NOVARA
CONSULTING GROUP

Der Beweis, der es nie gab und immer noch nicht gibt

Heather M. Grizzle · 4. August 2026

Eine Untersuchung der fehlenden unabhängigen Evidenz hinter den größten KI-Gebärdensprachen-Behauptungen des Jahres 2026, von SignGemma bis hin zu kommerziellen Übersetzungstools.

DER FÜNFTE PARAMETER

Novara Consulting Group · Novara Press · Zeitschrift für Governance und Evidenz in der Barrierefreiheits-KI

Der Fünfte Parameter, Bd. 1 · Ausgabe Nr. 13

AUTORIN Heather M. Grizzle

ORCID 0009-0005-2605-8256

STAND DER EVIDENZ Aktuell bis August 2026

SCHLÜSSELWÖRTER Gebärdensprach-KI; unabhängige Evaluierung; Evidenzklassen; SignGemma; Transparenz; Vertrauen; Deaf-Zugang

---

Die wichtigste Tatsache über KI und Gebärdensprache im Jahr 2026 ist nicht, was geschah.

Es ist das, was nicht geschehen ist.

Zwischen Januar und August kündigten Unternehmen Produkte an, demonstrierten Systeme, sammelten Geld ein und sprachen über die Zukunft der Kommunikation. Neue Werkzeuge wurden als schneller, leistungsfähiger und näher am alltäglichen Gebrauch präsentiert. Einige zogen begeisterte Schlagzeilen an. Andere kamen mit beeindruckenden Behauptungen über Trainingsdaten, Antwortzeiten oder die Anzahl der Gebärden, die sie erkennen konnten.

Was nicht auftauchte, war unabhängiger Nachweis dafür, dass diese Systeme so funktionierten, wie beworben. Kein Dritter veröffentlichte eine öffentliche Messung der

Genauigkeit, der Fehlerraten oder des Nutzerverständnisses für irgendeines der vom Index untersuchten Hauptsysteme. Recherchen in den erforderlichen Quellen fanden keinen unabhängigen Benchmark für Signapse, Sorenson AST, Rylo Sign, Sign-Speak, ChatSign, SignVrse, Talksign oder Googles SignGemma.

Die formale Formulierung des Befunds ist vorsichtig: Es wurden „bis August 2026“ keine Beweise der Klasse A „lokalisiert“. Das lässt die Möglichkeit offen, dass irgendwo außerhalb der öffentlichen Aufzeichnungen Beweise existieren. Sie könnten in einem Unternehmen liegen, bei einem Kunden gehalten werden oder in noch nicht veröffentlichter Forschung enthalten sein.

Aber Beweise, die nicht überprüft werden können, können keine öffentliche Behauptung untermauern. Diese Unterscheidung ist wichtig, denn die Branche war nicht still. Sie hat einen stetigen Strom von Ankündigungen, Partnerschaften, Demonstrationen und Finanzierungsgeschichten erzeugt. Der Eindruck ist der eines raschen Fortschritts. Doch die öffentlichen Beweise, die zur Beurteilung dieses Fortschritts nötig sind, haben nicht Schritt gehalten.

A field can be busy without becoming accountable. It can generate news without generating knowledge.

Drei Arten von Evidenz

Der Index unterteilt Beweise in drei breite Klassen.

- Klasse A ist unabhängiger Beweis. Er stammt von einem Dritten und umfasst Messungen, die überprüft werden können: Genauigkeitsergebnisse, Fehlerraten, Verständnisstudien, Vergleiche mit anderen Systemen oder Leistung auf einem veröffentlichten Benchmark.
- Klasse B ist Beweis, der von dem Unternehmen oder der Organisation erzeugt wird, die für das System verantwortlich ist. Er kann nützliche technische Details enthalten, aber die Organisation, die die Behauptung aufstellt, ist auch die Organisation, die bewertet wird.
- Klasse C besteht aus Behauptungen: Aussagen in Ankündigungen, Marketingmaterial, Interviews, Präsentationen oder Presseberichten, die nicht unabhängig überprüft werden können.

