Holes in the web

Huge swathes of human knowledge are missing from the internet. By definition, generative AI is shockingly ignorant too

A few years back, my dad was diagnosed with a tumour on his tongue – which meant we had some choices to weigh up. My family has an interesting dynamic when it comes to medical decisions. While my older sister is a trained doctor in Western allopathic medicine, my parents are big believers in traditional remedies. Having grown up in a small town in India, I am accustomed to rituals. My dad had a ritual too. Every time we visited his home village in southern Tamil Nadu, he'd get a bottle of thick, pungent, herb-infused oil from a *vaithiyar*, a traditional doctor practising Siddha medicine. It was his way of maintaining his connection with the kind of medicine he had always known and trusted.

Dad's tumour showed signs of being malignant, so the hospital doctors and my sister strongly recommended surgery. My parents were against the idea, worried it could affect my dad's speech. This is usually where I come in, as the expert mediator in the family. Like any good millennial, I turned to the internet for help in guiding the decision. After days of thorough research, I (as usual) sided with my sister and pushed for the surgery. The internet backed us up.

We eventually got my dad to agree and even set a date. But then, he slyly used my sister's pregnancy as a distraction to skip the surgery altogether. While we pestered him every day to get it

done, he was secretly taking his herbal concoction. And, lo and behold, after several months the tumour actually shrank and eventually disappeared. That whole episode definitely earned my dad some bragging rights.

At the time, I dismissed it as a lucky exception. But recently I've been wondering if I was too quick to dismiss my parents' trust in traditional knowledge, while easily accepting the authority of digitally dominant sources. I find it hard to believe my dad's herbal concoctions worked, but I have also since come to realise that the seemingly all-knowing internet I so readily trusted contains huge gaps – and in a world of AI, it's about to get worse.

The irony isn't lost on me that this dilemma has emerged through my research at a university in the United States, in a setting removed from my childhood and the very context where traditional practices were part of daily life. At Cornell University in New York, I study what it takes to design responsible AI systems. My work has been revealing to me how the digital world reflects profound power imbalances in knowledge, and how this is amplified by generative AI (GenAI). The early internet was dominated by the English language and Western institutions, and this imbalance has hardened over time, leaving whole worlds of human knowledge and experience undigitised. Now with the rise of GenAI – which is trained on this available digital corpus – that asymmetry threatens to become entrenched.

For many people, GenAI is becoming their primary way to learn about the world. A large-scale study published in September 2025, analysing how people have been using ChatGPT since its launch in November 2022, revealed that around half the queries were for practical guidance, or to seek information. These systems may appear neutral, but they are far from it. The most popular models privilege dominant epistemologies (typically Western and institutional) while marginalising alternative ways of

knowing, especially those encoded in oral traditions, embodied practice and the languages considered ‘low-resource’ in the computing world, such as Hindi or Swahili, both spoken by hundreds of millions. By amplifying these hierarchies, GenAI risks contributing to the erasure of systems of understanding that have evolved over centuries, disconnecting future generations from vast bodies of insights and wisdom that were never encoded yet remain essential to human ways of knowing. What’s at stake then isn’t just representation – it’s the resilience and diversity of knowledge itself.

GenAI is trained with massive datasets of text from sources like books, articles, websites and transcripts, hence the name ‘large language model’ (LLM). But this training data is far from the sum total of human knowledge. As well as oral cultures, many languages are underrepresented or absent.

To understand why this matters, we must first recognise that languages serve as vessels for knowledge – they are not merely communication tools, but repositories of specialised understanding. Each language carries entire worlds of human experience and insight developed over centuries: the rituals and customs that shape communities, distinctive ways of seeing beauty and creating art, deep familiarity with specific landscapes and natural systems, spiritual and philosophical worldviews, subtle vocabularies for inner experiences, specialised expertise in various fields, frameworks for organising society and justice, collective memories and historical narratives, healing traditions, and intricate social bonds.

When AI systems lack adequate exposure to a language, they have blind spots in their comprehension of human experience. For example, data from Common Crawl, one of the largest public sources of training data, reveals stark inequalities. It contains more than 300 billion web pages spanning 18 years, but English

dominates with 44 per cent of the content. What's even more concerning is the imbalance between how many people speak a language in the physical world and how much that language is represented in online data. Take Hindi, for example, the third most spoken language globally, spoken by around 7.5 per cent of the world's population. It accounts for only 0.2 per cent of Common Crawl's data. The situation is even more dire for Tamil, my own mother tongue. Despite being spoken by more than 86 million people worldwide, it represents just 0.04 per cent of the data. In contrast, English is spoken by approximately 20 per cent of the global population (including both native and non-native speakers), but it dominates the digital space by an exponentially larger margin. Similarly, other colonial languages such as French, Italian and Portuguese, with far fewer speakers than Hindi, are also better represented online.

