



观点与分析

编辑日期：2026年10月27日

互联网上没有的那些词

从芬兰到沙特：为什么最宏大的国家人工智能项目，或许会从一位祖母的声音、一本学生笔记和整个社会的记忆开始？

2020 年夏天，芬兰向国民提出了一个不同寻常的请求：请开口说话。

人们不必朗读正式文本，不必背诵实验室精心编写的句子，也不必使用学校里教授的标准芬兰语。

他们只需像平日生活中那样说话。

聊体育、食物和回忆，聊自己喜欢什么、对什么意见不一。带着自然的方言、迟疑和偶尔的口误，也带着交谈中的沉默、笑声，以及词典从未收录的表达。

这项计划名为“捐出你的话语”（Lahjoita puhetta, Donate Speech），由芬兰公共广播公司 Yle 联合芬兰语言银行 Kielipankki，以及研究与技术合作伙伴发起。

它原计划收集一万小时的自然芬兰语口语，用来开发真正听得懂人们日常说话方式的技术。

这场活动于 2024 年 3 月 6 日结束时，芬兰语言银行宣布已收集三千多小时的口语。2022 年发表的研究论文则称，2021 年版语料库约有 3600 小时、估算涉及两万多名说话者；人数依据应用客户端身份推算，并非经核实的个人身份。

这项计划没有达到最初设定的数量目标，却创造了比一批录音更重要的东西。

它让普通人成为未来技术所需知识基础设施的共同建设者。

论文 *Lahjoita puhetta: A Large-Scale Corpus of Spoken Finnish* 记录了日常口语如何成为自动语音识别 (Automatic Speech Recognition, ASR)、方言理解以及语言与社会特征分析的研究资源。

更打动我的是，芬兰人明白了一个再简单不过的道理：工程师再有才华，也无法坐在实验室里凭空发明数百万人真实说话的方式。

工程师可以设计向人学习的系统，但学习材料必须由人提供。

这个芬兰故事把我带到一个问题。我认为，它值得催生一项前所未有的沙特国家倡议：

如果沙特的每个家庭、每所学校和大学、每种职业和机构，都能参与建设一套规模宏大、真正代表我们的语言、文化与知识的人类数据基础设施，会发生什么？

它不靠收集互联网上别人如何书写我们，也不靠生成数百万段与现有模型输出相似的文本。

它从人本身出发。

从人们说出的词、理解的含义、知道如何阅读的文献出发，也从机器尚未学会的细节出发——因为那些细节从未写进机器能够触及的资料中。

互联网走过半个世纪，生活依然比它更辽阔

想象一下，你请语言模型 (Language Model) 解释阿拉伯语表达“يا بعد حيي” (ya ba‘d hayyi) 的含义。

它或许能给出尚可接受的词典释义，也许会说，这句话表达亲爱与珍重。

但知道一句话的含义是一回事，懂得它在人际关系中的位置又是另一回事。

老人可能这样称呼孙辈，女人可能这样称呼姐妹，朋友也可能在表达感激时对朋友说。语气和含义会随情境、地点以及说话者之间的关系而变化。

再想象一位工业厂站的工程师：他用同事们熟知的俗称描述一种反复出现的故障，而公司手册里根本没有这个叫法。

或者一位贝都因女性，她能分辨风势的细微变化和云层去向，而通用百科只能给出大致描述。

又或者一名护士，她熟悉患者描述疼痛时常用的说法，也知道这些民间表达与医生写进临床记录的词语有何不同。

这些知识没有出现，并不是因为它们毫无价值。

很多时候，它们之所以缺席，是因为从未进入机器赖以学习的数字系统。

由此浮现的问题，比阿拉伯语文本不足更深一层。

统计学中有一个问题叫抽样偏差（Sampling Bias）。如果数据来自无法代表整个社会的群体，模型或许会极其擅长理解这部分人，却不善于理解样本没有覆盖的人。

到了语言模型，这个问题更加复杂。

互联网并不是社会的一份均衡副本。

有些人比其他人写得更多；有些职业产生大量数字文本；有些群体只留下零星痕迹；有些方言只存在于私人交谈中，未经同意绝不应收集。

有时，掌握着最珍贵地方知识的人，恰恰在数字数据中最少出现。

因此，模型可能广知天下，却对使用者身边的真实人生所知甚少。

它知道的是已经发表的内容，未必是人们已经拥有的知识。

机会正在这里：沙特更需要的，不只是继续收集机器已经知道的东西，而是让人们把机器过去无从接触的知识教给它。

当民众成为训练机器的伙伴

芬兰并非唯一的例外。

在拥有众多语言和方言的印度，政府推出了 BHASHINI 计划。其中的语言捐赠项目 Bhasha Daan，允许国民贡献文本、录音和译文。

2026 年 3 月，印度电子和信息技术部宣布，已有一万多名贡献者加入 Bhashini Samudaye。这个平台汇聚个人、大学、企业和实验室，共同建设语言资源、开展数据标注（Data Annotation）。