Das bedeutet nicht, dass jede Behauptung der Klasse B oder C falsch ist. Ein Unternehmen kann sein Produkt ehrlich beschreiben. Ein Entwicklerteam kann ein reales

Ergebnis melden. Eine Demonstration kann echten technischen Fortschritt zeigen. Das Problem ist, dass nichts davon die zentrale Frage beantwortet: Wie schneidet das System ab, wenn jemand ohne Interesse an seinem Erfolg es misst?

Im Jahr 2026 hing fast jedes öffentliche Urteil über KI und Gebärdensprache immer noch von Material der Klasse B oder C ab. Die Unternehmen lieferten die Behauptungen, die Sprache und oft auch die Maßstäbe, mit denen diese Behauptungen verstanden wurden. Unabhängige Bewertung war weitgehend abwesend.

Das SignGemma-Problem

Googles SignGemma bietet das klarste Beispiel. Als es angekündigt wurde, erregte das Modell Aufmerksamkeit, weil es den Anschein erweckte, die Ressourcen eines großen Technologieunternehmens hinter Gebärdensprach-KI zu stellen. Damit verbundene Behauptungen umfassten ein Training mit 10.000 Stunden Daten und eine Latenz unter 200 Millisekunden. Diese Zahlen verbreiteten sich weithin, weil sie sowohl Umfang als auch Geschwindigkeit suggerierten.

Fünfzehn Monate später war das Modell immer noch nicht offiziell veröffentlicht. Google-Mitarbeiter bestätigten dies im AI Developers Forum des Unternehmens. Die Diskussion setzte sich bis Juni 2026 fort. Forscher aus Pakistan, Ägypten und anderswo baten um Zugang. Sie erhielten ihn nicht.

Dies schuf ein ungewöhnliches öffentliches Objekt: ein Modell, das diskutiert, aber nicht getestet werden konnte, zitiert, aber nicht inspiziert, erwartet, aber nicht benutzt werden konnte. Ohne Zugang zum Modell konnten Außenstehende seine berichteten Ergebnisse nicht reproduzieren. Sie konnten nicht untersuchen, wie seine Trainingsdaten zusammengestellt worden waren, seine Leistung über verschiedene Gebärdensprachen hinweg testen oder herausfinden, wie es sich außerhalb kontrollierter Beispiele verhielt. Die Behauptungen über Trainingsumfang und Latenz blieben daher Klasse C.

Googles Reputation ändert nichts an ihrem Status. Autorität mag eine Behauptung überzeugender machen, aber sie macht die Behauptung nicht unabhängig. Wiederholung verwandelt eine Behauptung nicht in eine Messung. Der Punkt ist nicht, dass SignGemma gescheitert sein muss. Der Punkt ist, dass der Öffentlichkeit kein zuverlässiger Weg gegeben wurde, zu wissen, ob es erfolgreich war.

Genauigkeit ist keine einzelne Zahl

Unabhängiges Testen ist in der Gebärdensprachtechnologie besonders wichtig, weil das Wort „Genauigkeit“ ebenso viel verschleiern kann, wie es offenbart. Ein System mag bei einem begrenzten Vokabular unter stabilen Lichtverhältnissen gut abschneiden, aber mit natürlicher Konversation kämpfen. Es mag Gebärden von Menschen erkennen, deren Erscheinung den Trainingsdaten ähnelt, während es bei anderen schlecht abschneidet. Es mag einzelne Gebärden korrekt identifizieren, aber die Bedeutung über einen Satz hinweg verlieren. Es mag für eine Gebärdensprache funktionieren und dennoch beworben werden, als funktioniere es für Gebärdensprache im Allgemeinen.

Gebärdensprachen sind keine kodierten Versionen gesprochener Sprachen. Sie haben ihre eigene Grammatik, regionale Variation und kulturellen Kontext. Bedeutung kann von Bewegung, Position, Zeitpunkt, Gesichtsausdruck und der Beziehung zwischen Gebärden abhängen. Ein System, das Hände erkennt, aber das Gesicht verpasst, mag Wörter erzeugen, während es den Satz verliert. Selbst eine hohe Genauigkeitsrate in der Schlagzeile würde uns wenig sagen, wenn wir nicht wüssten, was getestet wurde, wer teilnahm und was als korrektes Ergebnis galt.

