



OPINION AND ANALYSIS

Editorial date: 27 October 2026

The Words That Are Not on the Internet

From Finland to Saudi Arabia: Why the greatest national AI project might begin with a grandmother's voice, a student's notebook, and the memory of an entire society

In the summer of 2020, Finland made an unusual request of its citizens: it asked them to speak.

They were not asked to read formal texts, recite sentences carefully prepared in a laboratory, or use the ideal Finnish taught in schools.

They were asked to speak as they did in life.

About sport, food, memories, the things they loved, and the things about which they disagreed. They spoke in their natural dialects, with their hesitations and passing mistakes, and with the silences, laughter, and expressions that no dictionary records.

The initiative was called *Lahjoita puhetta*—Donate Speech. The Finnish public broadcaster Yle launched it in cooperation with the Language Bank of Finland, Kielipankki, and research and technology partners.

Its goal was to collect ten thousand hours of natural Finnish speech so that technologies could be developed to understand people as they actually speak.

When the campaign closed on 6 March 2024, the Language Bank of Finland announced that it had collected more than three thousand hours of speech. The research study published in 2022, however, describes the 2021 release of the corpus as containing about 3,600 hours and more than twenty thousand speakers—an estimate that treated each application-client identity as one speaker, rather than a count based on verified personal identities.

The initiative did not reach its original quantitative target, but it created something more important than a collection of audio recordings.

It turned ordinary citizens into participants in building the knowledge infrastructure that future technologies will need.

The study *Lahjoita puhetta: A Large-Scale Corpus of Spoken Finnish* documents how everyday speech became a resource for research in automatic speech recognition (ASR), dialect comprehension, and the analysis of linguistic and social features.

What is most striking is that the Finns understood something very simple: however brilliant an engineer may be, they cannot sit in a laboratory and invent the way millions of people speak.

An engineer can design the system that learns from them, but the people must provide the material from which it learns.

That Finnish story leads me to a question that I believe deserves an unprecedented Saudi national initiative:

What if every Saudi home, every school and university, every profession and institution could help create the largest authentic human-data infrastructure representing our language, culture, and knowledge?

It would not be built by gathering what others have written about us on the internet, or by generating millions of texts that resemble the output of today's models.

It would begin with people themselves.

With the words they say, the meanings they understand, the documents they know how to read, and the details a machine cannot learn because they have never been written in a source it can reach.

Half a Century of the Internet, and Life Is Still Larger

Imagine asking a language model to explain the Arabic expression “ya ba’d hayyi.”

It might provide an acceptable dictionary explanation, perhaps describing it as an expression of affection and esteem.

But knowing what an expression means is one thing. Understanding its place in a human relationship is another.

An older man might say it to his grandson, a woman to her sister, or one friend to another in a moment of gratitude. Its tone and meaning may shift with the context, the place, and the relationship between the speakers.

Now imagine an engineer at an industrial plant using a term familiar to colleagues for a recurring fault that does not appear under that name in the company manual.

Or a Bedouin woman who can name subtle changes in the wind and the direction of clouds, where a general encyclopedia offers only an approximate description.

Or a nurse who knows how patients describe pain and how that everyday language differs from the terms a physician enters in a clinical record.

This knowledge is not absent because it has no value.

In many cases, it is absent because it never entered the digital system from which machines learn.

That raises a question deeper than the shortage of Arabic text.

Statistics has a problem known as sampling bias. If data is gathered from a segment that does not represent the wider society, a model may become excellent at understanding that segment while remaining less capable of understanding everyone the sample left out.

With language models, the problem becomes more complex.

The internet is not a balanced copy of society.

Some people write more than others. Some professions produce abundant digital text. Some groups leave only limited traces. Some dialects appear only in private conversations that must not be collected without permission.

Sometimes the people who possess the most valuable local knowledge are the least present in digital data.

A model can therefore know a great deal about the world while knowing very little about a human experience close to the people who use it.

It knows what has been published, which is not necessarily the same as what has been known.

That is the opportunity: Saudi Arabia does not need only to collect more of what machines already know. It needs to enable people to teach them what they have never been able to reach.

When People Became Partners in Training Machines

Finland was not an isolated exception.

In India, home to a vast number of languages and dialects, the government launched the BHASHINI initiative. It includes Bhasha Daan, a language-donation program through which citizens can contribute texts, audio recordings, and translations.

