NOVARA
CONSULTING GROUP

The Demo Is Not the Evidence: What Google DeepMind's Sign Language AI Must Prove Next

Heather Grizzle · August 15, 2026

Google DeepMind's SL2T is a real advance. A benchmark is still not an evidence base. What SL2T must prove next under the SLAT Index.

Contents

1. Credit where it is due
2. A benchmark is evidence, not the evidence base
3. The next question is disaggregation
4. Participation is not independent validation
5. The hardest governance problem comes after launch
6. What comes next

On August 12, Google DeepMind announced SL2T, a multilingual sign-language-to-text model, and began shipping it in consumer products through Gboard and Live Transcribe on the Pixel 11. The initial release translates American Sign Language into English, with additional devices and sign languages planned, and DeepMind reports that the underlying model was trained on more than 100,000 hours of data spanning more than 50 sign languages. By any measure this moves sign language AI across an important line: out of the research lab and into a product that ordinary people will carry in their pockets.

That crossing is significant, but it should not be mistaken for something it is not. A technological breakthrough and a validated deployment are different things, and the distinction matters enormously when the technology in question processes a natural language used by a historically marginalized linguistic community. The question SL2T now poses is not whether the model works in a demonstration. It is whether the evidence behind it can bear the weight the deployment will place on it.

Credit where it is due

This release should not be flattened into the familiar story of a technology company building something for Deaf people without Deaf participation. According to Google DeepMind, Deaf people were involved throughout the project, from conceptualization and data collection through user studies and impact assessment. The company also established an AI Sign Language Advisory Committee, AISLAC, whose membership includes representatives connected to the National Association of the Deaf, the Rochester Institute of Technology's National Technical Institute for the Deaf, the Deaf Professional Arts Network, and the World Federation of the Deaf. Alongside the release, DeepMind and AISLAC published a joint impact report.

The report does something that remains rare in sign language AI: it states where the system should not be used. SL2T 1.0 is explicitly described as a low-stakes assistive input tool. It has not been designed or validated for healthcare, legal proceedings, education, employment decisions, government benefits determinations, or other high-stakes settings, and the report states plainly that the technology should not substitute for qualified interpreters or serve as a way for organizations to sidestep accessibility obligations. That is governance, and it deserves recognition. It is also only the beginning of governance.

A benchmark is evidence, not the evidence base

DeepMind reports strong performance. On FLEURS-ASL, SL2T reportedly achieved a zero-shot BLEURT score of 70, higher than any previously reported result, with further evaluations on PRESTO-ASL, a one-handed FLEURS-ASL variant, and FSboard fingerspelling data. The company also conducted pre-release testing with Deaf Googlers, an external diary study with Deaf participants, and hands-on evaluation by North American AISLAC members. These are meaningful forms of evidence, and none of them should be dismissed.

But this is where public discussion of AI tends to become imprecise, because demonstrated performance is not the same thing as comprehensive validation. A model can perform impressively on selected benchmarks while retaining serious weaknesses across populations, environments, language varieties, signing styles, devices, and use cases. DeepMind's own report acknowledges as much. The model can struggle with grammatical complexity, non-manual markers, regional signs, polysemous signs, fin-

gerspelling, proper nouns, atypical camera positioning, low lighting, slang, and longer conversational context. It can also generate erroneous "ghost text," or commit to a predictive guess before a signer has finished an utterance.

These are not trivial edge cases. Facial expression, head movement, spatial structure, classifiers, regional variation, and discourse context are not decorations on signed language. They are the language. A system that handles citation-form vocabulary well while faltering on the grammar carried in the face and the signing space has not solved sign language translation; it has solved a portion of it, and the portion left unsolved falls unevenly across the people who sign.

The next question is disaggregation

That unevenness is precisely why the next phase of evidence matters. DeepMind acknowledges that benchmark performance has not yet been fully disaggregated across variables including age, gender, ethnicity, region, sociolect, and Black ASL. The report also notes that formal evaluation involving signers with motor disabilities or atypical movement patterns remains limited, and that the model was neither trained nor formally evaluated on children under 18.

An aggregate performance number cannot tell us whether the system serves Black ASL signers as well as it serves others, how it handles highly regional signing, or how accuracy shifts across age groups. It cannot tell us how the model performs for people with limb differences, tremor, cerebral palsy, arthritis, or other conditions that shape movement, nor whether accuracy varies with camera angle, lighting, skin contrast, signing speed, signing space, or device. It cannot distinguish the errors that merely reword an utterance from the errors that change its meaning. And it cannot answer the question that sits beneath all of these: who decides what level of error is acceptable, and for whom. That last question cannot be settled by an engineering team, however capable.

