Les mots qu'on vous met dans la bouche : l'IA de traduction signe-vers-texte et l'erreur fluide

Heather Grizzle · 11 septembre 2026

L'IA de traduction signe-vers-texte peut inventer des phrases fluides qu'une personne Deaf n'a jamais signées, puis les consigner comme faisant foi. Pourquoi les scores BLEU ne peuvent pas voir l'échec qu'ils prétendent corriger.

Sommaire

Mauvaise traduction assurée dans les systèmes signe-vers-texte, et les benchmarks incapables de la voir

Le 9 septembre 2026, le média technologique chinois Quantum Bit a rapporté une méthode du vivo AI Lab appelée CAPO-SLT, issue d'un article intitulé « CAPO-SLT: Confidence-Aware Policy Optimization for Stable Sign Language Translation Generation », publié sur OpenReview, où les articles sont déposés pendant leur évaluation par les pairs. Le problème qu'il désigne est celui qui devrait inquiéter quiconque achète cette technologie. Un système de traduction peut produire une phrase fluide,

grammaticale et tout à fait plausible sans aucun appui dans ce que le signeur a réellement fait, et comme la phrase se lit bien, rien en elle n'annonce l'erreur. Comme le rapporte l'article, un seul mot localement plausible mais dépourvu de soutien visuel peut faire dévier toute une traduction, l'erreur s'accumulant tout au long de la phrase. La méthode ajuste les bornes à l'intérieur desquelles l'apprentissage par renforcement est autorisé à mettre à jour le modèle, jeton par jeton, selon le degré de confiance de la politique précédente : bornes plus serrées là où la confiance est élevée, proches de 0,2, et plus lâches là où elle est faible, jusqu'à 0,3, avec des plafonds supplémentaires sur les jetons porteurs d'un avantage négatif. L'objectif est d'empêcher le modèle de figer des suppositions confiantes et de laisser une marge pour que les suppositions incertaines soient ramenées vers la preuve visuelle.

Les résultats rapportés sont plus étroits que ne le laisse entendre le titre accrocheur, et cette différence compte. Sur le benchmark chinois CSL-Daily, les gains rapportés sont mesurés par rapport à deux systèmes nommés qui fonctionnent également à partir d'une entrée en pose : environ 1,26 BLEU-4 par rapport à Geo-Sign et 3,07 BLEU-4 par rapport à Uni-Sign, l'affirmation étant qu'il s'agit des meilleurs scores parmi les systèmes fondés uniquement sur la pose, plutôt que les meilleurs du domaine. Sur le benchmark de langue des signes américaine How2Sign, les chiffres rapportés sont 41,4 BLEU-1, 15,2 BLEU-4 et 34,9 ROUGE-L, soit des gains d'environ un point ou moins par rapport à Uni-Sign. L'article lui-même n'a pas pu être lu pour cette analyse, car le dépôt qui l'héberge exige une étape de vérification par navigateur ; ce qui suit examine donc le problème que ces travaux désignent, le sens de traduction dans lequel ils opèrent, et le type de preuve proposé, plutôt qu'une évaluation du système. Ces éléments méritent d'être examinés en eux-mêmes, car le problème est réel, le sens de traduction concerné est celui que les institutions gèrent le moins soigneusement, et le type de preuve ne peut pas démontrer ce qu'on lui demande de démontrer.

Dans quel sens va la traduction

La plupart des débats publics sur l'IA des langues signées concernent la production destinée aux personnes Deaf : un avatar sur un panneau de départs, une vidéo signante sur un site web, un robot dans une salle de classe. Le signe-vers-texte fonctionne dans l'autre sens. L'entrée est une personne Deaf en train de signer et la sortie est un texte que tout le monde d'autre lit, ce qui change la nature de l'erreur. Quand un avatar signe mal, une personne Deaf reçoit une mauvaise traduction et peut géné-

ralement dire que quelque chose ne va pas. Quand un système signe-vers-texte fabrique une phrase fluide, les mots de la personne Deaf ont été remplacés par des mots qu'elle n'a pas dits, et ces mots parviennent à un clinicien, un travailleur social, un enquêteur ou un sténographe judiciaire, qui n'a aucun moyen de savoir que quelque chose s'est produit.