In March 2026, India's Ministry of Electronics and Information Technology announced that more than ten thousand contributors had joined Bhashini Samudaye, a platform that brings together individuals, universities, companies, and laboratories to support the creation of language resources and data annotation.

The idea is not simply for a government to collect citizens' voices and place them in an archive.

It is to build a resource shared by universities, researchers, institutions, and companies, one from which different technologies can benefit.

That may be the difference between a short campaign to collect data and the establishment of a permanent national infrastructure.

In South Korea, the project took a broader institutional form.

The National Information Society Agency (NIA) created AI-Hub, which has made more than nine hundred datasets available for AI training since 2017, according to the agency.

The data covers Korean language, audio, images, medicine, industry, agriculture, and other fields.

The important detail is that Korea did not treat data annotation as a closed technical activity.

Within the government programs that built the training data made available through AI-Hub, direct jobs and paid crowdsourced work were created in data collection, processing, and quality verification. NIA has documented that some workers, with experience, were able to move from annotation into review and project management.

An important economic insight appears here: **teaching a machine can become a way to train people and equip them with new professional skills.**

Singapore took another approach that is no less interesting.

AI Singapore, working with Google, launched the Aquarium platform within the SEA-LION ecosystem to make it easier for communities and researchers to contribute to datasets for Southeast Asian languages.

The project was designed not merely to gather Malay, Thai, or Indonesian, but to improve access to resources that reflect the linguistic and cultural reality of the region, and to develop tools for participation, evaluation, and incentives.

Even its name, Aquarium, suggests a beautiful metaphor: language is not one large fish placed in a tank. It is a living system of diversity and relationships.

Estonia chose a different, educational entry point.

It launched AI Leap to introduce artificial intelligence into public education in a way that helps teachers and students develop critical thinking and learning skills.

During the 2025–2026 academic year, 154 of the country's 156 upper-secondary schools took part in the program, which involved about five thousand teachers and twenty thousand students.

This is not a project for collecting national training data in the sense I am proposing here.

But it shows that a country can turn its schools into a broad national network for learning and participating in the development of new technological practices.

In the United States, we find a remarkable example in an unexpected place: the Library of Congress.

The Library launched *By the People*, a project that invites volunteers to transcribe handwritten manuscripts and historical documents, review other people's transcriptions, and tag materials.

A volunteer might be a student, a retired person, or an amateur historian.

With every line they read, a historical document that a computer struggles to understand becomes searchable text that can be studied and used.

For artificial intelligence, this kind of work shows the value of turning unstructured archives into usable digital knowledge.

Five experiences, and five different methods, converge on one idea:

The data artificial intelligence needs is not always a product that can be bought ready-made. Sometimes it is knowledge scattered among millions of people, and the innovation lies in building the means to gather it, protect the rights of those who contributed it, and verify its quality.

“Diwan”: A National Project Written by the People

I imagine a Saudi initiative with a proposed name: the **National Saudi Diwan of Human Knowledge**.

I do not imagine it as an app that people are told to download and then forget, or as a seasonal competition that ends with prizes being handed out.

I imagine it as a permanent national participatory data infrastructure, with participation proposed for public and higher education, the private sector, nonprofit associations, research centers, and citizens and residents who choose to contribute.

Its purpose would be to create the largest documented Saudi resource of authentic, annotated human data, with carefully considered linguistic, cultural, professional, and social diversity.

At the heart of the proposed Diwan would be a digital portal offering different levels of participation.

A student could log in and find a task suited to their age and knowledge: explaining a local expression, writing a description of a scene, reading a sentence aloud, or reviewing text extracted automatically from a licensed document.

A teacher could help classify educational texts, determine their level of difficulty, and explain where they might be ambiguous.

A physician or nurse could review health terms and expressions using examples created specifically for training, or educational data anonymized in accordance with the Implementing Regulations and only after the risk of re-identification had been assessed, rather than exposed patient records.

An engineer could clarify professional vocabulary, classify images of equipment that may lawfully be used, or describe the steps of a general industrial procedure.

A woman interested in social history could, with her consent, record an explanation of a proverb, a local custom, or the name of an old household tool.

A Bedouin man could describe vocabulary connected with place, weather, and wildlife, while a city resident could contribute expressions that arose in a modern urban setting.

No dialect would have predetermined priority over another, and no social register would be treated as inherently superior.

The point would not be to impose a single image of a Saudi person on the model, but to let it see the diversity of the society from which it learns.

