The Words Put in Your Mouth: Sign-to-Text AI and Fluent Error

Heather Grizzle · September 11, 2026

Sign-to-text AI can invent fluent sentences a Deaf person never signed, then file them as record. Why BLEU scores cannot see the failure they claim to fix.

Contenuti

Confident mistranslation in sign-to-text systems, and the benchmarks that cannot see it

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 support 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.

The reported results are narrower than the headline framing suggests, and the difference matters. On the Chinese benchmark CSL-Daily, the reported gains are measured against two named systems that also work from pose input: roughly 1.26 BLEU-4 over Geo-Sign and 3.07 BLEU-4 over Uni-Sign, with the claim being the best scores among pose-only systems rather than the best in the field. On the American Sign Language benchmark How2Sign the reported figures are 41.4 BLEU-1, 15.2 BLEU-4 and 34.9 ROUGE-L, gains of about a point or less over Uni-Sign. The paper itself could not be read for this analysis, because the repository hosting it requires a browser verification step, so what follows examines the problem the work names, the direction of translation it operates in, and the kind of evidence being offered, rather than assessing the system. Those are worth examining on their own, because the problem is real, the direction is the one institutions handle least carefully, and the evidence type cannot demonstrate what it is being asked to demonstrate.

Which way the translation runs

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.

The asymmetry is the governance problem. The only person in the room who can detect the error is the person whose meaning was replaced, and that person is often the one who never sees the output. Text produced this way does not evaporate: it is typed into an intake note, a statement, a performance review, a case file. Once there, it car-

ries the authority of the record rather than the status of a machine translation, and the Deaf person must dispute a document that purports to quote them. A fluent error in this direction is not a failed access accommodation. It is an attribution, and institutions have almost no machinery for retracting one.

The metric cannot see the failure

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.

The size of the reported gain matters here too. Improvements of roughly one to three BLEU-4 points over named baselines are ordinary incremental progress, and they are being offered as evidence that a system fabricates less. Read against the Surrey finding, that is the weakest part of the case: a gain of a few points in a metric that its own field says does not track sign language understanding cannot establish that confident fabrication has been reduced. The claim may still be true. Demonstrating it would require a different instrument, and the paper, which is under review rather than published, is where that demonstration would have to be found.

Confidence is a training signal, not a disclosure

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.

What pose input does and does not carry

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 person'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.

Benchmarks are not deployments

CSL-Daily is a Chinese Sign Language corpus recorded in a laboratory on everyday topics. How2Sign is a large set of American Sign Language instructional videos. Both are serious research resources and neither resembles the conditions in which an institution would actually use sign-to-text: a person in pain, signing from a hospital bed at an angle the camera was not designed for, in a regional variety, with a hand immobilised by a cannula, interrupted, distressed, or shifting between registers. The reported How2Sign figure, 15.2 BLEU-4, is worth holding onto for that reason. On the more open-domain of the two benchmarks, under favourable conditions, the state of this field produces output whose overlap with a human reference translation remains modest. That is a research result, and a reasonable one. It is not a basis on which anything should be relied upon to record what a Deaf patient said.

What a buyer should require

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.

How the claim reached English

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.

This is worth naming because it is how most procurement officers will now encounter claims about this technology. A vendor claim enters the literature as a submitted paper, is reported accurately by a trade outlet, is rewritten by an aggregator for retrieval, is retrieved by an assistant and arrives in a briefing note with the numbers hardened and the citation gone. Each step is defensible on its own. The result is a figure with no denominator, no baseline and no document behind it, presented with more confidence than the original authors claimed. The discipline that protects a buyer here is the same one this article asks of the technology: trace the claim back to the thing it came from, and say plainly what you could not read.

I limiti di questa analisi

The account of CAPO-SLT here rests on Chinese-language trade reporting of the work, which names the paper and links it. The paper itself, hosted on a repository that requires a browser verification step, could not be read for this analysis, so the figures quoted are as reported rather than verified against the source, and nothing above should be read as an assessment of the system or of vivo AI Lab, whose stated problem is the right one and whose method may well do what it says. The argument stands on the problem statement, which the field has now named openly, and on the published work of others concerning what the standard benchmarks can and cannot show. No NCG

instrument has been applied to this system, and the wider point does not depend on this one method at all.

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.

Fonti

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

Come citare

Grizzle, H. M. (2026, September 11). The Words Put in Your Mouth: Sign-to-Text AI and Fluent Error. Novara Consulting Group. https://www.novaracg.com/it/2026/09/11/the-words-put-in-your-mouth-sign-to-text-ai-and-fluent-error/