NOVARA
CONSULTING GROUP

Ten dokument został przetłumaczony z języka angielskiego maszynowo i może zawierać błędy. Oryginał w języku angielskim znajduje się pod adresem https://www.novaracg.com/2026/08/15/the-demo-is-not-the-evidence-what-google-deepminds-sign-language-ai-must-prove-next/.

Demonstracja to nie dowód: co musi jeszcze udowodnić sztuczna inteligencja do języka migowego od Google DeepMind

Heather Grizzle · 15 sierpnia 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.

Spis treści

1. Uznanie tam, gdzie się należy
2. Test porównawczy to dowód, nie baza dowodowa
3. Kolejnym pytaniem jest dezagregacja danych
4. Udział nie jest niezależną walidacją
5. Najtrudniejszy problem zarządzania pojawia się po uruchomieniu
6. Co dalej

12 sierpnia Google DeepMind ogłosiło SL2T, wielojęzyczny model tłumaczący język migowy na tekst, i zaczęło wdrażać go w produktach konsumenckich za pośrednictwem Gboard i Live Transcribe na Pixelu 11. Wersja początkowa tłumaczy amerykański język migowy (ASL) na angielski, planowane są kolejne urządzenia i języki migowe, a DeepMind podaje, że model bazowy został wytrenowany na ponad 100 000 godzin danych obejmujących ponad 50 języków migowych. Bez wątpienia jest to ważny krok dla sztucznej inteligencji języka migowego: wyjście z laboratorium badawczego do produktu, który zwykli ludzie będą nosić w kieszeni.

To przekroczenie granicy jest istotne, ale nie należy go mylić z czymś, czym nie jest. Przełom technologiczny i zweryfikowane wdrożenie to dwie różne sprawy, a to rozróż-

nienie ma ogromne znaczenie, gdy dana technologia przetwarza naturalny język używany przez historycznie marginalizowaną społeczność językową. Pytanie, jakie stawia teraz SL2T, nie brzmi, czy model działa w demonstracji. Brzmi ono, czy dowody stojące za nim udźwignią ciężar, jaki nałoży na nie wdrożenie.

Uznanie tam, gdzie się należy

Tego wydania nie należy sprowadzać do znanej narracji o firmie technologicznej budującej coś dla społeczności Deaf bez jej udziału. Według Google DeepMind, osoby Deaf były zaangażowane na każdym etapie projektu, od konceptualizacji i zbierania danych po badania użytkowników i ocenę skutków. Firma powołała także Komitet Doradczy ds. Sztucznej Inteligencji i Języka Migowego, AISLAC, w którego skład wchodzą przedstawiciele powiązani z National Association of the Deaf, National Technical Institute for the Deaf przy Rochester Institute of Technology, Deaf Professional Arts Network oraz World Federation of the Deaf. Wraz z wydaniem produktu, DeepMind i AISLAC opublikowały wspólny raport dotyczący jego wpływu.

Raport robi coś, co wciąż jest rzadkością w dziedzinie sztucznej inteligencji dla języka migowego: wskazuje, gdzie system nie powinien być stosowany. SL2T 1.0 jest wyraźnie opisywany jako narzędzie wspomagające o niskiej stawce ryzyka. Nie został zaprojektowany ani zweryfikowany do zastosowań w opiece zdrowotnej, postępowaniach prawnych, edukacji, decyzjach zatrudnieniowych, ustalaniu świadczeń rządowych ani innych sytuacjach wysokiego ryzyka, a raport stwierdza wprost, że technologia ta nie powinna zastępować wykwalifikowanych tłumaczy ani służyć organizacjom jako sposób na obejście obowiązków w zakresie dostępności. To jest zarządzanie i zasługuje na uznanie. To także dopiero początek zarządzania.

Test porównawczy to dowód, nie baza dowodowa

DeepMind podaje wysokie wyniki. Na FLEURS-ASL SL2T uzyskało podobno wynik BLEURT zero-shot na poziomie 70, wyższy niż jakikolwiek wcześniej zgłoszony rezultat, wraz z dalszymi ocenami na PRESTO-ASL, jednorękiej wersji FLEURS-ASL, oraz na danych fingerspellingowych FSboard. Firma przeprowadziła również testy przedpremierowe z pracownikami Google z społeczności Deaf, zewnętrzne badanie dzienniczkowe z uczestnikami Deaf oraz praktyczną ocenę przez północnoamerykańskich członków AISLAC. To znaczące formy dowodów i żadnej z nich nie należy odrzucać.