Residents could also contribute their languages and experience, particularly in everyday interactions, work, services, and translation between languages.

That could help develop models with a broader understanding of social and professional contexts in the Kingdom.

The great advantage is that a contributor would not need to know programming or machine learning.

It would be enough to know something real that could be passed on to others.

A grandmother may not know what a transformer architecture is, but she knows stories and expressions that no transformer engineer can invent, however much computing power they possess.

At that point, the initiative becomes more than a source of data.

It becomes a social recognition that knowledge does not belong exclusively to people with academic credentials, and that everyday experience itself can be valuable material for developing technology.

The Day Correcting One Word Becomes a National Act

Imagine the Diwan's interface.

A volunteer sees an image of an old Arabic document that has been made available with the owner's permission.

An optical character recognition (OCR) tool tries to extract the text, but misreads a word because of the handwriting or damage to the paper.

The volunteer sees the word and corrects it.

The correction is then shown to a second contributor. If the two readings differ, the word is referred to a more experienced reviewer.

In a few minutes, the participants have performed work that could help create a dataset for improving systems that read Arabic documents.

Elsewhere, a licensed image of a traditional craft object or an agricultural tool is displayed.

The portal asks contributors to describe it, identify the name used in their region, explain its purpose, and perhaps add an alternative name from another dialect.

The image is then linked to the appropriate descriptions, turning it into training material for vision-language learning.

In a third task, a volunteer listens to a recording donated by someone with their consent, then reviews the transcript produced by an automatic speech recognition system.

Technically, this is not a hypothetical idea.

Google Crowdsource allows users to contribute to image label verification, handwriting recognition, translation review, image description, audio validation, and other tasks.

The platform awards different numbers of points according to the task and its difficulty, with extra points for some contributions confirmed by reviews from other participants.

Mozilla Foundation's Common Voice project allows people to record their voices and other volunteers to review those recordings, creating multilingual voice datasets that can support the development of speech technologies.

The idea has proved feasible.

But a Saudi initiative could go further than copying it.

It could connect contributions to fields of knowledge, create pathways to advanced levels, develop credentials in linguistic, audio, and visual annotation, and turn participation into a route toward specialization in data science.

Someone who begins by correcting one word in a manuscript might, years later, become a researcher developing a system capable of reading thousands of manuscripts.

Can a Machine Smell a Machine?

Here an unavoidable problem appears.

If participation is opened to a broad public, some contributors may try to use models such as GPT, Claude, or DeepSeek to generate ready-made text and then submit it as original human work.

This is not a marginal risk.

If the aim is to create a knowledge resource that originates with people, only for it to fill with text generated by commercial models, we may end up reintroducing knowledge that already existed in earlier models instead of broadening the representation of human experience.

I therefore propose that the initiative's portal include a dedicated research system for examining the provenance and quality of contributions.

The system could draw on n-gram analysis, statistical style signatures known as stylometric fingerprints, linguistic perplexity, and near-duplicate detection.

It could, for example, look for unusual repetition of particular constructions, long passages that match published material, or patterns that occur more frequently in machine output than in human reference sets.

But there is a scientific paradox here that must not be ignored.

You cannot prove that a text is human merely because it does not resemble synthetic text, just as you cannot accuse a person of using a machine merely because their style is consistent.

In the writing samples and seven detectors they tested, a 2023 study published in *Patterns* by Stanford researcher Weixin Liang and colleagues found that detectors of generated text could be biased against writing by people whose first language is not English, classifying human texts as generated.

A study led by Debora Weber-Wulff and published in the *International Journal for Educational Integrity* also found limited reliability among several widely used detection tools, with high false-positive and false-negative rates in some cases.

The proposed content-screening system should therefore never act as a judge that excludes a contribution on the strength of a single probability score.

It should instead form one part of a broader architecture for establishing data provenance.

That architecture would begin with the contribution itself: a clear declaration of how it was created, an optional documented editing trail, tasks completed directly inside the portal, independent human reviews, and random verification samples where necessary, all without intrusive surveillance of volunteers.

Originally human-authored data would be kept distinct from human-verified AI outputs and from synthetic data.

All three can be useful, but they are not the same, and their differences should not disappear inside a single repository.

I also propose a provenance record for every dataset, explaining how it was collected, the basis on which it was licensed, the nature of its review, and the level of confidence in its quality.

This is consistent with the idea advanced by Timnit Gebru and her colleagues in the well-known paper [Datasheets for Datasets](#), which called for datasets to be documented in the way industries document the characteristics of products and components.

