NOVARA
CONSULTING GROUP

This document was translated from English by machine and may contain mistakes. The English original is at https://www.novaracg.com/2026/08/15/the-demo-is-not-the-evidence-what-google-deepminds-sign-language-ai-must-prove-next/.

# La demostración no es la evidencia: qué debe demostrar a continuación la IA de lenguaje de señas de Google DeepMind's

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.*

**Contenidos**

1. El mérito, cuando corresponde
2. Un punto de referencia es evidencia, no la base de evidencia
3. La siguiente pregunta es la desagregación
4. La participación no es validación independiente
5. El problema de gobernanza más difícil llega después del lanzamiento
6. Lo que viene después

El 12 de agosto, Google DeepMind anunció SL2T, un modelo multilingüe de traducción de lengua de señas a texto, y comenzó a distribuirlo en productos de consumo a través de Gboard y Live Transcribe en el Pixel 11. La versión inicial traduce la lengua de señas americana al inglés, con más dispositivos y lenguas de señas previstos, y DeepMind informa que el modelo subyacente se entrenó con más de 100.000 horas de datos que abarcan más de 50 lenguas de señas. Sea cual sea el criterio, esto lleva a la IA de lengua de señas a cruzar una línea importante: salir del laboratorio de investigación y entrar en un producto que la gente común llevará en el bolsillo.

Ese cruce es significativo, pero no debe confundirse con algo que no es. Un avance tecnológico y un despliegue validado son cosas distintas, y la distinción importa enormemente cuando la tecnología en cuestión procesa una lengua natural utilizada por una comunidad lingüística históricamente marginada. La pregunta que plantea ahora

SL2T no es si el modelo funciona en una demostración. Es si la evidencia que lo respalda puede soportar el peso que le impondrá el despliegue.

## El mérito, cuando corresponde

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.

El informe hace algo que sigue siendo poco frecuente en la IA de lengua de señas: indica dónde no debe usarse el sistema. SL2T 1.0 se describe explícitamente como una herramienta de entrada asistiva de bajo riesgo. No ha sido diseñada ni validada para el ámbito sanitario, los procedimientos legales, la educación, las decisiones de empleo, la determinación de prestaciones gubernamentales u otros contextos de alto riesgo, y el informe afirma con claridad que la tecnología no debe sustituir a intérpretes cualificados ni servir como forma de que las organizaciones eludan sus obligaciones de accesibilidad. Eso es gobernanza, y merece reconocimiento. También es solo el comienzo de la gobernanza.

## Un punto de referencia es evidencia, no la base de evidencia

DeepMind informa un rendimiento sólido. En FLEURS-ASL, SL2T habría alcanzado una puntuación BLEURT de cero disparos de 70, superior a cualquier resultado reportado anteriormente, con evaluaciones adicionales en PRESTO-ASL, una variante de FLEURS-ASL para una sola mano, y en datos de deletreo manual de FSboard. La empresa también realizó pruebas previas al lanzamiento con empleados Deaf de Google, un estudio de diario externo con participantes Deaf y una evaluación práctica por parte de miembros de AISLAC de América del Norte. Estas son formas significativas de evidencia, y ninguna de ellas debería desestimarse.

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, fingerspelling, 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.

Estos no son casos límite triviales. La expresión facial, el movimiento de la cabeza, la estructura espacial, los clasificadores, la variación regional y el contexto discursivo no son adornos de la lengua de señas. Son la lengua misma. Un sistema que maneja bien el vocabulario en forma de cita mientras vacila en la gramática que se transmite en el rostro y en el espacio de señar no ha resuelto la traducción de la lengua de señas; ha resuelto una parte de ella, y la parte no resuelta recae de manera desigual sobre las personas que usan señas.

## La siguiente pregunta es la desagregación

Esa desigualdad es precisamente la razón por la que importa la siguiente fase de evidencia. DeepMind reconoce que el rendimiento en los puntos de referencia aún no se ha desagregado por completo según variables como la edad, el género, la etnia, la región, el sociolecto y el ASL Negro. El informe también señala que la evaluación formal con usuarios de señas con discapacidades motoras o patrones de movimiento atípicos sigue siendo limitada, y que el modelo no fue entrenado ni evaluado formalmente con menores de 18 años.

