The long read

What AI doesn't know: we could be creating a global 'knowledge collapse'

As GenAI becomes the primary way to find information, local and traditional wisdom is being lost. And we are only beginning to realise what we're missing

This article was originally published as 'Holes in the web' on Aeon.co

By Deepak Varuvel Dennison

A vegetable seller – and signs in Hindi – in Old Delhi, India. Photograph: Frank Bienewald/Alamy

Tue 18 Nov 2025 05.00 GMT

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 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. The whole episode 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 accepting the authority of digitally dominant sources. I find it hard to believe that my dad's herbal concoctions worked, but I have also come to realise that the seemingly all-knowing internet I so readily trusted contains huge gaps – and that, 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, New York, I study what it takes to design responsible AI systems. My work has been revealing, showing 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 emerging as the 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 ways of knowing (typically western and institutional) while marginalising alternatives, especially those encoded in oral traditions, embodied practice and languages considered "low-resource" in the computing world, such as Hindi or Swahili.

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 insight and wisdom that were never encoded yet remain essential, 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 on massive datasets of text from sources such as 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, with oral cultures and even languages underrepresented or absent.

To understand why this matters, we must first recognise that languages serve as vessels for knowledge. 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. Data from Common Crawl, one of the largest public sources of training data, reveals stark inequalities. It contains more than 300 billion webpages spanning 18 years, but English, which is spoken by approximately 19% of the global population, dominates, with 45% of the content. However, there can be an alarming imbalance between a language's demographic size and how well that language is represented in online data. Take Hindi, the third most popular language globally, spoken by about 7.5% of the world's population. It accounts for only 0.2% 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% of the data.

In the computing world, approximately 97% 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. A study from 2020 showed that 88% of the world's languages face such severe neglect in AI technologies that bringing them up to speed would be a herculean - perhaps impossible - effort.

To illustrate the kinds of knowledge missing, let's consider one example: our understanding of local ecologies. An environmentalist friend once told me something that has stayed with me - a community's connection with its ecology can be seen through the detailed and specific names it has for local plants. Because plant species are often regionally specific or ecologically unique, knowledge of these plants becomes equally localised. When a language becomes marginalised, the plant knowledge embedded within it often disappears as well.

A wattle-and-daub cottage designed by Indian architects Thannal, who specialise in natural building techniques. Photograph: Thannal

While writing this essay, I spoke to various people about the language gaps in GenAI

While writing this essay, I spoke to various people about the language gaps in GenAI – among them Dharan Ashok, chief architect at Thannal, an organisation dedicated to reviving natural building techniques in India. He agreed 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 relied on materials available in the surrounding environment.

Amid concerns over unsustainable and carbon-intensive construction, Dharan is actively working to recover the lost art of producing biopolymers from local plants. 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 an experience of missing the chance to learn how to make a specific type of limestone-based brick when 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 maintained not 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. This has normalised western knowledge as the standard, obscuring the 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 all around us, with a direct impact on our bodies and lived experiences. To understand how, 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 facades in the tropics.

Far from being neutral or purely aesthetic choices, glass buildings reflect a 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.

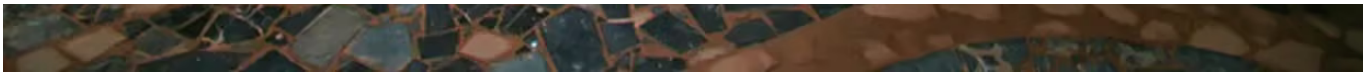
However, when this design is applied in tropical regions, it turns into an environmental contradiction. In places with intense sunlight, studies have shown that glass facades lead to significant indoor overheating and thermal discomfort,

even with modern glazing. Rather than conserving energy, these buildings demand more energy use to remain cool.

Yet glass facades have become the face of urban modernity, whether in San Francisco, Jakarta or Lagos – regardless of climate or cultural context. As climate breakdown accelerates, these glass buildings are gleaming reminders of the dangers of knowledge homogenisation. 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 in the drizzle, it seems to be a normal monsoon afternoon, except the rains arrived weeks early this year – 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 poor planning and unchecked urbanisation play their part, the issue also has epistemological roots.

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 groups, 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.



Interior of Thannal's wattle-and-daub cottage. Photograph: Thannal

But with modernisation, community-led water management gave way to centralised systems and individual solutions such as irrigation from far-off dams and bore wells. The "Green Revolution" of the late 1960s - when India embraced modern industrial agriculture - 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 Decolonizing Methodologies (1999), the Māori scholar Linda Tuhiwai Smith emphasises that colonialism profoundly disrupted local knowledge systems - and the cultural and intellectual foundations on 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, when asked "what's your favourite food?", the model is more likely to respond with "pizza" 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 or ideas in a way that distorts their original proportions. This phenomenon can be referred to as "mode amplification".

The glass facade of DLF's Gateway Tower in Gurugram, India. Photograph: Danny Lehman/Getty Images

Suppose the training data includes 60% references to pizza, 30% to pasta and 10% to biryani 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 such as biryani 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.

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, X's AI chatbot, 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.

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.

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. 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.

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 state government of Karnataka, home to the city of Bengaluru, has partnered with the US-based nonprofit Khan Academy to deploy Khanmigo, an AI-powered learning assistant, in 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 water ecologies.

All this means that, in a world where AI increasingly mediates access to knowledge, future generations may lose connection with vast bodies of experience, insight and wisdom. AI developers could 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.

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

Liquid-cooled servers at the Global Switch data centre, London. Photograph: Bloomberg/Getty Images

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 is 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 nonprofit 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 in India. 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. Potential 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 hold it. But ultimately, the loss is not just theirs to bear, but belongs to the world at large.

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 postcodes. 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

I 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 here. 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 or to exploit identity politics, generating misinformation without sincere inquiry. However, I'm equally wary of letting it disappear. We may lose something valuable, only to recognise its worth much later. And 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: are we capable of moving 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 I don't know might be the most honest place to start.

The long read

The Audio Long Read

The Pushkin job: unmasking the thieves behind an international rare books heist – podcast

The Audio Long Read

'The jobless should lead the attack': a radical Jamaican journalist in 1920s London – podcast

'We were forced to burn bodies': will survivors of the Tadamon massacres see justice?

5d ago

More from News

Exclusive

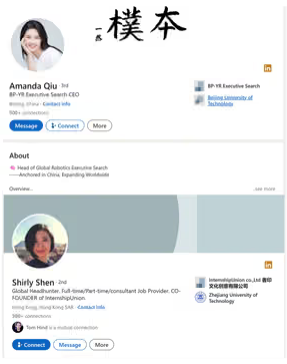
'Deeply shocking': Nigel Farage faces fresh claims of racism and antisemitism at school

4h ago

Espionage

MI5 names two LinkedIn headhunters in alert to MPs and peers about Chinese espionage

2h ago



Immigration and asylum

Mahmood faces calls for compassion and clarity over hardline asylum policies

3h ago

Most viewed