它的用意不只是由政府收集国民的声音，再把声音放进档案库。

更重要的是建设一项可由大学、研究人员、机构和企业共享，并能服务于多种技术的资源。

这或许正是短期数据收集活动与永久性国家基础设施之间的区别。

在韩国，这类项目采取了更广泛的制度形式。

韩国国家信息社会振兴院（National Information Society Agency, NIA）建立了 AI-Hub。该机构称，自 2017 年以来，平台已提供九百多个用于训练人工智能（Artificial Intelligence, AI）的数据集（Datasets）。

数据涵盖韩语、声音、图像、医疗、工业、农业等领域。

值得注意的是，韩国没有把数据标注当成封闭的技术活动。

在建设并通过平台提供训练数据的政府项目中，数据收集、处理和质量验证既创造了直接就业，也引入了有偿众包。该机构还记录了部分从业者随经验增长，从标注转向审核和项目管理的可能路径。

这里显现出一种重要的经济视角：教机器学习，也可以成为培训人、让人获得新职业技能的途径。

新加坡选择了另一条同样有意思的道路。

AI Singapore 与 Google 合作，在 SEA-LION 体系内推出 Aquarium 平台，让社区和研究人员更容易为东南亚语言数据集作出贡献。

项目不只是为了收集马来语、泰语或印度尼西亚语，还希望让更多人接触能够反映当地语言与文化现实的资源，并开发参与、评估和激励工具。

“Aquarium（水族箱）”这个名字本身也藏着一个优美的比喻：语言不是放进缸里的一条大鱼，而是由多样性与关系组成的活系统。

爱沙尼亚则从教育切入。

它推出 AI Leap，把人工智能引入公共教育，帮助教师和学生培养批判性思考与学习能力。

2025—2026 学年，全国 156 所高中中有 154 所参与该项目，覆盖约五千名教师和两万名学生。

严格来说，这不是我在这里设想的国家训练数据收集项目。

但它证明，国家可以让学校成为一张广泛的全国网络，用于学习并参与新技术实践的形成。

到了美国，我们会在一个意想不到的地方看到精彩案例：美国国会图书馆（Library of Congress）。

图书馆发起了 *By the People* 项目，邀请志愿者转录手写稿和历史文献、审核他人的转录结果，并为资料加上标签。

志愿者可以是学生、退休人士，也可以是业余历史爱好者。

他们每读出一行，就把一份计算机难以理解的历史文献，变成可以检索、研究和利用的文本。

对人工智能而言，这类工作体现了把非结构化档案（Unstructured Archives）转化为可用数字知识的价值。

五种实践、五条不同路径，最终汇聚到同一个认识：

人工智能需要的数据，并不总是一件可以买到的现成商品。有时，它是散落在数百万人身上的知识；真正的创新，是建立一套办法，把知识汇集起来，维护贡献者的权利，并验证其质量。

“迪万”：由人民共同书写的国家项目

我设想一项沙特倡议，暂拟名为“沙特人类知识迪万”。

在我的设想中，它不是一款要求人们下载后便被遗忘的应用，也不是颁完奖就结束的季节性比赛。

它是一套永久性的国家参与式数据基础设施（National Participatory Data Infrastructure），拟邀请普通教育与高等教育体系、私营部门、非营利组织、研究中心，以及自愿贡献的公民和居民参与。