Cette asymétrie constitue le problème de gouvernance. La seule personne dans la pièce capable de détecter l'erreur est celle dont le sens a été remplacé, et cette personne est souvent celle qui ne voit jamais le résultat. Le texte produit de cette manière ne s'évapore pas : il est saisi dans une note d'admission, une déclaration, une évaluation de performance, un dossier. Une fois là, il porte l'autorité du dossier officiel plutôt que le statut d'une traduction automatique, et la personne Deaf doit alors contester un document qui prétend la citer. Une erreur fluide dans ce sens n'est pas un échec d'accommodement d'accès. C'est une attribution, et les institutions n'ont presque aucun mécanisme pour la rétracter.

La métrique ne peut pas voir l'échec

Chaque chiffre rapporté pour cette méthode, et pour pratiquement tout système comparable, est un score BLEU ou ROUGE. Les deux fonctionnent en comparant la sortie de la machine à une traduction de référence et en mesurant le chevauchement des séquences de mots. Ils ont été conçus pour la traduction automatique de langues parlées et mesurent la ressemblance à une référence, ce qui n'est pas la même chose que la fidélité à ce que le signeur a exprimé. Une phrase fabriquée qui partage par hasard une formulation courante avec la référence obtient un bon score. Une traduction correcte formulée différemment obtient un mauvais score. Le mode d'échec que la méthode est censée corriger, une sortie fluide non étayée par la preuve visuelle, est précisément ce que ces métriques sont le moins capables de détecter, car la fluidité est ce qu'elles récompensent.

Il ne s'agit pas d'une critique venue de l'extérieur du domaine. Un article du Centre for Vision, Speech and Signal Processing de l'université de Surrey, publié ce mois-ci, évalue six systèmes de traduction et conclut que « les gains en BLEU-4 ne constituent pas en eux-mêmes une preuve d'une meilleure compréhension de la langue des signes », notant que le contexte multimodal à faibles ressources « permet aux modèles d'exploiter des corrélations fallacieuses et des a priori issus des langues parlées, plutôt que d'apprendre des représentations plus solides des signes ». Leur protocole alternatif,

qui teste si le contenu du passage a survécu, reclasse le domaine et révèle un écart entre systèmes que le BLEU-4 ne pouvait absolument pas voir. Lu à la lumière de cette conclusion, un gain rapporté en BLEU-4 constitue une preuve que la sortie ressemble davantage aux références. Ce n'est pas encore une preuve que le système a compris les signes, et ce n'est certainement pas une preuve que la fabrication assurée a été réduite.

L'ampleur du gain rapporté compte également ici. Des améliorations d'environ un à trois points de BLEU-4 par rapport à des références nommées relèvent d'un progrès incrémental ordinaire, et elles sont pourtant présentées comme la preuve qu'un système fabrique moins. À la lumière de la conclusion de Surrey, c'est là le point le plus faible de l'argument : un gain de quelques points sur une métrique dont le domaine lui-même dit qu'elle ne reflète pas la compréhension de la langue des signes ne peut pas établir que la fabrication assurée a été réduite. L'affirmation peut néanmoins être vraie. La démontrer exigerait un autre instrument, et l'article, qui est en cours d'évaluation plutôt que publié, est l'endroit où cette démonstration devrait être trouvée.

La confiance est un signal d'entraînement, pas une divulgation

L'élément le plus intéressant de cette approche, pris au pied de la lettre, est qu'elle traite l'incertitude propre du modèle comme une information sur laquelle agir. C'est le bon instinct, et il pointe vers ce que les acheteurs devraient exiger, quoique pas tout à fait sous la forme que fournit cette méthode. Une valeur de confiance utilisée à l'intérieur de l'entraînement change la façon dont le modèle apprend. Elle n'atteint pas la personne qui lit la sortie, et elle ne change pas l'apparence de la sortie lorsque le modèle est incertain. Le système produit toujours une phrase fluide, car produire des phrases fluides est ce qu'il fait. Rien dans le texte n'indique que trois de ses mots étaient des suppositions.

Deux choses changeraient cela, et aucune n'est exotique. La première est l'abstention : le système devrait pouvoir refuser. Un système signe-vers-texte capable de dire « ce passage n'a pas été clairement capté » et de marquer le vide est considérablement plus sûr qu'un système qui renvoie toujours une phrase complète, car un vide visible invite un humain à le combler, tandis qu'une invention fluide ne le fait pas. La seconde est l'exposition : la confiance, lorsque le système la possède, devrait être visible au point d'utilisation, et les segments à faible confiance devraient être marqués dans le texte

qui parvient au lecteur. Un vendeur annonçant une précision de 95 % décrit une moyenne, et une moyenne ne dit rien à un acheteur sur la nature des cinq pour cent restants. Cinq pour cent de bruit est un désagrément. Cinq pour cent de fabrication assurée, répartis de manière imprévisible dans un dossier médical ou une déclaration de témoin, est un produit d'une tout autre nature, et le seul chiffre que la plupart des vendeurs publient ne permet pas de les distinguer.

