yang, Y., and Lee, H. Code-switching red-teaming: Llm evaluation for safety and multilingual understanding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13392–13413, 2025b.
- Zeighami, S., Shankar, S., and Parameswaran, A. Cut costs, not accuracy: Llm-powered data processing with guarantees. *Proceedings of the ACM on Management of Data*, 3(6):1–26, 2025.
- Zhu, X., Cheng, D., Li, H., Zhang, K., Hua, E., Lv, X., Ding, N., Lin, Z., Zheng, Z., and Zhou, B. How to synthesize text data without model collapse? In *Forty-second International Conference on Machine Learning*, 2025.