它的目标，是建设规模最大、有据可查的沙特原创标注人类数据资源，并审慎体现语言、文化、职业与社会多样性。

拟议中的迪万以一座数字门户为中心，提供不同层级的参与方式。

学生登录后，可以找到适合自己年龄和知识的任务：解释一个本地表达、描写一个场景、朗读一句话，或审核从获得授权的文献中自动提取的文本。

教师可以协助分类教育文本、判断难度，并解释其中容易混淆的地方。

医生或护士可以审核健康术语和表达，所用材料应是专为训练编写的示例，或依照实施条例完成匿名化并评估过重新识别风险的教育数据，绝不能是暴露身份的患者记录。

工程师可以解释专业词汇、标注依法可以使用的设备图像，或描述一项通用工业流程的步骤。

一位关注社会史的女性，可以在同意之后录下自己对谚语、地方习俗或老式家用器具名称的解释。

一位来自贝都因地区的男性可以描述有关地点、天气和野生动物的词汇；城市居民也可以贡献现代都市生活中形成的表达。

任何方言都不应预先高于另一种，任何社会语言风格也不应被视为天然优越。

目的不是把单一的“沙特人形象”强加给模型，而是让模型看见它所学习的这个社会有多么多样。

居民也可以贡献自己的语言与经验，尤其是日常交往、工作、服务以及语言互译方面的知识。

这将为开发更广泛理解沙特社会与职业情境的模型打开可能。

它最大的优点，是贡献者无需懂编程或机器学习（Machine Learning）。

只要他确实知道一些可以传给别人的东西，就足够了。

祖母也许不知道什么是 Transformer 架构（Transformer Architecture），但她知道的故事和表达，再强大的算力也无法让一位 Transformer 工程师凭空发明。

到这里，这项倡议就不再只是一种数据来源。

它还意味着一种社会承认：知识并非学位持有者的专属，日常经验本身也可以成为技术发展的珍贵材料。

当纠正一个词成为国家行动

不妨想象一下迪万的界面。

志愿者面前出现一份旧阿拉伯语文献的图像，权利方已经授权使用。

光学字符识别（Optical Character Recognition, OCR）工具尝试提取文字，却因笔迹特点或纸张破损而读错一个词。

志愿者看出问题，改正了它。

这项修改会交给第二名贡献者审核。两人的读法若不一致，则转交经验更丰富的审核者。

短短几分钟，他们完成的工作就可能为改进阿拉伯语文献阅读系统的数据集作出贡献。

另一个任务会展示一件传统手工艺品或农具的授权图像。

门户请参与者描述它，说出自己所在地区对它的称呼和它的用途，或许还可以补充另一种方言中的别名。

图像再与恰当的描述关联起来，成为视觉语言学习（Vision-Language Learning）的训练材料。

在第三项任务中，志愿者听取一段由本人同意捐出的录音，再审核自动语音识别系统生成的文本。

从技术角度看，这并非空想。

[Google Crowdsourcing](#) 允许用户参与图像标签验证（Image Label Verification）、手写识别（Handwriting Recognition）、译文审核、图像描述、录音验证等任务。

平台会根据任务类型和难度给予不同积分；对于经其他参与者审核确认的部分贡献，还会增加积分。

Mozilla Foundation 的 Common Voice 项目允许人们录下自己的声音，再由其他志愿者审核录音，从而形成可用于语音技术开发的多语言声音数据集。

这条思路已经证明可以实行。

但沙特的拟议倡议可以超越对这些项目的简单复制。

它可以把贡献与知识领域相连，设置进阶路径，发展语言、声音和视觉标注方面的能力证书，并让参与成为走向数据科学专业化的一扇门。

一个最初只在手稿中改正了一个词的人，几年后或许会成为研究人员，开发出能够阅读成千上万份手稿的系统。

机器能闻到机器的气味吗？

这里有一个绕不开的问题。

如果向广泛公众开放参与，一些贡献者可能会使用 GPT、Claude 或 DeepSeek 等模型生成现成文本，再把它作为原创的人类贡献提交。

这绝不是边缘风险。

如果目标是建设源自人类的知识资源，结果资源中却充满商业模型生成的文本，我们可能只是把早已存在于旧模型里的知识重新灌入其中，并没有扩大人类经验的代表范围。

因此，我建议在倡议门户中加入一套专门研究贡献来源与质量的系统。

系统可以结合 N 元语法分析（N-gram Analysis）、文体统计指纹（Stylometric Fingerprints）、语言困惑度（Perplexity），以及近重复检测（Near-Duplicate Detection）。

例如，它可以寻找某些句式异常反复的现象、与已发表材料完全一致的长文本，或在机器输出中远多于人类参照数据集的模式。

但这里存在一个不能忽视的科学悖论。

一段文字不像合成文本，并不能证明它出自人类；同样，一个人的文风规整，也不能成为指控他使用机器的理由。

斯坦福大学的 Weixin Liang 等人在 2023 年发表于《Patterns》的研究发现，在他们测试的样本和七种检测器范围内，生成文本检测器可能对英语非母语者的写作产生偏差，把人写的文字判作机器生成。