The underrepresentation of Hindi and Tamil, troubling as it is, represents just the tip of the iceberg. In the computing world, approximately 97 per cent of the world's languages are classified as 'low-resource'. This designation is misleading when applied beyond computing contexts: many of these languages boast millions of speakers and carry centuries-old traditions of rich linguistic heritage. They are simply underrepresented online or in accessible datasets. In contrast, 'high-resource' languages have abundant and diverse digital data available. A study from 2020 showed that 88 per cent of the world's languages face such severe neglect in AI technologies that bringing them up to speed would require herculean – perhaps impossible – efforts. It wouldn't be surprising if the status quo is not too different even now.

To illustrate the kinds of knowledge missing, let's consider just one example: our understanding of local ecologies. An environmentalist friend once told me something that has stayed with me – a community's connection with their ecology can be

seen through the names they have for their local plants. The more intimate their relationship with their environment, the more detailed and specific their botanical vocabulary becomes. Because plant species are often regionally specific or ecologically unique, knowledge of these plants becomes equally localised. This insight proves remarkably accurate when we examine the research. For instance, one study on medicinal plants in North America, northwest Amazonia and New Guinea found that more than 75 per cent of the 12,495 distinct uses of plant species were unique to just one local language. When a language becomes marginalised, the plant knowledge embedded within it often disappears as well.

Exterior and interior views of a wattle and daub cottage in Bengaluru, India by natural building pioneer firm Thannal

While writing this essay, I spoke to various people about the language gaps in GenAI. One of them was Dharan Ashok, chief architect at Thannal, an organisation dedicated to reviving natural building techniques in India. He echoed that there is a strong connection between language and local ecological knowledge, and that this in turn underpins Indigenous architectural knowledge. While modern construction is largely synonymous with concrete and steel, Indigenous building methods were deeply ecological, he told me. They relied on materials available in the surrounding environment, with biopolymers derived from native plants playing a significant role.

Amid concerns over unsustainable and carbon-intensive contemporary construction practices, Dharan is actively working to recover the lost art of producing biopolymers from local plants. However, he noted that the greatest challenge lies in the fact that this knowledge is largely undocumented and has been passed down orally through native languages. It is often held by just a few elders, and when they pass away, it is lost. Dharan recounted a recent experience of missing the chance to learn

how to make a specific type of limestone-based brick after the last artisan with that knowledge died.

To understand how certain ways of knowing rise to global dominance, often at the expense of Indigenous knowledge, it helps to consider the idea of cultural hegemony developed by the Italian philosopher Antonio Gramsci.

Gramsci argued that power is not maintained solely through force or economic control, but also through the shaping of cultural norms and everyday beliefs. Over time, epistemological approaches rooted in Western traditions have come to be seen as objective and universal, rather than culturally situated or historically contingent. This has normalised Western knowledge as the standard, obscuring the specific historical and political forces that enabled its rise. Institutions such as schools, scientific bodies and international development organisations have helped entrench this dominance.

Epistemologies are not just abstract and cognitive. They are physically embodied around us, with a direct impact on our bodies and lived experiences. To understand why, let's consider an example that contrasts sharply with the kind of Indigenous construction practices that Dharan seeks to revive: high-rise buildings with glass façades in the tropics.

Far from being neutral or purely aesthetic choices, glass buildings reflect a particular epistemological tradition rooted in Western architectural modernism. Originally designed for colder, low-light climates, these buildings were praised for their perceived energy efficiency, allowing ample daylight into interiors and reducing reliance on artificial lighting.

However, when this design is applied in tropical regions, it turns into an environmental contradiction. In places with intense sunlight, studies have shown that glass façades lead to significant indoor overheating and thermal discomfort, even with modern glazing. Rather than conserving energy, these buildings demand it to remain cool.

Yet glass façades have become the ubiquitous face of urban modernity, be it San Francisco, Jakarta or Lagos, regardless of climate or cultural context.