Ce que l'entrée en pose transporte et ne transporte pas

Le système rapporté fonctionne à partir de données de pose, c'est-à-dire des points-clés corporels suivis plutôt que la vidéo brute. Il y a là un véritable avantage en matière de confidentialité, puisque les points-clés identifient moins qu'un enregistrement du visage d'un signeur, et cela compte dans un domaine où les données concernent le corps d'une personne. Mais cela soulève aussi une question à laquelle on ne peut répondre sans l'article : quels points-clés. Les langues signées portent la grammaire sur le visage. Un hochement de tête peut nier la phrase qu'il accompagne, la position des sourcils peut marquer la différence entre une affirmation et une question, et les configurations de la bouche désambiguïsent des signes que les mains rendent identiques. Si un ensemble de points-clés échantillonne le visage de manière éparse, un système peut être structurellement incapable de percevoir les traits qui inversent un sens, et l'échec qui en résulte est du genre fluide : une phrase déclarative bien construite là où le signeur a nié, ou une affirmation là où le signeur a posé une question. Dans un contexte clinique ou judiciaire, la différence entre une phrase et sa négation constitue tout le contenu.

Les benchmarks ne sont pas des déploiements

CSL-Daily est un corpus de langue des signes chinoise enregistré en laboratoire sur des sujets du quotidien. How2Sign est un vaste ensemble de vidéos pédagogiques en langue des signes américaine. Les deux sont des ressources de recherche sérieuses, et ni l'une ni l'autre ne ressemble aux conditions dans lesquelles une institution utiliserait réellement le signe-vers-texte : une personne souffrante, signant depuis un lit d'hôpital sous un angle que la caméra n'était pas conçue pour capter, dans une variante régionale, avec une main immobilisée par une canule, interrompue, en détresse, ou passant d'un registre à l'autre. Le chiffre rapporté pour How2Sign, 15,2 BLEU-4, mérite d'être retenu pour cette raison. Sur celui des deux benchmarks qui est

le plus ouvert, dans des conditions favorables, l'état de ce domaine produit des résultats dont le chevauchement avec une traduction de référence humaine reste modeste. C'est un résultat de recherche raisonnable. Ce n'est pas une base sur laquelle on devrait s'appuyer pour consigner ce qu'un patient Deaf a dit.

Ce qu'un acheteur devrait exiger

Six exigences en découlent, propres à ce sens de traduction. La première est l'abstention avec un vide visible, afin que l'incertitude apparaisse comme un vide plutôt que comme de la prose. La deuxième est le marquage de la confiance dans le texte livré, accessible au lecteur et pas seulement au modèle. La troisième est la rétro-traduction vers le signeur avant tout enregistrement, afin que la personne Deaf voie, dans sa propre langue, ce que le système s'apprête à dire en son nom, et puisse l'arrêter. La quatrième est la traçabilité dans le dossier : le texte issu d'une traduction automatique devrait être étiqueté comme tel partout où il est conservé, avec la source conservée, afin qu'un litige ultérieur porte sur la sortie d'une machine plutôt que sur la crédibilité de la personne.

La cinquième est une voie de correction qui atteint le dossier lui-même plutôt que le service d'assistance du fournisseur, car une phrase non corrigée dans un dossier survit à tout ticket d'incident. La sixième est une preuve adaptée à l'affirmation : une fidélité sémantique mesurée par des protocoles testant si le contenu a survécu, des tests contradictoires sur la négation, les questions et le changement de rôle, une désagrégation par variante signée et conditions de capture, et une compréhension vérifiée auprès de signeurs Deaf plutôt qu'inférée du chevauchement avec les références. La position de NCG est que le niveau de preuve exigé augmente avec les enjeux du contexte, et qu'un système dont les erreurs entrent dans un dossier officiel attribué à une personne se situe au sommet de cette échelle, pas au milieu.