The aim is not to build a tool that boasts of detecting the “smell of AI.”

It is to build a knowledge resource in which we know the history of every part, what we can learn from it, and the limits of our confidence in it.

A smell is an impression.

Documented provenance is knowledge.

Material	Origin	What should be documented?
Human-authored data	Created by a person	Source, creation process and rights
Human-verified AI outputs	Machine output followed by human review	Generator, reviewer’s role and review limits
Synthetic data	Machine generation	Generation method, purpose and use limits

A conceptual summary of the distinctions in this section; no category is automatically accurate or suitable.

When Disagreement Is Not an Error in the Data

One of the most important principles for the Diwan should be that not every difference among contributors is treated as an error to be erased.

Two people may see the same image and describe it differently.

Two speakers from different regions may disagree about an expression because the same word carries different meanings in their respective communities.

Two researchers may read an ambiguous phrase and arrive at two reasonable interpretations.

Many conventional annotation systems try to reach one final ground-truth label, perhaps through majority voting or the decision of a specialist reviewer.

That is appropriate for questions with a clear objective answer, such as reading a number in an undamaged document or determining whether a particular object appears in an image.

It may be less appropriate for language and culture.

Not every disagreement means that one side is wrong.

In their 2022 paper *Investigating Reasons for Disagreement in Natural Language Inference*, Nan-Jiang Jiang and Marie-Catherine de Marneffe studied why reviewers differed in their interpretations of relationships between texts. They found that some disagreements arose from genuine possibilities in meaning rather than random annotation errors.

That matters enormously for a project intended to represent Saudi society.

Regional differences in the use of a word may be precisely the information the machine needs.

Differences between generations in the meaning of an expression may be evidence of a linguistic shift worth documenting.

The annotation system should therefore be able to retain multiple answers when they are meaningful and connect them to the relevant geographic, age, or professional context without turning unnecessary personal data into exposed records.

This does not mean assuming that every opinion is equally valid in every subject.

Medical, scientific, and legal questions require specialist verification, while plurality may be legitimate when describing linguistic usage or a social tradition.

Some good data does not teach a model a single answer. It teaches the model why people differ in their answers.

That is one of the greatest possible gains from moving beyond the collection of text to the collection of annotated human knowledge.

How a Volunteer Hour Could Become a Knowledge Asset

I now come to an element that I consider decisive to the initiative's success: how do we make participation a social habit rather than a temporary burst that ends with a media campaign?

Saudi Arabia's National Volunteer Work Platform could potentially provide part of the answer. It already offers mechanisms for organizing volunteer opportunities, documenting hours, and issuing certificates.

Saudi Vision 2030 reports show a substantial expansion in volunteer work: by the 2025 report, the national platform had recorded more than 189 million cumulative volunteer hours.

Imagine one of the approved opportunities becoming “help document Saudi dialects,” “review historical Arabic texts,” “annotate educational documents,” or “create voice data to improve the understanding of Arabic.”

A volunteer would enter the portal, complete tasks suited to their level, and have the contributions credited after quality review.

Eligible hours would be calculated according to published standards for time and professional effort, not merely by the number of clicks or words.

Through a potential institutional integration with the relevant authorities, those hours and certificates could eventually appear in government digital services, including the Tawakkalna application, if such a link were formally approved.

There would be no need to create a parallel system for everything that already exists.

The more intelligent course would be for the proposed initiative to connect with the national volunteer and education infrastructure, subject to the approval and participation of the relevant ministries and authorities, and add a new field: **digital knowledge volunteering**.

But we must understand a sensitive aspect of the incentive system.

If people are rewarded only for the number of words they produce, we may receive a great many words of little value.

If they are rewarded for the number of labels they apply, some may answer quickly and without care.

This recalls Goodhart’s law in economics and measurement: when a metric becomes the target of a reward, it may lose its ability to represent the quality it was intended to measure.

Rewards should therefore depend on the quality and difficulty of a contribution and the outcome of its review.

A volunteer might earn experience points (XP), advance through knowledge levels, unlock more specialized tasks, receive skill badges and accredited certificates, and gain training opportunities in university laboratories.

Universities, schools, and regions might compete, not over the raw number of contributions, but over the quality, diversity, and research impact of the resources they produce.

Incentives could include research awards, training opportunities, grants for student projects, and recognition for outstanding contributors.