Debora Weber-Wulff 领衔、发表于《International Journal for Educational Integrity》的研究也指出，多种常用检测工具的可靠性有限，在一些情况下会产生较高的假阳性和假阴性错误。

因此，拟议的内容筛查系统不应仅凭一个概率分数，就像法官一样拒绝某项贡献。

它应当只是更广泛的数据来源证明架构（Data Provenance）的一部分。

这套架构从贡献本身开始：清楚声明内容如何产生；可以选择保留有据可查的编辑轨迹；任务直接在门户内完成；由人独立审核；必要时随机抽样核验；全过程不得对志愿者进行侵入式监控。

应把人类原创数据（Human-Authored Data）、经人类核验的人工智能输出（Human-Verified AI Outputs）和合成数据（Synthetic Data）分开。

三类数据都可能有用，但它们并不是同一种东西，彼此的差异不应在同一个存储库里消失。

我还建议为每个数据集建立来源记录，说明它如何收集、基于什么获得许可、接受了何种审核，以及对其质量可以有多大信心。

这与 Timnit Gebru 等人在知名论文 [*Datasheets for Datasets*](#) 中提出的思路一致：记录数据集，就像工业领域记录产品与部件的特性一样。

目标不是制造一款工具，炫耀自己能闻出“人工智能的气味”。

目标是建设一种知识资源，让我们知道其中每个部分的来历、能从中学到什么，以及对它的信任边界在哪里。

气味只是一种印象。

有据可查的来源才是知识。

材料类型	来源	应记录什么？
人类原创数据（Human-Authored Data）	由人创作	来源、创作过程与权利
经人类验证的人工智能输出（Human-Verified AI Outputs）	机器输出后经人工审查	生成工具、审查者的作用与审查边界
合成数据（Synthetic Data）	机器生成	生成方法、目的与使用限制

对本节区别的概念性归纳；任何一类都不会自动保证准确或适用。

当人们意见不同，并不意味着数据有错

迪万最重要的原则之一，应是不把参与者之间的所有差异都当成必须抹去的错误。

两个人看同一幅图，可能会用不同方式描述。

来自两个地区的说话者，可能会对某个表达的含义意见不一，因为同一个词在各自环境中不同用法。

两名研究人员也可能阅读一句含义不唯一的话，各自得出合理解释。

许多传统数据标注系统试图得出唯一的最终真值标签（Ground Truth Label），可能依据多数票，也可能由专业审核者决定。

对于有明确客观答案的问题，这样做很合适，例如读取完好文献中的一个数字，或判断图像中是否存在某个特定物体。

但面对语言和文化，它未必同样合适。

分歧并不总能说明其中一方有错。

Nan-Jiang Jiang 与 Marie-Catherine de Marneffe 在 2022 年发表的论文 *Investigating Reasons for Disagreement in Natural Language Inference* 中，研究了审核者为何会对文本之间的关系产生不同解释。他们发现，部分分歧源于含义本身确实存在多种可能，而不只是标注时的随机错误。