Comment l'affirmation est arrivée en anglais

Il y a ici une seconde histoire, qui trouve sa place dans un article consacré à la preuve. Le rapport de Quantum Bit est correctement sourcé : il nomme la méthode, décrit ce qu'elle fait, rapporte des gains par rapport à des références nommées et renvoie vers l'article. Au moment où cela est parvenu aux lecteurs anglophones, le contenu était passé par un site dont la signature est un « département éditorial IA », qui se décrit lui-même comme un portail d'optimisation pour moteurs génératifs, c'est-à-dire un

contenu écrit pour être récupéré et répété par des outils de recherche IA, et qui note ses propres articles pour leur qualité. Cette version a supprimé le lien vers l'article et transformé des gains mesurés par rapport à des références particulières en scores absolus plats. Rien de tout cela n'est frauduleux. C'est simplement le résumé d'un résumé, optimisé pour être repris par des machines, et ce qui a été discrètement retiré, c'est le chemin de retour vers la preuve.

Il vaut la peine de le nommer, car c'est ainsi que la plupart des responsables des achats rencontreront désormais les affirmations sur cette technologie. L'affirmation d'un vendeur entre dans la littérature sous forme d'article soumis, est rapportée fidèlement par un média spécialisé, réécrite par un agrégateur en vue de sa récupération, récupérée par un assistant, et arrive dans une note de synthèse avec les chiffres durcis et la citation disparue. Chaque étape est défendable en elle-même. Le résultat est un chiffre sans dénominateur, sans référence et sans document derrière lui, présenté avec plus d'assurance que les auteurs originaux n'en revendiquaient. La discipline qui protège un acheteur ici est la même que celle que cet article exige de la technologie : remonter l'affirmation jusqu'à sa source, et dire clairement ce qu'on n'a pas pu lire.

Les limites de cette analyse

Le récit de CAPO-SLT présenté ici repose sur un reportage professionnel en langue chinoise sur ces travaux, qui nomme l'article et le lie. L'article lui-même, hébergé sur un dépôt exigeant une étape de vérification par navigateur, n'a pas pu être lu pour cette analyse ; les chiffres cités sont donc rapportés tels quels plutôt que vérifiés par rapport à la source, et rien ci-dessus ne doit être lu comme une évaluation du système ou du vivo AI Lab, dont le problème énoncé est le bon et dont la méthode fait peut-être bien ce qu'elle annonce. L'argument repose sur l'énoncé du problème, que le domaine a désormais nommé ouvertement, et sur les travaux publiés d'autres chercheurs concernant ce que les benchmarks standards peuvent et ne peuvent pas montrer. Aucun instrument de NCG n'a été appliqué à ce système, et l'argument plus large ne dépend en rien de cette seule méthode.

Il vaut la peine de conclure sur ce qu'il y a d'encourageant ici, car il y en a. Un laboratoire de recherche commercial a publiquement identifié la fabrication assurée comme le défaut central de la traduction signe-vers-texte, plutôt que de rapporter un énième gain incrémental en laissant l'échec sans nom. C'est la version honnête du problème, et le nommer est la condition préalable pour le mesurer. Le risque maintenant est le

risque familier : que le fait de nommer le problème se propage, que la solution soit tenue pour acquise, et que quelque part une institution saisisse une phrase fluide dans le dossier d'une personne Deaf en croyant que le problème a été résolu par un article qu'elle n'a jamais lu.

Sources

1. Quantum Bit via 36kr, « CAPO-SLT obtient les meilleurs scores sur trois métriques chinoises », rapportant la méthode du vivo AI Lab, 9 septembre 2026. https://36kr.com/p/3975775861813507
2. vivo AI Lab, « CAPO-SLT: Confidence-Aware Policy Optimization for Stable Sign Language Translation Generation », OpenReview (en cours d'évaluation ; non lu pour cette analyse). https://openreview.net/forum?id=l4bCsrSkYx
3. Zgeo (département éditorial IA), récit réécrit de ce qui précède, 9 septembre 2026. https://www.zgeo.com.cn/news/capo-slt-sign-language-translation-vivo-ai-lab
4. Oline Ranum, Edward Fish, Simon Hadfield et Richard Bowden, « Beyond BLEU: A Case for Redefining Sign Language Translation Benchmarks », université de Surrey (CVSSP), arXiv:2609.03734, septembre 2026. https://arxiv.org/abs/2609.03734
5. « RVLF: A Reinforcing Vision-Language Framework for Gloss-Free Sign Language Translation », arXiv:2512.07273 (résultats rapportés sans gloses sur CSL-Daily). https://arxiv.org/abs/2512.07273

Comment citer

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/fr/2026/09/11/the-words-put-in-your-mouth-sign-to-text-ai-and-fluent-error/