Work that requires advanced expertise or sustained professional effort should have paid pathways, as in some Korean programs, rather than relying indefinitely on volunteers.

Volunteer work is not an excuse to obtain specialized labor for free.

An initiative that respects its contributors will be better able to preserve its quality and their trust over the long term.

The School That Becomes a Knowledge Laboratory

I can imagine schools and universities at the heart of this initiative.

That must not mean forcing students to produce data or automatically turning school assignments into training sources.

It could instead mean carefully designed, optional educational pathways that make participation a way to understand language, research, and artificial intelligence.

At the intermediate-school level, a student might learn how to describe an image, distinguish fact from interpretation, and understand why machines do not always agree when they read texts.

At secondary school, a student could take part in a project documenting local expressions or scientific vocabulary and learn how evaluation questions are designed and results reviewed.

A university student could move on to more difficult tasks: reviewing specialized data, studying inter-annotator agreement, detecting dataset bias, and analyzing model errors.

This work could develop into graduation projects, research papers, and start-ups.

Teachers and faculty members would need to play an essential part in designing and evaluating tasks, rather than merely overseeing their administration.

Children's participation would require special safeguards: parental consent where required, minimization of personal data, a preference for nonsensitive educational tasks, and no automatic use of a child's voice, image, or personal writing for open training.

Saudi Arabia's Personal Data Protection Law (PDPL), together with policies governing children's data, makes the importance of lawful purpose, appropriate consent, data minimization, and protection clear.

If the initiative depends on public contribution, respect for the right not to participate is part of its legitimacy.

No student should lose a grade or an educational opportunity because they chose not to contribute their data.

We want children to learn how knowledge is made, not to become raw material gathered without understanding where it goes.

Who Owns What People Donate?

This is, in my view, the most difficult and important question facing the proposed initiative.

If a million people contribute texts, voices, images, and explanations, and a company then uses them to develop a commercial model, what returns to society?

Does the data become the exclusive property of one organization? Can an independent researcher access it? Can contributors withdraw personal data that they no longer want processed, within the appropriate legal rights and limits?

Can rare cultural material be turned into a commercial product without respecting the community that provided it?

The reassuring phrase “national data” is not enough.

Being national does not cancel the rights of individuals and communities in their data.

I believe the initiative needs a governance model close to the concept of a community data trust, in which a responsible body manages resources under a clear mandate and under legal, scientific, and public oversight.

Every dataset would have a statement defining its purpose, its data-licensing scope, contributors’ rights, and its level of access.

Some datasets could be published as open data, particularly when they are free of personal information and the relevant rights have been preserved.

Others could be made available to researchers under licenses and controls.

Still others could remain inside secure research environments that prevent raw data from being copied and contributors’ identities from being viewed.

It would be useful to draw on the CARE Principles for Indigenous Data Governance, developed by international networks concerned with Indigenous data governance and centered on collective benefit, authority to control, responsibility, and ethics.

I do not mean that these principles should be transferred literally into the Saudi context. I mean that we can learn from their general philosophy: cultural knowledge should not be extracted from a community and its benefits then severed from that community.

A woman might donate a recording in which she explains how a traditional craft is made. That does not necessarily mean she intends to give a company the right to use her voice to create a model that imitates her vocal identity.

Someone might allow a local story to be documented without permitting the disclosure of relatives' names or details of their lives.

The platform should therefore separate consent to collect material, consent to use it in research, consent to use it for training models, and consent to make it available commercially.

Consent is not a small word placed at the bottom of a screen.

It is the ethical relationship that will determine whether society trusts the initiative and continues to sustain it with its knowledge.

Data gathered with people's trust may be more valuable than a larger volume gathered without their understanding of where it will go.

Why This Project Could Matter More Than a Single Model

Consider what the initiative could produce over time.

A substantial resource of Saudi voices and dialects that could support speech-recognition technologies.

Collections of original, documented Arabic texts to support language modeling.

Annotated images and documents for document understanding and computer vision.

Professional terminology databases from medicine, engineering, education, and industry.

Blind evaluation sets that universities could use to test how well models understand society.

These are different resources. They should not all be poured into one container without distinction.

Some data is suitable for pretraining, some for fine-tuning, and some only for evaluation.

Evaluation data must be carefully isolated so it does not leak into training. Otherwise, the model becomes like a student who knew the examination questions before entering the room.

If these resources are managed as shared scientific infrastructure, however, a broad ecosystem of research, companies, and products could grow on top of them.

