This document was translated from English by machine and may contain mistakes. The English original is at https://www.novaracg.com/2026/09/11/the-words-put-in-your-mouth-sign-to-text-ai-and-fluent-error/.

Las palabras que te ponen en la boca: la IA de lengua de señas a texto y el error fluido

Heather Grizzle · September 11, 2026

La IA de signos a texto puede inventar oraciones fluidas que una persona Sorda nunca firmó, y luego archivarlas como registro. Por qué las puntuaciones BLEU no pueden ver el fallo que afirman resolver.

Contenidos

1. En qué sentido va la traducción
2. La métrica no puede ver el fallo
3. La confianza es una señal de entrenamiento, no una divulgación
4. Lo que la entrada de pose transmite y lo que no
5. Los parámetros de referencia no son despliegues
6. Qué debería exigir un comprador
7. Cómo la afirmación llegó al inglés
8. Los límites de este análisis
9. Fuentes

Traducción confiada errónea en los sistemas de lengua de signos a texto, y los parámetros de referencia que no pueden verla

On 9 September 2026 the Chinese technology outlet Quantum Bit reported a method from vivo AI Lab called CAPO-SLT, from a paper titled "CAPO-SLT: Confidence-Aware Policy Optimization for Stable Sign Language Translation Generation", posted on OpenReview, where papers sit while under peer review. The problem it names is the one that should worry anyone buying this technology. A translation system can produce a sentence that is fluent, grammatical and entirely plausible while having no sup-

port in what the signer actually did, and because the sentence reads well, nothing about it announces the error. As the reporting puts it, a single word that is locally plausible but lacking visual support can pull an entire translation off course, with the error accumulating across the sentence. The method adjusts the bounds within which reinforcement learning is allowed to update the model, token by token, according to how confident the previous policy was: tighter bounds where confidence is high, near 0.2, and looser ones where it is low, up to 0.3, with additional caps on tokens carrying negative advantage. The aim is to stop the model from locking in confident guesses and to leave room for uncertain ones to be pulled back toward the visual evidence.

Los resultados reportados son más limitados de lo que sugiere el enfoque del titular, y la diferencia importa. En el parámetro de referencia chino CSL-Daily, las ganancias reportadas se miden frente a dos sistemas nombrados que también trabajan a partir de datos de pose: aproximadamente 1,26 BLEU-4 sobre Geo-Sign y 3,07 BLEU-4 sobre Uni-Sign, siendo la afirmación que se trata de las mejores puntuaciones entre los sistemas basados únicamente en pose, y no las mejores del campo en general. En el parámetro de referencia de lengua de signos americana How2Sign, las cifras reportadas son 41,4 BLEU-1, 15,2 BLEU-4 y 34,9 ROUGE-L, ganancias de alrededor de un punto o menos frente a Uni-Sign. El propio artículo no pudo leerse para este análisis, porque el repositorio que lo aloja exige un paso de verificación del navegador, así que lo que sigue examina el problema que el trabajo señala, la dirección de traducción en la que opera y el tipo de evidencia que se ofrece, en lugar de evaluar el sistema. Vale la pena examinar esos aspectos por sí mismos, porque el problema es real, la dirección es la que las instituciones manejan con menos cuidado, y el tipo de evidencia no puede demostrar aquello que se le pide que demuestre.

En qué sentido va la traducción

Most public argument about signed language AI concerns output to Deaf people: an avatar on a departure board, a signing video on a website, a robot in a classroom. Sign-to-text runs the other way. The input is a Deaf person signing and the output is text that everyone else reads, and that changes what a mistake is. When an avatar signs badly, a Deaf person receives a poor translation and can usually tell that something is wrong. When a sign-to-text system fabricates a fluent sentence, the Deaf person's words have been replaced by words they did not say, and those words go to a

clinician, a caseworker, an investigator or a court reporter, who has no way of knowing anything happened.

La asimetría es el problema de gobernanza. La única persona en la sala capaz de detectar el error es aquella cuyo significado fue reemplazado, y a menudo esa persona es la que nunca ve la salida. El texto producido de esta manera no se evapora: se escribe en una nota de admisión, una declaración, una evaluación de desempeño, un expediente. Una vez allí, adquiere la autoridad del registro en lugar del estatus de una traducción automática, y la persona Deaf debe impugnar un documento que pretende citarla. Un error fluido en este sentido no es una adaptación de acceso fallida. Es una atribución, y las instituciones apenas tienen mecanismos para retractarse de una.

La métrica no puede ver el fallo

Every number reported for this method, and for essentially every system like it, is BLEU or ROUGE. Both work by comparing the machine's output with a reference translation and measuring overlap of word sequences. They were built for spoken-language machine translation and they measure resemblance to a reference, which is not the same as fidelity to what the signer expressed. A fabricated sentence that happens to share common phrasing with the reference scores well. A correct translation phrased differently scores badly. The failure mode the method is designed to fix, fluent output unsupported by the visual evidence, is precisely the failure these metrics are least able to detect, because fluency is what they reward.