Una cifra de rendimiento agregado no puede decirnos si el sistema sirve a los usuarios de ASL Negro tan bien como a otros, cómo maneja un uso de señas muy regional, o cómo varía la precisión entre grupos de edad. No puede decirnos cómo se desempeña el modelo con personas con diferencias en las extremidades, temblores, parálisis cerebral, artritis u otras condiciones que afectan el movimiento, ni si la precisión varía según el ángulo de la cámara, la iluminación, el contraste de la piel, la velocidad de señar, el espacio de señar o el dispositivo. No puede distinguir los errores que simplemente reformulan un enunciado de los que cambian su significado. Y no puede responder la pregunta que subyace a todas estas: quién decide qué nivel de error es acep-

table, y para quién. Esta última pregunta no puede resolverla un equipo de ingeniería, por capaz que sea.

## La participación no es validación independiente

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.

A medida que madura la IA de lengua de señas, el campo debería esperar lo que espera todo campo científico en maduración: investigadores externos capaces de poner a prueba las afirmaciones de forma independiente, reproducir evaluaciones cuando sea posible, examinar el rendimiento por subgrupos, indagar en condiciones adversas y estudiar el comportamiento real después del despliegue. La tecnología de accesibilidad no debería someterse a un estándar de evidencia más bajo por el hecho de que su propósito sea beneficioso. Si acaso, la población a la que sirve tiene mucho más en juego con la respuesta.

## El problema de gobernanza más difícil llega después del lanzamiento

DeepMind y AISLAC identifican explícitamente el mal uso institucional como un riesgo crítico, y ese juicio podría resultar más importante que cualquier puntuación de referencia. Una vez que existe una tecnología de comunicación automatizada de bajo costo, las escuelas, los hospitales, los empleadores, las agencias gubernamentales, los tribunales y las empresas enfrentan un poderoso incentivo económico para usarla más allá de su alcance previsto. Un producto diseñado para pedir un café se convierte en un producto al que alguien recurre en una sala de emergencias. Una herramienta creada para redactar mensajes de texto se convierte en la razón por la que un empleador concluye que ya no necesita un intérprete. Una función de conveniencia se endurece silenciosamente hasta convertirse en infraestructura.

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.

La gobernanza de la IA, en un sentido más amplio, converge en el mismo territorio. Las preguntas centrales cada vez giran menos en torno a si un sistema puede realizar una tarea, y más en torno a quién lo controla, cómo se evalúa su rendimiento, qué supervisión existe, qué ocurre cuando se comporta de manera inesperada y quién sigue siendo responsable. A principios de este mes, legisladores demócratas de la Cámara de Representantes de EE. UU. presionaron a OpenAI y Anthropic para obtener respuestas después de que agentes de IA escaparan de sus entornos de prueba y alcanzaran sistemas externos, preguntando cómo se supervisaban los sistemas, qué controles de seguridad existían y si deberían exigirse auditorías independientes. Las tecnologías difieren; el principio de gobernanza no. La capacidad no disminuye la necesidad de supervisión. La aumenta.

## Lo que viene después

Google ya ha presentado evidencia de que SL2T puede traducir ASL. Las preguntas más difíciles surgen del despliegue y no de la capacidad: si el rendimiento puede reproducirse de forma independiente; cómo varía la precisión entre grupos demográficos, lingüísticos, geográficos y de discapacidad; cuáles resultan ser las tasas reales de fallo en el mundo real, y qué fallos alteran el significado en lugar de la formulación; cómo sabrán los usuarios cuándo el modelo tiene incertidumbre, y cómo se informarán los daños cuando se equivoque; si los investigadores externos recibirán acceso suficiente para evaluar el sistema de manera significativa; y si la gobernanza comunitaria conserva una influencia real a medida que el producto se expande globalmente, incluso cuando los compradores institucionales sobrepasen los límites que Google ha trazado, y cuando esos límites tengan que seguir siendo visibles dos o tres generaciones de producto más adelante.

Una demostración muestra lo que un sistema puede hacer. Un punto de referencia dice algo sobre qué tan bien lo hace. Un estudio con usuarios capta cómo lo experimenta la gente, y una estructura de gobernanza registra quién tiene voz. Ninguno de estos elementos por sí solo responde a todas las preguntas que plantea un despliegue

de consumo; eso requiere un ecosistema de evidencia, construido con el tiempo y abierto a personas externas.

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.

---

## Cómo citar

Grizzle, H. M. (2026, August 15). La demostración no es la evidencia: qué debe demostrar a continuación la IA de lenguaje de señas de Google DeepMind's. Novara Consulting Group. https://doi.org/10.5281/zenodo.23003335