NOVARA
CONSULTING GROUP

被塞进你嘴里的话:手语转文字人工智能与流畅型错误

Heather Grizzle · 2026年9月11日

文本转手语人工智能可能会编造出聋人从未打过的流畅句子,然后将其作为记录归档。这就是为什么*BLEU*分数看不出它们声称要解决的失败。

目录

手语转文字系统中的自信型误译,以及无法察觉它的基准测试

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.

报道中的结果比标题所暗示的要窄得多,而这一差异很重要。在中文基准测试CSL-Daily上,所报告的提升是相对于另外两个同样基于姿态输入的命名系统而言的:相较Geo-Sign约提升1.26个BLEU-4,相较Uni-Sign约提升3.07个BLEU-4,其说法是在仅使用姿态输入的系统中取得最佳成绩,而非在整个领域中最佳。在美国手语基准测试How2Sign上,报告的数字为41.4 BLEU-1、15.2 BLEU-4和34.9 ROUGE-L,相较Uni-Sign的提升约为一个点或更少。本分析未能读到论文原文,因为托管论文的仓库需要浏览器验证步骤,因此以下内容审视的是该研究所指出的问题、其所处理的翻译方向,以及所提供证据的类型,而非对该系统本身作出评估。这些内容本身就值得审视,因为这个问题是真实存在的,其所涉及的翻译方向恰恰是各机构处理最不谨慎的方向,而所提供的证据类型也无法证明它被要求证明的东西。

翻译方向为何

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 When a sign-to-text system fabricates a fluent sentence, the Deaf person's words have been replaced by words they did not say, and those words go to a clinician, a caseworker, an investigator or a court reporter, who has no way of knowing anything happened.

这种不对称性正是治理层面的问题所在。房间里唯一能察觉这一错误的人,是其原意被替换的那个人,而这个人往往正是从未看到输出结果的那个人。以这种方式产生的文字并不会凭空消失:它会被打进接诊记录、陈述书、绩效评估或案卷之中。一旦进入这些文件,它便获得了正式记录的权威性,而不再只是机器翻译的状态,聋人必须去反驳一份声

称在引用自己原话的文件。这个方向上的流畅型错误,并非一次访问便利措施的失败,而是一种归属认定,而各机构几乎没有任何机制来撤销这种认定。

该指标看不到失败

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.

这里所报告的提升幅度本身也很重要。相较命名基准系统约一到三个BLEU-4点的提升,是普通的渐进式进步,而它们却被当作系统编造程度降低的证据来提出。结合萨里大学的发现来看,这正是整个论证中最薄弱的一环:在一个连该领域自身都承认无法反映手语理解能力的指标上,提升几个点,不足以证明自信型编造已经减少。这一说法或许仍然成立,但要证明它,需要用不同的检验方式,而这种证明理应能在论文中找到——只是这篇论文尚处于审稿阶段,并未正式发表。

置信度是训练信号,而非披露信息

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.

姿态输入能承载什么、不能承载什么

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.

基准测试不等于实际部署

CSL-Daily是一个在实验室环境中录制的中国手语语料库,内容涉及日常话题。How2Sign是一套大型的美国手语教学视频集合。二者都是严肃的研究资源,但都不代表机构实际使用手语转文字技术时的真实情况:一个处于疼痛之中的人,从病床上以摄像机原本未曾设计适应的角度打手语,使用地区性变体,一只手因插着输液导管而无法活动,时而被打断,情绪激动,或在不同语域之间切换。正因如此,报告中的How2Sign数字——15.2 BLEU-4——值得牢记。在这两个基准测试中相对更为开放领域的那一个上,在有利条件下,该领域现有水平所产生的输出,与人工参考翻译的重合度依然有限。这是一个合理的研究结果。但它不应成为任何依赖于记录聋人患者所说内容的依据。

采购方应当要求什么

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.

这一说法是如何传入英语世界的

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.

这一点值得特别指出,因为如今大多数采购官员正是通过这样的路径接触到关于这项技术的说法。一项厂商说法,首先作为投稿论文进入学术文献,继而被行业媒体准确报道,再被聚合平台改写以便于检索,然后被某个人工智能助手检索到,最终出现在一份简报中,数字被硬化为确凿的结论,引用来源却已消失不见。每一步单独来看都说得过去。而最终结果是一个没有分母、没有基准、也没有背后文献支撑的数字,却比原作者所声称的更为斩钉截铁地呈现出来。能够保护采购方的那种审慎态度,正是本文对这项技术所要求的同一种态度:把说法追溯回它的出处,并坦率说明哪些内容自己未能核实。

本分析的局限性

本文对CAPO-SLT的介绍,依据的是对这项工作进行报道的中文行业媒体报道,该报道指名了论文并附有链接。论文本身托管在一个需要浏览器验证步骤的平台上,本分析未能读到原文,因此文中所引用的数字均为转述而非核对原始文献所得,以上内容均不应被解读为对该系统或对vivo AI Lab的评估——该实验室所提出的问题方向是正确的,其方法也很可能确如其所言那样有效。本文的论证立足于该领域如今已公开指出的问题本身,以及他人已发表的、关于标准基准测试能够证明什么、不能证明什么的研究成果。NCG尚未对该系统应用任何自有工具进行评估,而本文更广泛的论点,也完全不依赖于这一种具体方法。

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.

资料来源

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

引用方式

Grizzle, H. M. (2026, September 11). 被塞进你嘴里的话:手语转文字人工智能与流畅型错误. Novara Consulting Group. https://www.novaracg.com/zh/2026/09/11/the-words-put-in-your-mouth-sign-to-text-ai-and-fluent-error/