这一点对于希望代表沙特社会的项目至关重要。

不同地区的人对某个词的用法存在差异，或许恰好就是机器需要的信息。

不同世代对某个表达的理解不同，也可能显示出值得记录的语言变迁。

因此，当多种答案确有意义时，标注系统应能够把它们保留下来，并与恰当的地域、年龄或职业情境相连，同时避免把不必要的个人数据变成公开记录。

这并不等于假定所有领域里的每种意见都同样正确。

医学、科学和法律问题需要专业核验；描述语言用法或社会传统时，多元答案则可能是合理的。

有些优质数据并不教给模型一个唯一答案，而是教它理解人们为什么会有不同答案。

这是从收集文本转向收集标注人类知识后，最宝贵的潜在收获之一。

一小时志愿服务，或许会成为知识资产

接下来这个要素，在我看来决定着倡议能否成功：怎样让参与成为一种社会习惯，而不是随着宣传活动结束就消退的一时热情？

沙特国家志愿工作平台或许可以提供部分答案。它已经具备组织志愿机会、记录服务时数和颁发证书的机制。

沙特“2030 愿景”的报告显示，志愿服务大幅增长；截至 2025 年度报告，该平台累计记录的志愿服务时数已超过 1.89 亿小时。

想象一下，获批的志愿机会中出现了“参与记录沙特方言”“审核阿拉伯语历史文本”“标注教育文献”，或“创建语音数据以提升阿拉伯语理解能力”。

志愿者进入门户，完成适合自己能力水平的任务；贡献通过质量审核后，才获得认可。

可计入的服务时数应依据公开的时间与专业标准核算，而不是只看点击次数或字数。

如果未来与主管机构达成制度性整合，并正式批准相关连接，这些时数和证书就可能显示在政府数字服务中，包括 Tawakkalna 应用。

没有必要为已经存在的一切另建一套平行系统。

更明智的做法，是让拟议倡议在有关部委和机构批准并参与的前提下，与国家志愿服务和教育基础设施连接，再增添一个新领域：数字知识志愿服务（Digital Knowledge Volunteering）。

不过，激励机制中有一个敏感面向必须看清。

如果只按字数奖励，人们可能会贡献大量价值不高的文字。

如果只按标签数量奖励，有些人可能为了赶速度而草率作答。

这让人想起经济学与测量领域的古德哈特定律（Goodhart's Law）：一项指标一旦成为奖励目标，就可能失去衡量原本质量的能力。

因此，奖励应取决于贡献的质量、难度和审核结果。

志愿者可以获得经验值（Experience Points, XP），逐步提升知识等级，解锁更专业的任务，取得技能徽章（Skill Badges）和经认可的证书，并获得进入大学实验室培训的机会。

大学、学校和地区之间也可以开展竞赛，比的不是贡献的原始数量，而是所产资源的质量、多样性与研究影响。

激励方式还可以包括研究奖励、培训机会、学生项目资助，以及对杰出贡献者的表彰。

至于需要高水平专业能力或长期稳定投入的工作，应像韩国的部分实践一样设置有偿岗位，不能永远依赖志愿服务。

志愿工作不能成为无偿获取专业劳动的借口。

尊重贡献者的倡议，才更有能力长期维持质量与信任。

当学校成为知识实验室

在我的设想中，学校和大学可以成为倡议的核心。

但这不应靠强迫学生生产数据，也不能把学校作业自动变成训练来源。

它应通过经过审慎设计的自愿教育路径，让参与成为理解语言、研究和人工智能的一种方式。

初中阶段，学生可以学习如何描述图像、区分事实与解释，并理解机器为什么不总能对文本得出相同读法。

高中阶段，学生可以参与记录地方表达或科学词汇的项目，学习如何设计评估问题、如何审核结果。

大学生则可以进入更困难的任务：审核专业数据、研究标注者间一致性（Inter-Annotator Agreement）、发现数据集偏差（Dataset Bias），以及分析模型错误。

这些工作还可能发展为毕业项目、研究和创业项目。

教师与高校教职人员的参与，应体现在任务设计和评估上，而不只是行政监督。

儿童参与必须受到特别保护：在必要时取得监护人同意；尽量少收集个人数据；优先选择不敏感的教育任务；不得自动把儿童的声音、图像或个人写作开放用于训练。

沙特《个人数据保护法》（Personal Data Protection Law，PDPL）及有关儿童数据的政策，清楚表明合法目的、适当同意、数据最小化和保护的重要性。

如果倡议依靠公众贡献，那么尊重人们不参与的权利，就是它正当性的一部分。

学生不应因为不愿贡献个人数据而失去分数或教育机会。

我们希望孩子学习知识如何产生，而不是在不知道数据去向的情况下，自己变成被收集的原料。

人们捐出的东西归谁所有？

在我看来，这是拟议倡议面对的最困难、也最重要的问题。

如果一百万人贡献了文字、声音、图像和解释，之后一家公司用它们开发商业模型，社会能得到什么回报？

数据会成为某个机构的独占财产吗？独立研究人员能否使用？在适用的法律权利与限制范围内，贡献者能否撤回自己不愿继续被处理的个人数据？

稀有文化材料可以在不尊重贡献它的社群时，被转化为商品吗？

一句听来安心的“国家数据”并不足够。

“国家”二字并不会取消个人与社群对自身数据享有的权利。

我认为，这项倡议需要一种接近社区数据信托（Data Trust）的治理模式：由负责任的机构依据明确授权管理资源，并接受法律、科学与社会监督。

每个数据集都应附有说明，明确使用目的、数据许可（Data Licensing）范围、贡献者权利和访问级别。

有些数据集可以作为开放数据（Open Data）发布，尤其是在不含个人信息、相关权利也得到保护的情况下。

另一些可以依据许可与管控向研究人员开放。

还有一些则应留在安全研究环境（Secure Research Environments）中，既不允许复制原始数据，也不允许查看贡献者身份。

这里可以参考原住民数据治理 CARE 原则（CARE Principles for Indigenous Data Governance）。这些原则由关注原住民数据治理的国际网络制定，强调集体利益、控制权、责任与伦理。