Potential beneficiaries could include existing national models, research laboratories, deep-technology companies, universities, and start-ups.

Some of the datasets could also become valuable resources outside the Kingdom for training or testing multilingual systems or developing specialized technologies.

The achievement would then be more than possessing the largest number of Arabic tokens.

It would be possessing data whose sources, diversity, limits, and contribution to model capability are known.

Bigger is not always better, and rarer does not always mean smaller in number.

One thousand hours of documented natural speech from an environment absent from global data may be immensely valuable, while ten million similar sentences may be little more than an additional burden on training.

A New Kind of Capital Called Society

I believe one of the most important economic shifts in artificial intelligence will be a move from focusing on the volume of data alone to examining its authenticity, origin, usage rights, and knowledge value.

As it becomes easier and cheaper for models to generate text, it becomes harder to tell whether some new data represents a human experience that models have not learned before, or merely recycles the probabilities of the model itself.

The relative value of original human data may therefore rise, especially in underrepresented fields, languages, and experiences.

This is not because humans always produce more accurate data than machines. It is because they live in the world the machine is trying to represent.

A physician knows how patients describe their symptoms in practice. A teacher knows the mistakes students repeatedly make. An engineer knows the working vocabulary that has arisen inside a factory. A farmer knows the language of the land that appears in no technical

manual.

This experience is distributed among people, and no central laboratory can gather it alone.

That brings us back to the Austrian economist Friedrich Hayek's idea of dispersed knowledge, explored in his famous 1945 essay *The Use of Knowledge in Society*.

Hayek observed that a great deal of practical knowledge is distributed among members of society and connected to local conditions that no central body can fully comprehend.

We may not agree with all his economic conclusions, but that observation seems especially relevant in the age of artificial intelligence.

No data center, by itself, knows what an entire society knows.

If the proposed initiative can build methods for contribution, review, and preservation, it will have created a channel for turning dispersed knowledge into a resource that scientists and engineers can use.

That is a new economic asset.

It is more than a vast collection of files. It is **an infrastructure for the continuous production of knowledge, sourced from society and returning some of its benefit to society.**

A Beginning That Does Not Have to Wait for the Future

I do not imagine the initiative beginning by asking for a million contributions on its first day.

It would be better to begin with defined projects whose effects researchers can actually measure.

Initial tracks could focus on dialects, natural speech, Arabic documents, cultural images, and professional terminology, with the potential participation of universities, schools, nonprofit organizations, and research laboratories.

The data could then be divided into training and test sets, and model results compared before and after the human contributions are added.

Did recognition of a particular dialect improve? Did table-extraction errors decline? Did the model become more capable of understanding professional terminology? Did greater diversity among contributors reduce bias?

If clear positive effects appeared, the initiative could expand to new regions and fields.

If the results did not improve, we should understand why before gathering more data.

The purpose is not to collect text whose sheer volume makes us proud.

The purpose is to gather useful knowledge.

Ambitious goals could be set over several years: attracting one million voluntary contributors, building annotated human datasets in multiple fields, and creating Saudi evaluation benchmarks recognized in research.

But success should be measured by data quality, diversity, impact, and contributors' rights, not by participant numbers alone.

Perhaps the most beautiful measure of success would come five years later, when a Saudi university discovered that a dataset created with contributions from students, teachers, and volunteers across different regions had become a reference used by international laboratories.

At that moment, society would have helped produce knowledge that traveled beyond the place where it began.

When People Donate What a Machine Cannot Invent

I return to Finland.

The initiative did not ask an ordinary citizen to become a computer scientist or understand the mathematical equations behind speech-recognition models.

It asked for something simpler: speak.

Behind that request was a deep recognition that some of the most important resources of the future already exist within society, but have not yet found a route that turns them into learnable knowledge.

I believe Saudi Arabia has an opportunity to shape a broader version of this idea.

We have extensive public and higher education, a developing nonprofit sector, national platforms for volunteer work, research institutions, technology companies, and a society rich in geographic, social, professional, and linguistic diversity.

The question is not whether we can persuade citizens to produce millions of sentences.

It is whether we can build a project that makes every small contribution part of a large scientific undertaking and preserves each contributor's right to know where their contribution went and how it was used.

Imagine a child describing a picture of a tool they have seen only in their grandfather's home.

The grandfather then explains its name and its old use.