As climate change accelerates, these glass buildings are gleaming reminders of the dangers of knowledge homogenisation and epistemic hierarchies. Ironically, I'm writing this from inside one of those very buildings in Bengaluru in southern India. I sit in cooled air with the soft hum of the air conditioner in my ears. Outside, people saunter through a gentle drizzle. It looks like a normal monsoon afternoon – except the rains arrived weeks ahead of schedule this year, yet another sign of growing climate unpredictability.

In Bengaluru, I see yet another example of the impacts of lost knowledge: water management. How can a city flood severely in May, submerging cars, yet scramble for water even for domestic use in March? While this can be attributed to factors like poor planning and unchecked urbanisation, it also has deep epistemological roots.

A flooded road after heavy rains in Bengaluru, India, 22 October 2024. Photo by Sayan Hazra/Reuters

Bengaluru was once celebrated for its smart water management system, fed by a series of interconnected cascading lakes. For centuries, these lakes were managed by dedicated communities, such as the Neeruganti community (*neeru* means ‘water’ in the Kannada language), who controlled water flow and ensured fair distribution. Depending on the rains, they guided farmers on which crops to grow, often suggesting water-efficient varieties. They also handled upkeep: desilting tanks, planting vegetation to prevent erosion, and clearing feeder channels.

But with modernisation, community-led water management gave way to centralised systems and individual solutions like irrigation from far-off dams and borewells. The Green Revolution of the late 1960s added to this shift, pushing water- and fertiliser-heavy crops developed in Western labs. The Neerugantis were sidelined, and many moved on in search of other work. Local

lakes and canals declined, and some were even built over – replaced with roads, buildings or bus stops.

Experts have realised that the key to saving Bengaluru from its water crisis lies in bringing these lake systems back to life. A social worker I spoke with, who's been involved in several of these projects, said they often turn to elders from the Neeruganti community for advice. Their insights are valuable but their local knowledge is not written down, and their role as community water managers has long been delegitimised. Knowledge exists only in their native language, passed on orally, and is mostly absent from digital spaces – let alone AI systems.

While all my examples so far are drawn from India due to personal familiarity, such hierarchies are widespread, rooted in the global history of imperialism and colonialism. In her book *Decolonising Methodologies* (1999), the Māori scholar Linda Tuhiwai Smith emphasises that colonialism profoundly disrupted local knowledge systems – and the cultural and intellectual foundations upon which they were built – by severing ties to land, language, history and social structures. Smith's insights reveal how these processes are not confined to a single region but form part of a broader legacy that continues to shape how knowledge is produced and valued. It is on this distorted foundation that today's digital and GenAI systems are built.

I recently worked with Microsoft Research, examining several GenAI deployments built for non-Western populations. Observing how these AI models often miss cultural contexts, overlook local knowledge, and frequently misalign with their target community has brought home to me just how much they encode existing biases and exclude marginalised knowledge.

The work has also brought me closer to understanding the technical reasons why such inequalities develop inside the models. The problem is far deeper than gaps in training data. By design, LLMs also tend to reproduce and reinforce the most statistically prevalent ideas, creating a feedback loop that narrows the scope of accessible human knowledge.

Why so? The internal representation of knowledge in an LLM is not uniform. Concepts that appear more frequently, more prominently, or across a wider range of contexts in the training data tend to be more strongly encoded. For example, if pizza is commonly mentioned as a favourite food across a broad set of training texts, the model is more likely to respond with 'pizza' when asked 'What's your favourite food?' Not because the LLM likes pizza, but because that association is more statistically prominent.

More subtly, the model's output distribution does not directly reflect the frequency of ideas in the training data. Instead, LLMs often amplify dominant patterns in a way that distorts their original proportions. This phenomenon can be referred to as 'mode amplification'. Suppose the training data includes 60 per cent references to pizza, 30 per cent to pasta, and 10 per cent to biriyani as favourite foods. One might expect the model to reproduce this distribution if asked the same question 100 times. However, in practice, LLMs tend to overproduce the most frequent answer. Pizza may appear more than 60 times, while less frequent items like biriyani may be underrepresented or omitted altogether. This occurs because LLMs are optimised to predict the most probable next 'token' (the next word or word fragment in a sequence), which leads to a disproportionate emphasis on high-likelihood responses, even beyond their actual prevalence in the training corpus. Together, these two principles – uneven internal knowledge representation and mode

amplification in output generation – help explain why LLMs often reinforce dominant cultural patterns or ideas.