我不是要把这些原则原封不动地移植到沙特，而是借鉴其基本理念：文化知识不应从社群中被抽走，知识所产生的利益也不应与社群割裂。

一位女性可能捐出录音，解释某种传统手艺的做法；但这并不必然表示，她愿意让某家公司使用自己的声音，创造模仿其声纹身份的模型。

一个人可能同意记录地方故事，却没有允许公开家人的姓名或生活细节。

因此，平台必须把同意收集材料、同意用于研究、同意用于训练模型，以及同意提供商业使用分别处理。

同意不是放在屏幕底部的一个小词。

它是一种伦理关系，决定社会是否会信任这项倡议，并继续用自己的知识滋养它。

在人们信任之下收集的数据，或许比规模更大、却让贡献者不明白去向的数据更珍贵。

为什么这个项目可能比单个模型更重要？

想想这项倡议经过数年可以产出什么。

它可以形成规模可观的沙特声音与方言资源，供语音识别（Speech Recognition）技术使用。

可以形成一批原创且有据可查的阿拉伯语文本，支持语言建模（Language Modeling）。

可以形成标注图像和文献，用于文档理解（Document Understanding）与计算机视觉（Computer Vision）。

可以建立来自医学、工程、教育和工业领域的专业术语库。

还可以建立盲测评估集（Blind Evaluation Sets），让大学用来检验模型理解社会的能力。

这些资源各不相同，不应不加区分地倒进同一个容器。

有些数据适合预训练（Pretraining），有些适合微调（Fine-Tuning），有些则只能用于评估（Evaluation）。

评估数据必须谨慎隔离，避免泄漏到训练数据中；否则，模型就像进考场前已经看过试题的学生。

但如果把这些资源作为共享科学基础设施管理，其上就可能生长出由研究、企业和产品组成的广泛生态。

现有国家模型、研究实验室、深科技企业、大学和初创企业都可能从中受益。

部分数据集也可能成为沙特境外所需的资源，用于训练或测试多语言系统，或开发专门技术。

届时，成就将不只是拥有最多的阿拉伯语词元（Arabic Tokens）。

更在于拥有这样一批数据：我们知道它们从何而来、有何多样性、边界在哪里，以及能为模型能力增加什么。

更大并不总是更好，更稀有也不一定意味着数量更少。

一千小时来自全球数据未能代表的环境、且来源有据可查的自然口语，可能极其珍贵；一千万个彼此相似的句子，却可能只是训练的额外负担。

一种名为“社会”的新资本

我认为，人工智能领域最重要的经济变化之一，将是注意力从单纯的数据规模，转向数据的真实性、来源、使用权和知识价值。

模型生成文本越容易、成本越低，我们就越难判断，一批新数据究竟代表模型此前没有学过的人类经验，还是只在循环利用模型自身的概率。

因此，原创人类数据的相对价值可能上升，尤其是在代表不足的领域、语言和经验中。

这不是因为人产生的数据永远比机器更准确，而是因为人生活在机器试图表征的世界里。

医生知道患者现实中怎样描述症状；教师知道学生经常犯什么错；工程师知道工厂内部形成的实用术语；农民知道技术手册从未记录的土地词汇。

这些经验分散在人群之中，任何一个中央实验室都无法独自收齐。

这让我们回到奥地利经济学家 Friedrich Hayek 的“分散知识”（Dispersed Knowledge）概念。他在 1945 年的名篇 *The Use of Knowledge in Society* 中讨论了这一思想。

