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 démonstration n'est pas la preuve : ce que l'IA de langue des signes de Google DeepMind doit encore prouver

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.

Sommaire

1. À chacun son dû
2. Un benchmark est une preuve, pas la base de preuves
3. La prochaine question est celle de la désagrégation
4. La participation n'est pas une validation indépendante
5. Le problème de gouvernance le plus difficile survient après le lancement
6. Ce qui vient ensuite

Le 12 août, Google DeepMind a annoncé SL2T, un modèle multilingue de traduction de la langue des signes vers le texte, et a commencé à le déployer dans des produits grand public via Gboard et Live Transcribe sur le Pixel 11. La version initiale traduit l'ASL en anglais, avec d'autres appareils et langues des signes prévus, et DeepMind indique que le modèle sous-jacent a été entraîné sur plus de 100 000 heures de données couvrant plus de 50 langues des signes. À tous égards, cela fait franchir une étape importante à l'IA en langue des signes : elle sort du laboratoire de recherche pour entrer dans un produit que des gens ordinaires porteront dans leur poche.

Ce franchissement est significatif, mais il ne faut pas le confondre avec autre chose qu'il n'est pas. Une percée technologique et un déploiement validé sont deux choses différentes, et cette distinction compte énormément lorsque la technologie en ques-

tion traite une langue naturelle utilisée par une communauté linguistique historiquement marginalisée. La question que pose désormais SL2T n'est pas de savoir si le modèle fonctionne dans une démonstration. C'est de savoir si les preuves qui le sous-tendent peuvent supporter le poids que le déploiement va y faire peser.

À chacun son dû

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.

Ce rapport fait quelque chose qui reste rare dans l'IA en langue des signes : il indique où le système ne devrait pas être utilisé. SL2T 1.0 est explicitement décrit comme un outil d'aide à la saisie à faible enjeu. Il n'a pas été conçu ni validé pour les soins de santé, les procédures judiciaires, l'éducation, les décisions d'emploi, les déterminations de prestations gouvernementales, ou d'autres contextes à enjeux élevés, et le rapport indique clairement que la technologie ne doit pas se substituer à des interprètes qualifiés ni servir de moyen pour les organisations de contourner leurs obligations en matière d'accessibilité. Cela relève de la gouvernance, et cela mérite d'être reconnu. Ce n'est aussi que le début de la gouvernance.

Un benchmark est une preuve, pas la base de preuves

DeepMind fait état de performances solides. Sur FLEURS-ASL, SL2T aurait obtenu un score BLEURT en zero-shot de 70, supérieur à tout résultat précédemment rapporté, avec des évaluations supplémentaires sur PRESTO-ASL, une variante à une main de FLEURS-ASL, et des données de dactylologie FSboard. L'entreprise a également mené des tests préalables au lancement avec des employés Sourds de Google, une étude externe de type journal de bord avec des participants Sourds, et une évaluation pratique par des membres nord-américains de l'AISLAC. Ce sont des formes de preuves significatives, et aucune d'elles ne doit être écartée.



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.

Ce ne sont pas des cas limites triviaux. L'expression faciale, le mouvement de la tête, la structure spatiale, les classificateurs, la variation régionale et le contexte discursif ne sont pas des ornements du langage signé. Ils sont la langue elle-même. Un système qui gère bien le vocabulaire sous forme de citation tout en trébuchant sur la grammaire portée par le visage et l'espace de signature n'a pas résolu la traduction de la langue des signes ; il en a résolu une partie, et la partie non résolue affecte inégalement les personnes qui signent.

La prochaine question est celle de la désagrégation

Cette inégalité est précisément la raison pour laquelle la prochaine phase de preuves compte. DeepMind reconnaît que les performances des benchmarks n'ont pas encore été pleinement désagrégées selon des variables telles que l'âge, le genre, l'origine ethnique, la région, le sociolecte, et l'ASL noir américain. Le rapport note également que l'évaluation formelle impliquant des signeurs présentant des handicaps moteurs ou des schémas de mouvement atypiques reste limitée, et que le modèle n'a été ni entraîné ni formellement évalué sur des enfants de moins de 18 ans.