This is not a complaint from outside the field. A paper from the Centre for Vision, Speech and Signal Processing at the University of Surrey, posted this month, evaluates six translation systems and concludes that "gains in BLEU-4 are not on their own evidence of better sign language understanding," noting that the low-resource multimodal setting "allows models to exploit spurious correlations and spoken-language priors, rather than learning stronger sign representations." Their alternative protocol, which tests whether the content of the passage survived, reorders the field and reveals a gap between systems that BLEU-4 could not see at all. Read alongside that finding, a reported improvement in BLEU-4 is evidence that output resembles references more closely. It is not yet evidence that the system understood the signing, and it is certainly not evidence that confident fabrication has been reduced.

El tamaño de la ganancia reportada también importa aquí. Mejoras de aproximadamente uno a tres puntos BLEU-4 frente a líneas de base nombradas son un progreso



incremental ordinario, y se están ofreciendo como evidencia de que un sistema fabrica menos. Leído frente al hallazgo de Surrey, esa es la parte más débil del argumento: una ganancia de unos pocos puntos en una métrica que el propio campo dice que no rastrea la comprensión de la lengua de signos no puede establecer que la fabricación confiada se haya reducido. La afirmación puede seguir siendo cierta. Demostrarlo requeriría un instrumento distinto, y el artículo, que está en revisión y no publicado, es donde tendría que encontrarse esa demostración.

La confianza es una señal de entrenamiento, no una divulgación

The most interesting thing about the approach, taken at face value, is that it treats the model's own uncertainty as information worth acting on. That is the right instinct and it points at the thing procurement should be asking for, though not quite in the form the method provides. A confidence value used inside training changes how the model learns. It does not reach the person reading the output, and it does not change what the output looks like when the model is uncertain. The system still produces a fluent sentence, because producing fluent sentences is what it does. Nothing in the text says that three of its words were guesses.

Two things would change that, and neither is exotic. The first is abstention: the system should be able to decline. A sign-to-text system that can say "this passage was not clearly captured" and mark the gap is considerably safer than one that always returns a complete sentence, because a visible gap invites a human to fill it and a fluent invention does not. The second is exposure: confidence, where the system has it, should be visible at the point of use, and low-confidence spans should be marked in the text that reaches the reader. A vendor stating 95 percent accuracy is describing an average, and an average tells a buyer nothing about the character of the remaining five percent. Five percent noise is an inconvenience. Five percent confident fabrication, distributed unpredictably through a medical history or a witness statement, is a different product altogether, and the one figure most vendors publish cannot distinguish between them.

Lo que la entrada de pose transmite y lo que no

The reported system works from pose data, meaning tracked body keypoints rather than raw video. There is a real privacy advantage in that, since keypoints are less identifying than footage of a signer's face, and it matters in a field where the data is a per-

son's body. But it also raises a question that cannot be answered without the paper: which keypoints. Signed languages carry grammar on the face. A headshake can negate the sentence it accompanies, eyebrow position can mark the difference between a statement and a question, and mouth patterns disambiguate signs that the hands render identically. If a keypoint set samples the face sparsely, a system may be structurally unable to see the features that reverse a meaning, and the failure this produces is the fluent kind: a clean declarative sentence where the signer negated, or a statement where the signer asked. In a clinical or legal setting, the difference between a sentence and its negation is the whole of the content.

Los parámetros de referencia no son despliegues

CSL-Daily es un corpus de lengua de signos china grabado en un laboratorio sobre temas cotidianos. How2Sign es un amplio conjunto de vídeos instructivos de lengua de signos americana. Ambos son recursos de investigación serios y ninguno se asemeja a las condiciones en las que una institución realmente usaría la conversión de signos a texto: una persona con dolor, signando desde una cama de hospital en un ángulo para el que la cámara no fue diseñada, en una variedad regional, con una mano inmovilizada por una cánula, interrumpida, angustiada o cambiando de registro. La cifra reportada de How2Sign, 15,2 BLEU-4, merece tenerse en cuenta por esa razón. En el más abierto de los dos parámetros de referencia, en condiciones favorables, el estado de este campo produce una salida cuya coincidencia con una traducción de referencia humana sigue siendo modesta. Ese es un resultado de investigación, y razonable. No es una base sobre la cual debería confiarse nada para registrar lo que dijo un paciente Deaf.

Qué debería exigir un comprador

Six requirements follow, and they are specific to this direction of translation. The first is abstention with a visible gap, so that uncertainty appears as a gap rather than as prose. The second is confidence marking in the delivered text, available to the reader and not only to the model. The third is back-translation to the signer before anything is recorded, so that the Deaf person sees, in their own language, what the system is about to say in their name, and can stop it. The fourth is provenance in the record: text derived from machine translation should be labelled as such wherever it is stored,

with the source retained, so that a later dispute is about a machine's output rather than about the person's credibility.