Ale właśnie tu publiczna dyskusja o sztucznej inteligencji zwykle traci precyzję, ponieważ wykazana skuteczność nie jest tym samym co wszechstronna walidacja. Model może radzić sobie imponująco w wybranych testach porównawczych, jednocześnie zachowując poważne słabości w odniesieniu do różnych populacji, środowisk, odmian języka, stylów migania, urządzeń i zastosowań. Sam raport DeepMind to potwierdza. Model może mieć trudności ze złożonością gramatyczną, znacznikami niemanualnymi, znakami regionalnymi, znakami wieloznacznymi, fingerspellingiem, nazwami własnymi, nietypowym ustawieniem kamery, słabym oświetleniem, slangiem oraz dłuższym kontekstem konwersacyjnym. Może też generować błędny "tekst-widmo" lub podejmować przewidywane domysły, zanim osoba migająca zakończy wypowiedź.

To nie są błahe przypadki brzegowe. Mimika twarzy, ruch głowy, struktura przestrzenna, klasyfikatory, zróżnicowanie regionalne i kontekst dyskursu nie są ozdobnikami języka migowego. To jest język. System, który dobrze radzi sobie ze słownictwem w formie cytacyjnej, ale słabnie przy gramatyce niesionej przez twarz i przestrzeń migową, nie rozwiązał problemu tłumaczenia języka migowego; rozwiązał jedynie jego część, a nierozwiązana część rozkłada się nierówno wśród osób migających.

Kolejnym pytaniem jest dezagregacja danych

Ta nierówność jest właśnie powodem, dla którego kolejny etap gromadzenia dowodów ma znaczenie. DeepMind przyznaje, że wyniki testów porównawczych nie zostały jeszcze w pełni zdezagregowane pod względem zmiennych takich jak wiek, płeć, pochodzenie etniczne, region, socjolekt oraz Black ASL. Raport zauważa również, że formalna ocena obejmująca osoby migające z niepełnosprawnościami ruchowymi lub nietypowymi wzorcami ruchu wciąż jest ograniczona oraz że model nie był ani trenowany, ani formalnie oceniany na dzieciach poniżej 18 roku życia.

Zbiorczy wskaźnik skuteczności nie mówi nam, czy system służy osobom migającym w Black ASL tak samo dobrze jak innym, jak radzi sobie z silnie regionalnym miganiem, ani jak dokładność zmienia się w zależności od grupy wiekowej. Nie mówi nam, jak model działa u osób z różnicami kończyn, drżeniem, mózgowym porażeniem dziecięcym, artretyzmem lub innymi schorzeniami wpływającymi na ruch, ani czy dokładność zmienia się w zależności od kąta kamery, oświetlenia, kontrastu skóry, prędkości migania, przestrzeni migowej czy urządzenia. Nie potrafi odróżnić błędów, które jedynie inaczej formułują wypowiedź, od błędów, które zmieniają jej znaczenie. I nie potrafi odpowiedzieć na pytanie leżące u podstaw tego wszystkiego: kto decyduje, jaki

poziom błędu jest akceptowalny i dla kogo. Tego ostatniego pytania nie może rozstrzygnąć zespół inżynierski, choćby najbardziej kompetentny.

Udział nie jest niezależną walidacją

Warto zachować dodatkowe rozróżnienie. AISLAC jest istotnym mechanizmem zarządzania; udział organizacji Deaf, badania użytkowników i wspólna ocena wpływu mają znaczenie. Ale udział doradczy i niezależna ocena techniczna pełnią różne funkcje. Opublikowane dowody obecnie składają się z ocen przeprowadzonych przez twórcę, testów z udziałem użytkowników i doradców zebranych w ramach własnej struktury zarządzania projektu oraz raportu współautorstwa Google DeepMind i uczestników AISLAC. To rzeczywista baza dowodowa, ale nie jest to niezależna replikacja.

W miarę dojrzewania sztucznej inteligencji dla języka migowego dziedzina ta powinna oczekiwać tego, czego oczekuje każda dojrzewająca dziedzina naukowa: zewnętrznych badaczy zdolnych do niezależnego testowania twierdzeń, odtwarzania ocen tam, gdzie to możliwe, badania skuteczności w podgrupach, testowania warunków adwersarialnych oraz badania zachowania systemu w warunkach rzeczywistych po wdrożeniu. Technologia dostępności nie powinna podlegać niższym standardom dowodowym tylko dlatego, że jej cel jest korzystny. Wręcz przeciwnie, populacja, której służy, ma tu jeszcze więcej do stracenia.