Un chiffre de performance agrégé ne peut pas nous dire si le système sert aussi bien les signeurs d'ASL noir américain que les autres, comment il gère une signature fortement régionale, ni comment la précision varie selon les groupes d'âge. Il ne peut pas nous dire comment le modèle se comporte pour les personnes ayant des différences de membres, des tremblements, une paralysie cérébrale, de l'arthrite, ou d'autres conditions qui influencent le mouvement, ni si la précision varie selon l'angle de la caméra, l'éclairage, le contraste de la peau, la vitesse de signature, l'espace de signature, ou l'appareil. Il ne peut pas distinguer les erreurs qui reformulent simplement un énoncé de celles qui en changent le sens. Et il ne peut pas répondre à la question qui sous-



tend toutes les autres : qui décide du niveau d'erreur acceptable, et pour qui. Cette dernière question ne peut être tranchée par une équipe d'ingénierie, aussi compétente soit-elle.

La participation n'est pas une validation indépendante

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.

À mesure que l'IA en langue des signes mûrit, le domaine devrait s'attendre à ce que tout domaine scientifique en maturation attend : des chercheurs externes capables de tester les affirmations de manière indépendante, de reproduire les évaluations lorsque cela est possible, d'examiner les performances par sous-groupe, de sonder les conditions adverses, et d'étudier le comportement réel après le déploiement. La technologie d'accessibilité ne devrait pas être soumise à une norme de preuve inférieure au motif que son objectif est bénéfique. Si quelque chose, la population qu'elle sert a encore plus en jeu dans la réponse.

Le problème de gouvernance le plus difficile survient après le lancement

DeepMind et l'AISLAC identifient explicitement le détournement institutionnel comme un risque critique, et ce jugement pourrait s'avérer plus important que n'importe quel score de benchmark. Une fois qu'une technologie de communication automatisée peu coûteuse existe, les écoles, les hôpitaux, les employeurs, les agences gouvernementales, les tribunaux et les entreprises font face à une puissante incitation économique à l'utiliser au-delà de son champ d'application prévu. Un produit conçu pour commander un café devient un produit auquel quelqu'un recourt aux urgences. Un outil conçu pour rédiger des messages texte devient la raison pour laquelle un employeur conclut qu'un interprète n'est plus nécessaire. Une fonctionnalité pratique se transforme discrètement en infrastructure.

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 gouvernance de l'IA, plus largement, converge vers le même territoire. Les questions centrales sont de moins en moins de savoir si un système peut accomplir une tâche, et de plus en plus de savoir qui le contrôle, comment ses performances sont évaluées, quelle surveillance existe, ce qui se passe lorsqu'il se comporte de manière inattendue, et qui reste responsable. Plus tôt ce mois-ci, des démocrates de la Chambre des représentants américaine ont interpellé OpenAI et Anthropic pour obtenir des réponses après que des agents d'IA ont échappé à leurs environnements de test et atteint des systèmes externes, demandant comment les systèmes étaient surveillés, quels contrôles de sécurité existaient, et si des audits indépendants devraient être exigés. Les technologies diffèrent ; le principe de gouvernance ne diffère pas. La capacité ne diminue pas le besoin de surveillance. Elle l'augmente.

Ce qui vient ensuite

Google a déjà produit des preuves que SL2T peut traduire l'ASL. Les questions les plus difficiles découlent du déploiement plutôt que de la capacité : les performances peuvent-elles être reproduites de manière indépendante ; comment la précision varie-t-elle selon les groupes démographiques, linguistiques, géographiques et de handicap ; quels seront les taux d'échec réels, et quels échecs modifient le sens plutôt que la formulation ; comment les utilisateurs sauront-ils quand le modèle est incertain, et comment les préjudices seront-ils signalés lorsqu'il se trompe ; les chercheurs externes recevront-ils un accès suffisant pour évaluer le système de manière significative ; et la gouvernance communautaire conservera-t-elle une influence réelle à mesure que le produit se développe mondialement, y compris lorsque des acheteurs institutionnels dépasseront les limites que Google a fixées, et lorsque ces limites devront rester visibles deux ou trois générations de produits plus tard.

Une démonstration montre ce qu'un système peut faire. Un benchmark dit quelque chose sur la qualité avec laquelle il le fait. Une étude utilisateur capture la manière dont les gens en font l'expérience, et une structure de gouvernance enregistre qui a voix au chapitre. Aucun de ces éléments seul ne répond à toutes les questions que sou-

lève un déploiement grand public ; cela nécessite un écosystème de preuves, construit dans le temps et ouvert aux acteurs extérieurs.

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.

Comment citer

Grizzle, H. M. (2026, August 15). La démonstration n'est pas la preuve : ce que l'IA de langue des signes de Google DeepMind doit encore prouver. Novara Consulting Group. https://doi.org/10.5281/zenodo.23003335