The fifth is a correction route that reaches the record itself rather than the vendor's support desk, because an uncorrected sentence in a case file outlives any ticket. The sixth is evidence appropriate to the claim: semantic fidelity measured by protocols that test whether content survived, adversarial testing on negation, questions and role shift, disaggregation by signing variety and capture conditions, and comprehension checked with Deaf signers rather than inferred from reference overlap. NCG's position is that the evidence required rises with the stakes of the setting, and that a system whose errors enter an official record attributed to a person sits at the top of that scale, not in the middle of it.

Cómo la afirmación llegó al inglés

There is a second story here, and it belongs in a piece about evidence. The Quantum Bit report is properly sourced: it names the method, describes what it does, reports gains against named baselines and links the paper. By the time this reached English-speaking readers, it had passed through a site whose byline is an "AI editorial department", which describes itself as a portal for generative engine optimisation, meaning content written to be retrieved and repeated by AI search tools, and which scores its own articles for quality. That version dropped the link to the paper and converted gains measured against particular baselines into flat absolute scores. Nothing about it is fraudulent. It is simply a summary of a summary, optimised to be picked up by machines, and the thing it quietly removed was the route back to the evidence.

Vale la pena señalar esto porque así es como la mayoría de los responsables de contratación se encontrarán ahora con afirmaciones sobre esta tecnología. La afirmación de un proveedor entra en la literatura como un artículo presentado, es reportada con precisión por un medio especializado, es reescrita por un agregador para su recuperación, es recuperada por un asistente y llega a una nota informativa con las cifras endurecidas y la cita desaparecida. Cada paso es defendible por sí solo. El resultado es una cifra sin denominador, sin línea de base y sin documento detrás, presentada con más confianza de la que reclamaron los autores originales. La disciplina que protege aquí a un comprador es la misma que este artículo exige a la tecnología: rastrear la afirmación hasta aquello de lo que procede, y decir con claridad qué no se pudo leer.

Los límites de este análisis

El relato de CAPO-SLT aquí se basa en el reportaje comercial en chino sobre el trabajo, que nombra el artículo y lo enlaza. El propio artículo, alojado en un repositorio que exige un paso de verificación del navegador, no pudo leerse para este análisis, así que las cifras citadas son las reportadas, no verificadas contra la fuente, y nada de lo anterior debe leerse como una evaluación del sistema ni de vivo AI Lab, cuyo problema declarado es el correcto y cuyo método bien puede hacer lo que dice. El argumento se sostiene sobre el planteamiento del problema, que el campo ha nombrado ahora abiertamente, y sobre el trabajo publicado por otros acerca de lo que los parámetros de referencia estándar pueden y no pueden mostrar. No se ha aplicado a este sistema ningún instrumento de NCG, y el argumento más amplio no depende en absoluto de este método concreto.

It is worth ending on what is encouraging here, because there is something. A commercial research lab has publicly identified confident fabrication as the central defect in sign-to-text translation rather than reporting another incremental gain and leaving the failure unnamed. That is the honest version of the problem, and naming it is the precondition for measuring it. The risk now is the familiar one: that the naming travels, the fix is assumed, and an institution somewhere types a fluent sentence into a Deaf person's file in the belief that the problem was solved by a paper it never read.

Fuentes

1. Quantum Bit via 36kr, "CAPO-SLT achieves top scores on three Chinese metrics," reporting vivo AI Lab's method, 9 September 2026. https://36kr.com/p/3975775861813507
2. vivo AI Lab, "CAPO-SLT: Confidence-Aware Policy Optimization for Stable Sign Language Translation Generation," OpenReview (under review; not read for this analysis). https://openreview.net/forum?id=l4bCsrSkYx
3. Zgeo (AI editorial department), rewritten account of the above, 9 September 2026. https://www.zgeo.com.cn/news/capo-slt-sign-language-translation-vivo-ai-lab
4. Oline Ranum, Edward Fish, Simon Hadfield and Richard Bowden, "Beyond BLEU: A Case for Redefining Sign Language Translation Benchmarks," University of Surrey (CVSSP), arXiv:2609.03734, September 2026. https://arxiv.org/abs/2609.03734
5. "RVLF: A Reinforcing Vision-Language Framework for Gloss-Free Sign Language

Translation," arXiv:2512.07273 (reported CSL-Daily gloss-free results). https://arxiv.org/abs/2512.07273

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, September 11). Las palabras que te ponen en la boca: la IA de lengua de señas a texto y el error fluido. Novara Consulting Group. https://doi.org/10.5281/zenodo.23003302