Najtrudniejszy problem zarządzania pojawia się po uruchomieniu

DeepMind i AISLAC wyraźnie wskazują instytucjonalne nadużycie jako krytyczne ryzyko, a ta ocena może okazać się ważniejsza niż jakikolwiek wynik testu porównawczego. Gdy tylko powstanie niedroga zautomatyzowana technologia komunikacyjna, szkoły, szpitale, pracodawcy, agencje rządowe, sądy i firmy staną przed silną ekonomiczną pokusą, by używać jej poza zamierzonym zakresem. Produkt zaprojektowany do zamawiania kawy staje się produktem, po który ktoś sięga na oddziale ratunkowym. Narzędzie stworzone do pisania wiadomości tekstowych staje się powodem, dla którego pracodawca uznaje, że tłumacz nie jest już potrzebny. Funkcja wygody po cichu przekształca się w infrastrukturę.

Raport DeepMind przewiduje ten dryf i wprost odrzuca zastępowanie tłumaczy. Otwartym pytaniem pozostaje, czy ta granica przetrwa zderzenie z zamówieniami pu-

blicznymi, budżetowaniem, rozszerzaniem produktu oraz nieuniknionym momentem, gdy administrator zapyta, czy tańsza zautomatyzowana opcja jest wystarczająco dobra. Zadeklarowane granice są konieczne. Egzekwowane granice mają znaczenie.

Zarządzanie sztuczną inteligencją w szerszym ujęciu zmierza w tym samym kierunku. Kluczowe pytania coraz rzadziej dotyczą tego, czy system potrafi wykonać zadanie, a coraz częściej tego, kto go kontroluje, jak ocenia się jego skuteczność, jaki istnieje monitoring, co się dzieje, gdy zachowuje się nieoczekiwanie, i kto pozostaje odpowiedzialny. Na początku tego miesiąca demokraci z Izby Reprezentantów USA naciskali na OpenAI i Anthropic o odpowiedzi po tym, jak agenty AI wymknęły się ze swoich środowisk testowych i dotarły do systemów zewnętrznych, pytając, jak systemy te były monitorowane, jakie istniały zabezpieczenia oraz czy należy wymagać niezależnych audytów. Technologie się różnią; zasada zarządzania nie. Możliwości nie zmniejszają potrzeby nadzoru. Zwiększają ją.

Co dalej

Google udowodniło już, że SL2T potrafi tłumaczyć ASL. Trudniejsze pytania pojawiają się w związku z wdrożeniem, a nie z możliwościami: czy skuteczność może zostać niezależnie odtworzona; jak dokładność zmienia się w zależności od grup demograficznych, językowych, geograficznych i związanych z niepełnosprawnością; jaki okaże się rzeczywisty odsetek błędów i które błędy zmieniają znaczenie, a nie tylko formę; skąd użytkownicy będą wiedzieć, kiedy model jest niepewny, i jak będą zgłaszane szkody, gdy się myli; czy zewnętrzni badacze otrzymają wystarczający dostęp, by rzeczywiście ocenić system; oraz czy zarządzanie przez społeczność zachowa rzeczywisty wpływ w miarę globalnego skalowania produktu, w tym gdy instytucjonalni nabywcy przekroczą granice wyznaczone przez Google, i gdy te granice będą musiały pozostać widoczne za dwie lub trzy generacje produktu.

Demonstracja pokazuje, co system potrafi zrobić. Test porównawczy mówi coś o tym, jak dobrze to robi. Badanie użytkowników rejestruje, jak ludzie go doświadczają, a struktura zarządzania zapisuje, kto ma głos. Żadne z nich samo w sobie nie odpowiada na wszystkie pytania, jakie rodzi wdrożenie konsumenckie; to wymaga ekosystemu dowodowego, budowanego z czasem i otwartego na zewnętrznych obserwatorów.

SL2T może okazać się najważniejszym dotychczasowym osiągnięciem w konsumenckiej sztucznej inteligencji dla języka migowego, a Google podjęło kroki w zakresie zarządzania, które zasługują na poważną uwagę: udział społeczności Deaf, wyraźne gra-

nice wdrożenia, publiczne ujawnienie ograniczeń oraz zadeklarowaną odmowę zastępowania tłumaczy. Odpowiednią reakcją nie jest ani celebracja, ani odrzucenie. Jest nią kontrola, ponieważ gdy sztuczna inteligencja języka migowego przechodzi z laboratorium badawczego do czyjejś kieszeni, zmienia się standard. Demonstracja już nie wystarczy. Teraz muszą pojawić się dowody.

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.

Jak cytować

Grizzle, H. M. (2026, August 15). The Demo Is Not the Evidence: What Google DeepMind's Sign Language AI Must Prove Next. Novara Consulting Group. https://doi.org/10.5281/zenodo.23003335