Hayek 指出，大量实用知识分散在社会成员之间，并与任何中央机构都无法完全掌握的地方情境相连。

我们未必赞同他的所有经济结论，但这一观察在人工智能时代显得格外贴切。

没有任何一个数据中心能独自知道整个社会所知道的一切。

如果拟议倡议能够建立贡献、审核与保存机制，它就创造了一条通道，把分散的知识转化为科学家和工程师可以利用的资源。

这是一种新的经济资产。

它不只是一大批文件，而是一套持续生产知识的基础设施：来源是社会，产生的部分利益也回到社会。

不必等到未来才开始

我不认为这项倡议应该在第一天就索求一百万份贡献。

更好的起点，是一批范围明确、让研究人员真正能够衡量成效的项目。

初期路径可以聚焦方言、自然口语、阿拉伯语文献、文化图像和专业术语，并拟邀请大学、学校、非营利机构和研究实验室参与。

随后把数据分成训练集与测试集，比较加入人类贡献前后的模型结果。

某种方言的识别能力提高了吗？表格提取错误减少了吗？模型更能理解专业术语了吗？贡献者更多样后，偏差降低了吗？

如果出现明确的积极效果，倡议可以扩展到新的地区与领域。

如果结果没有改善，就先弄清原因，再继续增加数据收集。

目的不是收集一批让我们因规模而自豪的文字。

目的是收集有用的知识。

项目可以设定跨越数年的宏大目标，例如吸引一百万名自愿贡献者，在多个领域建立标注人类数据集，并创建获得研究界认可的沙特评估基准。

但衡量成功标准，应是数据的质量、多样性、影响和贡献者权利，而不能只有参与人数。

或许最美的成功指标是：五年后，一所沙特大学发现，由不同地区的学生、教师和志愿者共同创建的数据集，已经成为国际实验室使用的参考资源。

那一刻，社会共同生产的知识就越过了它诞生之地的边界。

当人们捐出机器无法发明的东西

让我们回到芬兰。

那项计划没有要求普通人变成计算机科学家，也没有要求他们理解语音识别模型背后的数学方程。

它只提出了一个更简单的请求：说话。

这个请求背后有一种深刻认识：未来最重要的部分资源，本就存在于社会内部，只是还没有找到一条把它们转化为可学习知识的道路。

我认为，沙特有机会塑造这套理念的更广阔版本。

我们拥有广泛的普通教育与高等教育体系、正在发展的非营利部门、国家志愿服务平台、研究机构 and 科技企业，也拥有一个在地域、社会、职业和语言方面都极为多样的社会。

问题不在于，我们能否说服国民生产数百万个句子。

问题在于，我们能否建设一个项目，让每一份微小贡献都成为大型科学工作的一部分，同时维护贡献者了解内容去向与使用方式的权利。

想象一个孩子描述一件工具的图片，他只在祖父家见过它。

接着，祖父解释它的名称和过去的用途。

教师审核文字，语言研究者改正其中的一些用词，再由专业人员把它与有据可查的历史材料联系起来。

四个人只用几分钟互动，就创造出一种未必以这种形式存在于任何网站上的知识。

这份知识可以成为数据集中的样本、评估基准的一部分、数字词典中的一条记录，或用于开发能够理解物件历史与文化的模型。

这就是我认为配得上拟议名称“沙特人类知识迪万”的图景。

它不是一座把人们捐出的东西一股脑倒进去的巨大存储库。

它是一种善于倾听的机构，通过人来教机器，也回馈给人技能、认可，以及他们协助创造的知识所带来的利益中的公平份额。

到那时，那位解释地方表达的母亲，就不再只是数据库里一名微不足道的贡献者。

那位在手稿中改正一个词的学生，也不再只是一个在屏幕前花了五分钟的志愿者。

他们都以自己的方式，参与建设了更宏大的事物。

我们希望机器能够学习我们的文化，而这样的未来，或许取决于尚未有人对着麦克风说出的词、尚未被计算机读过的文献，以及某位女性或男性所拥有的知识——他们还不知道，自己掌握的东西有一天会成为珍贵的科学资源。

我所希望的，不是让社会为了填满机器的记忆而说得更多。

而是建设一个系统，真正倾听这个社会已经知道的东西。

一个国家，并不会仅仅因为把别人已经知道的内容翻译给人工智能，就把自己的语言交给它。只有当儿女们成为共同书写机器学习材料的伙伴时，这个国家才真正把自身的一部分交给了人工智能。

也许，沙特最宏大的人工智能项目不会从一间摆满服务器的房间开始，而会从一个普通家庭开始：祖母对孙女说，“来，我教你一个他们不知道的词。”

孙女学会了。

之后，机器也学会了。

而人，始终是最初的老师。

参考文献与一手资料

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.