A teacher reviews the text, a linguist corrects some of its wording, and a specialist connects it to documented historical material.

A few minutes of interaction among four people have become knowledge that may not exist in this form anywhere on the internet.

That knowledge could become an example in a dataset, part of an evaluation benchmark, an entry in a digital dictionary, or material for developing a model that understands the history and culture of objects.

This is the vision that I believe deserves the proposed name “National Saudi Diwan of Human Knowledge.”

It would not be a vast repository into which we throw whatever people donate.

It would be an institution that listens well, teaches machines through people, and gives those people skill, recognition, and a fair share of the benefit created by the knowledge they helped produce.

The mother who explained a local expression would no longer be merely a minor contributor to a database.

The student who corrected a word in a manuscript would no longer be merely a volunteer who spent five minutes in front of a screen.

Each would have helped, in their own way, to build something larger.

The future in which we want machines to learn our culture may depend on words no one has yet spoken into a microphone, documents no computer has yet read, and knowledge carried by a woman or man who does not know that what they possess may one day become a precious scientific resource.

My hope is not that we ask society to speak more simply to fill a machine’s memory.

It is that we build a system that listens to what society actually knows.

Nations do not give artificial intelligence their language merely by translating into it what others already know. They give it something of themselves when their sons and daughters become partners in writing the knowledge from which it will learn.

Perhaps the greatest Saudi artificial intelligence project will begin not in a room full of servers, but in an ordinary home, when a grandmother says to her granddaughter, “Come, let me teach you a word they do not know.”

The granddaughter learns.

Then the machine learns.

And the human being remains the first teacher.

References and primary sources

1. [Donate Speech \(Lahjoita puhetta\)](#) — The Language Bank of Finland (Kielipankki).
2. [Lahjoita puhetta: a large-scale corpus of spoken Finnish with some benchmarks](#) — Language Resources and Evaluation / Springer Nature.
3. [Over 10,000 contributors onboarded on Bhashini Samudaye](#) — Press Information Bureau, Government of India / Ministry of Electronics & IT.
4. [AI Data](#) — National Information Society Agency, Republic of Korea.
5. [AI-Hub data catalog](#) — AI-Hub / National Information Society Agency, Republic of Korea.
6. [Data Dam project begins, being key to Digital New Deal](#) — Ministry of Science and ICT, Republic of Korea.
7. [Data labeler established as a new future-oriented job](#) — National Information Society Agency, Republic of Korea.
8. [Introducing Aquarium: An Open Data Platform for Southeast Asian Languages](#) — AI Singapore / SEA-LION.
9. [We're expanding our presence in Singapore to advance AI in the Asia-Pacific region](#) — Google DeepMind.
10. [TI-Hüpe jõuab sügisest kutseharidusse ja põhikooliõpetajateni](#) — Estonian Ministry of Education and Research.
11. [AI Leap](#) — e-Estonia / Estonian Government.
12. [By the People: LC Crowdsourcing](#) — Library of Congress.
13. [Achievements \(Points\)](#) — Google Crowdsourcing Help.
14. [How does Google use my answers?](#) — Google Crowdsourcing Help.
15. [Common Voice](#) — Mozilla Common Voice.
16. [GPT detectors are biased against non-native English writers](#) — Patterns / Cell Press.
17. [Testing of detection tools for AI-generated text](#) — International Journal for Educational Integrity / Springer Nature.

18. [Datasheets for Datasets](#) — arXiv (preprint).
19. [Datasheets for Datasets](#) — Communications of the ACM (published article).
20. [Investigating Reasons for Disagreement in Natural Language Inference](#) — Transactions of the Association for Computational Linguistics / MIT Press.
21. [Saudi Vision 2030 Annual Report 2025](#) — Saudi Vision 2030.
22. [Ministry platforms — Volunteer Platform](#) — Saudi Ministry of Human Resources and Social Development.
23. [Personal Data Protection Law](#) — Saudi Data & AI Authority.
24. [Implementing Regulations of the Personal Data Protection Law](#) — Saudi Data & AI Authority / National Data Management Office.
25. [Children and Incompetents Data Protection Policy](#) — Saudi Data & AI Authority / National Data Management Office.
26. [The CARE Principles for Indigenous Data Governance](#) — Global Indigenous Data Alliance.
27. [The Use of Knowledge in Society](#) — American Economic Review / American Economic Association.
28. [Conference Volumes — Papers in Monetary Economics](#) — Reserve Bank of Australia.