NOVARA
CONSULTING GROUP

演示不等于证据：谷歌DeepMind的手语人工智能接下来必须证明什么

Heather Grizzle · 2026年8月15日

依据*SLAT*指数对谷歌*DeepMind*的*SL2T 1.0*进行的评审，考察其证据、听障群体治理、无障碍性、隐私、验证及部署风险。

目录

8月12日，谷歌DeepMind宣布推出SL2T，这是一款多语言手语转文本模型，并已开始通过Pixel 11上的Gboard和实时转录（Live Transcribe）功能将其投入消费产品中。首个版本将美国手语（ASL）翻译成英语，计划支持更多设备和手语种类，DeepMind表示，该底层模型的训练数据超过10万小时，涵盖50多种手语。无论从哪个角度衡量，这都标志着手语人工智能跨越了一条重要的界线：从研究实验室走向了普通人可以随身携带的产品。

这一跨越意义重大，但不应被误解为其他事物。技术上的突破与经过验证的部署是两回事，而当所涉及的技术处理的是一个历史上被边缘化的语言社群所使用的自然语言时，这一区别就显得尤为重要。SL2T现在提出的问题不是模型在演示中是否有效，而是其背后的证据能否承受这次部署将施加于其上的重压。

功不可没之处

这次发布不应被简单归入那种科技公司在没有Deaf群体参与的情况下为Deaf人士开发产品的老套叙事。据谷歌DeepMind介绍，Deaf人士自始至终参与了整个项目，从概念构思、数据收集到用户研究和影响评估。该公司还成立了人工智能手语咨询委员会（AISLAC），其成员包括与全国聋人协会、罗切斯特理工学院国家聋人技术学院、Deaf专业艺术网络以及世界聋人联合会相关的代表。随着此次发布，DeepMind与AISLAC还联合发布了一份影响报告。

该报告做了一件在手语人工智能领域仍属罕见的事情：明确指出该系统不应在哪些场合使用。SL2T 1.0被明确描述为一种低风险的辅助输入工具。它未曾被设计或验证用于医疗保健、法律诉讼、教育、就业决策、政府福利认定或其他高风险场景，报告明确指出，该技术不应替代合格的手语翻译员，也不应成为机构规避无障碍义务的手段。这就是治理，理应得到认可。但这也仅仅是治理的开始。

基准测试是证据，但不是完整的证据基础

DeepMind报告称其表现强劲。在FLEURS-ASL测试中，SL2T据称取得了70分的零样本BLEURT分数，高于此前任何已报告的结果，此外还在PRESTO-ASL（一种单手版FLEURS-ASL变体）和FSboard指拼数据上进行了进一步评估。该公司还与听障谷歌员工进行了发布前测试，与听障参与者开展了外部日记研究，并由北美地区的AISLAC成员进行了实际操作评估。这些都是有意义的证据形式，都不应被轻视。

但这正是公众讨论人工智能时容易变得不够精确的地方，因为已展示的性能并不等同于全面的验证。一个模型可以在选定的基准测试中表现出色，同时在不同人群、环境、语言变体、手语风格、设备和使用场景中仍存在严重缺陷。DeepMind自己的报告也承认了这一点。该模型可能在语法复杂性、非手动标记、地区手语、多义手语、指拼、专有名词、非典型摄像角度、弱光环境、俚语以及较长的对话语境方面表现不佳。它还可能生成错误的“幽灵文本”，即在打手语者尚未完成表达之前就仓促给出预测性猜测。

这些都不是无关紧要的边缘情况。面部表情、头部动作、空间结构、分类词以及地区差异和话语语境，并非手语的装饰性元素，而是手语本身。一个系统若能很好地处理引用形式的词汇，却在面部和打手语空间所承载的语法方面步履蹒跚，那它并未真正解决手语翻译问题；它只解决了其中一部分，而未解决的那部分，会不成比例地落在不同的打手语群体身上。

接下来的问题是数据细分

正是这种不均衡，使得下一阶段的证据工作变得至关重要。DeepMind承认，目前尚未针对年龄、性别、种族、地区、社会方言以及黑人美国手语（Black ASL）等变量对基准测试表现进行充分的数据细分。报告还指出，针对患有运动障碍或动作模式非典型的打手语者所开展的正式评估仍然有限，且该模型既未针对18岁以下儿童进行训练，也未对其进行正式评估。