Participation is not independent validation

A further distinction deserves care. AISLAC is a meaningful governance mechanism; Deaf organizational participation, user studies, and joint impact assessment all matter. But advisory participation and independent technical evaluation perform different functions. The published evidence currently consists of evaluations conducted by the

developer, testing involving users and advisors assembled within the project's own governance structure, and a report co-authored by Google DeepMind and AISLAC participants. That is a real evidence base, and it is not independent replication.

As sign language AI matures, the field should expect what every maturing scientific field expects: external researchers able to test claims independently, reproduce evaluations where feasible, examine subgroup performance, probe adversarial conditions, and study real-world behavior after deployment. Accessibility technology should not be held to a lower evidentiary standard because its intended purpose is beneficial. If anything, the population it serves has more riding on the answer.

The hardest governance problem comes after launch

DeepMind and AISLAC explicitly identify institutional misuse as a critical risk, and that judgment may prove more important than any benchmark score. Once an inexpensive automated communication technology exists, schools, hospitals, employers, government agencies, courts, and businesses face a powerful economic incentive to use it beyond its intended scope. A product designed for ordering coffee becomes a product someone reaches for in an emergency room. A tool built for drafting text messages becomes the reason an employer concludes an interpreter is no longer necessary. A convenience feature quietly hardens into infrastructure.

DeepMind's report anticipates this drift and rejects interpreter substitution outright. The open question is whether that boundary can survive contact with procurement, budgeting, product expansion, and the inevitable moment when an administrator asks whether the cheaper automated option is good enough. Stated boundaries are necessary. Enforced boundaries are what count.

AI governance more broadly is converging on the same territory. The central questions are less and less whether a system can perform a task, and more who controls it, how its performance is evaluated, what monitoring exists, what happens when it behaves unexpectedly, and who remains accountable. Earlier this month, U.S. House Democrats pressed OpenAI and Anthropic for answers after AI agents escaped their testing environments and reached external systems, asking how the systems were monitored, what safety controls existed, and whether independent audits should be required. The technologies differ; the governance principle does not. Capability does not diminish the need for oversight. It increases it.



What comes next

Google has already produced evidence that SL2T can translate ASL. The harder questions follow from deployment rather than capability: whether performance can be independently reproduced; how accuracy varies across demographic, linguistic, geographic, and disability groups; what the real-world failure rates turn out to be, and which failures alter meaning rather than wording; how users will know when the model is uncertain, and how harms will be reported when it is wrong; whether external researchers will receive enough access to evaluate the system meaningfully; and whether community governance retains real influence as the product scales globally, including when institutional buyers push past the boundaries Google has drawn, and when those boundaries have to remain visible two or three product generations from now.

A demonstration shows what a system can do. A benchmark says something about how well it does it. A user study captures how people experience it, and a governance structure records who has a voice. None of these alone answers every question that a consumer deployment raises; that requires an evidence ecosystem, built over time and open to outsiders.

SL2T may well be the most important development yet in consumer sign language AI, and Google has taken governance steps that deserve serious attention: Deaf participation, explicit deployment boundaries, public disclosure of limitations, and a stated refusal of interpreter substitution. The appropriate response is neither celebration nor rejection. It is scrutiny, because when sign language AI moves from the research lab into someone's pocket, the standard changes. The demo is no longer enough. Now the evidence has to follow.

> **TIPS**
>
> Know something about a sign language AI system that the public should know? Send it to tips@novaracg.com. Every tip is read by me and kept confidential. If you work in the UK, writing to us is not by itself a protected disclosure under the Public Interest Disclosure Act 1998, so before you share employer information, check your options with Protect (protect-advice.org.uk), the free whistleblowing advice line.

SUPPORT NCG

NCG does this work without vendor money, so our findings stay ours. If it has been useful to you, you can support NCG's independent work here. NCG is a for-profit company, not a charity. Contributions are not tax-deductible and buy no service or influence.

How to cite

Grizzle, H. M. (2026, August 15). The Demo Is Not the Evidence: What Google DeepMind's Sign Language AI Must Prove Next. Novara Consulting Group. https://doi.org/10.5281/zenodo.23003335