This uneven encoding gets further skewed through reinforcement learning from human feedback (RLHF), where GenAI models are fine-tuned based on human preferences. This inevitably embeds the values and worldviews of their creators into the models themselves. Ask ChatGPT about a controversial topic and you'll get a diplomatic response that sounds like it was crafted by a panel of lawyers and HR professionals who are overly eager to please you. Ask Grok the same question and you might get a sarcastic quip followed by a politically charged take that would fit right in at a certain tech billionaire's dinner party.

Commercial pressures add another layer entirely. The most lucrative users – English-speaking professionals willing to pay $20-200 monthly for premium AI subscriptions – become the implicit template for 'superintelligence'. These models excel at generating quarterly reports, coding in Silicon Valley's preferred languages, and crafting emails that sound appropriately deferential to Western corporate hierarchies. Meanwhile, they stumble over cultural contexts that don't translate to quarterly earnings.

It should not come as a surprise that a growing body of studies shows how LLMs predominantly reflect Western cultural values and epistemologies. They overrepresent certain dominant groups in their outputs, reinforce and amplify the biases held by these groups, and are more factually accurate on topics associated with North America and Europe. Even in domains such as travel recommendations or storytelling, LLMs tend to generate richer and more detailed content for wealthier countries compared with poorer ones. There are at least 50 more similar studies I could cite here.

And beyond merely *reflecting* existing knowledge hierarchies, GenAI has the capacity to *amplify* them, as human behaviour changes alongside it. The integration of AI overviews in search engines, along with the growing popularity of AI-powered search engines such as Perplexity, underscores this shift. A recent study found that US Google users were less likely to click on search results when an AI summary appeared alongside, indicating a change in people's behaviour. Previously, people had to browse multiple links to compare viewpoints and gather comprehensive information. Now, they can read AI-generated summaries.

As AI-generated content has started to fill the internet, it adds another layer of amplification to ideas that are already popular online. The internet, as the primary source of knowledge for AI models, becomes recursively influenced by the very outputs those models generate. With each training cycle, new models increasingly rely on AI-generated content, reinforcing prevailing narratives and further marginalising less prominent perspectives. This risks creating a feedback loop where dominant ideas are continuously amplified while long-tail or niche knowledge fades from view.

The AI researcher Andrew Peterson describes this phenomenon as 'knowledge collapse', a gradual narrowing of the information humans can access, along with a declining awareness of alternative or obscure viewpoints. As LLMs are trained on data shaped by previous AI outputs, underrepresented knowledge can become less visible – not because it lacks merit, but because it is less frequently retrieved or cited. Peterson also warns of the 'streetlight effect', named after the joke where a person searches for lost keys under a streetlight at night because that's where the light is brightest. In the context of AI, this would be people searching where it's *easiest* rather than where it's most *meaningful*. Over time, this would result in a degenerative narrowing of the public knowledge base, driven not by

censorship but convenience and algorithms.

Across the globe, GenAI is also becoming part of formal education, used to generate learning content and support self-paced education through AI tutors. For example, the Karnataka state government, home to the city of Bengaluru, has partnered with the US-based non-profit Khan Academy to deploy Khanmigo, an AI-powered learning assistant, into schools and colleges. I would be surprised if Khanmigo holds the insights of elder Neerugantis, grounded in local knowledge and practices, needed to teach school students in Karnataka how to care for their own water ecologies.

All this means that, in a world where AI increasingly mediates access to knowledge, future generations might lose connection with vast bodies of experience, insight and wisdom. AI developers might argue that this is simply a data problem, solvable by incorporating more diverse sources into training datasets. While that might be technically possible, the challenges of data sourcing, prioritisation and representation are far more complex than such a solution implies.

A conversation I had with a senior leader involved in designing and overseeing the development of an AI chatbot that serves more than 8 million farmers across four countries in Asia and Africa brought this into focus. The system provides agricultural advice based mostly on databases from government advisories and international development organisations, which tend to rely on research literature. The senior leader acknowledged how many local practices that might be effective are *still* excluded from the chat responses since they are not documented in the research literature.

The rationale isn't that research-backed advice is always right or risk-free. It's that it offers a defensible position if something goes

wrong. In a system this large, leaning on recognised sources is seen as the safer bet, protecting an organisation from liability while sidelining knowledge that hasn't been vetted through institutional channels. So the decision is more than just technical. It's a compromise shaped by the structural context, not based on what's most useful or true.

This structural context doesn't just shape institutional choices. It also shapes the kinds of challenges I heard about in my conversation with Perumal Vivekanandan, founder of the non-profit organisation Sustainable-agriculture and Environmental Voluntary Action (SEVA). His experiences highlight the uphill battle faced by those working to legitimise Indigenous knowledge.