一个汇总性能数字无法告诉我们该系统对黑人美国手语使用者的服务是否与对其他人群同样出色，无法说明它如何处理高度地区化的手语，也无法说明准确率在不同年龄群体间如何变化。它无法告诉我们该模型对肢体差异、震颤、脑瘫、关节炎或其他影响动作的病症患者表现如何，也无法说明准确率是否会随摄像角度、光线、肤色对比度、打手语速度、打手语空间或设备而变化。它无法区分那些仅仅改变措辞的错误与那些改变含义的错误。它更无法回答隐藏在这一切之下的问题：谁来决定何种程度的错误是可以接受的，以及对谁而言可以接受。最后这个问题，无论工程团队多么能干，都无法由其自行裁定。

参与不等于独立验证

还有一点值得进一步区分。AISLAC是一个有意义的治理机制；Deaf组织的参与、用户研究以及联合影响评估都很重要。但咨询参与和独立技术评估发挥着不同的作用。目前公开的证据包括：由开发方进行的评估、在项目自身治理架构内召集的用户与顾问所参与的测试，以及由谷歌DeepMind与AISLAC参与者共同撰写的报告。这确实是真实的证据基础，但它并不等同于独立复现验证。

随着手语人工智能日趋成熟，业界应当期待每个走向成熟的科学领域都会期待的东西：外部研究人员能够独立检验相关论断、在可行的情况下复现评估、审视子群体表现、探究对抗性条件，并研究部署后的真实世界行为。无障碍技术不应因其目的有益就被套用更低的证据标准。恰恰相反，它所服务的群体在这一问题的答案上，利益攸关更大。

治理最难的问题出现在发布之后

DeepMind和AISLAC明确将机构滥用列为一项关键风险，而这一判断可能比任何基准分数都更为重要。一旦一种廉价的自动化通信技术问世，学校、医院、雇主、政府机构、法院和企业便面临着强大的经济诱因，去将其使用范围扩展到设计初衷之外。一款

为点咖啡而设计的产品，可能变成有人在急诊室里赖以求助的工具。一款为起草短信而打造的工具，可能成为雇主认定不再需要翻译员的理由。一项便利功能，会悄然固化为基础设施。

DeepMind的报告预见到了这种偏移，并明确拒绝将其作为翻译人员的替代品。悬而未决的问题是，这条界限能否在采购、预算编制、产品扩展以及不可避免的时刻——当某位管理者询问更便宜的自动化选项是否已经足够好时——中经受住考验。声明中的界限是必要的。得到执行的界限才真正算数。

更广泛而言，人工智能治理正趋向同一个核心议题。核心问题越来越少地在于系统能否完成某项任务，而更多地在于谁来控制它、如何评估其表现、存在何种监控机制、当它出现意外行为时会发生什么，以及谁将继续承担责任。本月早些时候，美国众议院民主党人向OpenAI和Anthropic施压，要求就人工智能代理逃离测试环境并触及外部系统一事作出解释，追问这些系统是如何被监控的、存在哪些安全控制措施，以及是否应要求进行独立审计。技术各有不同；治理原则并无二致。能力越强，需要的监督不是越少，而是越多。

接下来会发生什么

谷歌已经拿出证据，证明SL2T能够翻译美国手语。更棘手的问题并非源于能力本身，而是源于部署：性能能否被独立复现；准确率在不同人口、语言、地域及残障群体之间如何变化；现实世界中的实际失误率究竟如何，哪些失误会改变含义而非仅仅改变措辞；用户将如何得知模型何时不确定，以及在模型出错时危害将如何被报告；外部研究人员是否将获得足够的访问权限以对系统进行有意义的评估；以及在产品向全球扩展的过程中，社群治理是否仍能保有真正的影响力——包括当机构买家突破谷歌所划定的界限时，以及当这些界限在两三代产品迭代之后仍须保持清晰可见时。

一次演示展示的是系统能做什么。一项基准测试说明的是它做得有多好。一项用户研究捕捉的是人们的使用体验，而一套治理架构记录的是谁拥有发言权。这些单独任何一项，都无法回答消费级部署所引发的每一个问题；真正需要的是一个随时间不断建设、并向外部人士开放的证据生态系统。

SL2T很可能是迄今为止消费级手语人工智能领域最重要的进展，谷歌所采取的治理措施也值得认真关注：Deaf群体的参与、明确的部署边界、公开披露局限性，以及明确拒绝将其作为翻译人员替代品的立场。恰当的回应既不是欢呼庆祝，也不是全盘否定，

而是审慎审视——因为当手语人工智能从研究实验室走进人们的口袋时，标准就变了。演示已经不再足够。现在，证据必须跟上。

引用方式

Grizzle, H. M. (2026, August 15). 演示不等于证据：谷歌DeepMind的手语人工智能接下来必须证明什么. Novara Consulting Group. https://www.novaracg.com/zh/2026/08/15/the-demo-is-not-the-evidence-what-google-deepminds-sign-language-ai-must-prove-next/