Formed in 1992, SEVA focuses on preserving and disseminating Indigenous knowledge in agriculture, animal husbandry and the conservation of agricultural biodiversity. Over the years, Vivekanandan has documented more than 8,600 local practices and adaptations, travelling village to village.

Still, the work constantly runs into systemic roadblocks. Funders often withhold support, questioning the scientific legitimacy of the knowledge SEVA seeks to promote. When SEVA turns to universities and research institutions to help validate this knowledge, they often signal a lack of incentives to engage. Some even suggest that SEVA should fund the validation studies itself. This creates a catch-22: without validation, SEVA struggles to gain support; but without support, it can't afford validation. The process reveals a deeper challenge: finding ways to validate Indigenous knowledge within systems that have historically undervalued it.

SEVA's story shows that while GenAI may be accelerating the erasure of local knowledge, it is not the root cause. The marginalisation of local and Indigenous knowledge has long been driven by entrenched power structures. GenAI simply puts this process on steroids.

We often frame the loss of Indigenous knowledge as a tragedy only for the local communities who held it. But ultimately, the loss is not just theirs to bear, but belongs to the world at large too.

In our social structures, we may assign hierarchical value to certain communities or types of knowledge, but natural ecology reveals a different logic. Every local element plays a vital role in sustaining global balance. As the forester Peter Wohlleben illustrates in *The Hidden Life of Trees* (2015), natural systems are deeply interdependent, often in ways that are invisible to the casual observer. He offers a powerful example from Yellowstone National Park in the US. When wolves were eradicated from the park in the early 20th century, it led to a series of unexpected ecological consequences. Without wolves to keep their numbers in check, deer populations exploded. The deer overgrazed vegetation and altered the landscape. Riverbanks eroded, tree growth stalled, and the broader ecosystem suffered. When wolves were reintroduced decades later, the system began to heal. Vegetation rebounded, songbirds returned, and even the behaviour of rivers changed.

Wohlleben's broader point is that the health of a system depends on the presence of all its parts, even those that might seem inconsequential. The same principle applies to human knowledge. The disappearance of local knowledge is not a trivial loss. It is a disruption to the larger web of understanding that sustains both human and ecological wellbeing. Just as biological

species have evolved to thrive in specific local environments, human knowledge systems are adapted to the particularities of place. When these systems are disrupted, the consequences can ripple far beyond their point of origin.

Wildfire smoke doesn't care about transgressing zip codes. Polluted water doesn't pause at state lines. Rising temperatures ignore national borders. Infectious germs don't have visa waiting periods. Whether we acknowledge it or not, we are enmeshed in shared ecological systems where local wounds inevitably become global aches.

The biggest contradiction for me in writing this essay is that I'm trying to convince readers of the legitimacy and importance of local knowledge systems while I myself remain unconvinced about my dad's herbal concoctions. This uncertainty feels like a betrayal of everything I've argued for. Yet maybe it's exactly the kind of honest complexity we need to navigate.

I have my doubts about whether Indigenous knowledge truly works as claimed in every case. Especially when influencers and politicians invoke it superficially for likes, views or to exploit identity politics, generating misinformation without sincere enquiry. However, I'm equally wary of letting it disappear. We might lose something valuable, only to recognise its worth much later – perhaps with the aid of artificial superintelligence. But what's the collateral damage of that process? An ecological collapse we could have prevented?

The climate crisis is revealing cracks in our dominant knowledge paradigms. Yet at the same time, AI developers are convinced that their technology will accelerate scientific progress and solve our greatest challenges. I really want to believe they're right. But several questions remain: can we move towards this

technological future while authentically engaging with the knowledge systems we've dismissed, with genuine curiosity beyond tokenism? Or will we keep erasing forms of understanding through the hierarchies we've built, and find ourselves scrambling to colonise Mars because we never learned to listen to those who knew how to live sustainably on Earth?

Maybe the intelligence we most need is the capacity to see beyond the hierarchies that determine which knowledge counts. Without that foundation, regardless of the hundreds of billions we pour into developing superintelligence, we'll keep erasing knowledge systems that took generations to develop.

I don't know if my dad's herbal concoctions worked. But I'm learning that acknowledging that I don't know might be the most honest place to start.

Sincere gratitude to my PhD advisors Aditya Vashistha and Rene Kizilcec, my collaborators at Microsoft Research, and my most amazing friends without whom this essay wouldn't have been possible.

aeon.co 13 October 2025