Wurde das System an isolierten Gebärden oder an fortlaufender Konversation getestet? Waren die Teilnehmer dem Modell bekannt? Beurteilten gehörlose Gebärdennutzer, ob die Ausgabe Sinn ergab? Wurden Fehler lediglich gezählt, oder wurden sie auf möglichen Schaden untersucht? Umfasste der Test unterschiedliche Altersgruppen, Hautfarben, Gebärdenstile und regionale Varianten? Dies sind keine nebensächlichen Details. Sie bestimmen, was die Zahl bedeutet.

Deshalb sind Verständnisstudien neben technischen Benchmarks wichtig. Ein System kann eine beeindruckende Punktzahl erreichen und dabei trotzdem eine Ausgabe erzeugen, die Nutzer verwirrend, unnatürlich oder irreführend finden. Der letzte Test ist nicht, ob ein Modell Bewegung erkennt. Er ist, ob Menschen verstehen und sich auf das verlassen können, was es vermittelt.

Der Preis der Vorwegnahme

Das wiederholte Versprechen, dass ein wirksames System kurz vor der Ankunft steht, hat Konsequenzen. Ein unveröffentlichtes Modell kann Raum einnehmen, den funktionierende Dienste sonst füllen könnten. Es kann Finanzierungsentscheidungen prägen, politische Diskussionen beeinflussen und Institutionen ermutigen, Investitionen in menschliche Bereitstellung aufzuschieben. Es kann aktuelle Mängel vorübergehend

erscheinen lassen, selbst wenn kein Zeitplan oder öffentlicher Beweis diese Annahme unterstützt.

Dies ist die Politik der Erwartung. Hilfe kommt immer, also wird das Fehlen von Hilfe in der Gegenwart als weniger dringend behandelt. Für gehörlose Nutzer sind die Kosten praktisch. Ein Krankenhaus, eine Universität, ein Arbeitgeber oder eine Behörde kann auf aufkommende Technologie als Beweis dafür verweisen, dass sich die Barrierefreiheit verbessert. Aber eine vielversprechende Ankündigung kann keine medizinische Konsultation dolmetschen. Eine Demonstration kann keinen Zugang zu einem Klassenzimmer garantieren. Ein Modell, das Forschern nicht zur Verfügung steht, steht auch den Menschen nicht zur Verfügung, deren Leben zur Rechtfertigung seiner Bedeutung herangezogen wird.

Es gibt weitere Kosten. Wenn sich öffentliche Aufmerksamkeit um Systeme sammelt, die nicht unabhängig getestet wurden, können kleinere Organisationen, die weniger dramatische, aber zuverlässigere Formen der Unterstützung bieten, aus dem Blickfeld verschwinden. Menschliche Dolmetscher, gemeinschaftsgeführte Dienste und etablierte Barrierefreiheitspraktiken erhalten selten dieselbe Begeisterung wie ein neues KI-Modell. Sie werden als gegenwärtige Ausgaben beurteilt, während Technologie als zukünftige Möglichkeit geschätzt wird. Der Vergleich ist ungleich. Bestehende Angebote müssen sich jetzt für ihre Einschränkungen verantworten. Das versprochene System darf perfekt bleiben, weil es noch nicht angekommen ist.

Was Transparenz erfordern würde

Das Fehlen unabhängiger Beweise ist im Prinzip nicht schwer zu beheben. Entwickler könnten qualifizierten Forschern Zugang gewähren, bevor sie umfassende Leistungsbehauptungen aufstellen. Bewertungen könnten mit gehörlosen Gebärdennutzern gestaltet werden, anstatt lediglich an ihnen durchgeführt zu werden. Testsätze, Methoden und Definitionen könnten veröffentlicht werden. Ergebnisse könnten nach Sprache, Kontext und Nutzergruppe aufgeschlüsselt werden, statt in eine einzige Schlagzeilenzahl komprimiert zu werden.

Forscher sollten außerdem in der Lage sein, Fehlschläge zu melden, ohne von der Erlaubnis des bewerteten Unternehmens abhängig zu sein. Nichts davon würde die Unsicherheit beseitigen. Unabhängige Studien können schlecht konzipiert sein. Benchmarks können enge Formen der Leistung belohnen. Die veröffentlichten Ergebnisse können veralten, wenn sich Systeme ändern, doch unvollkommene Prüfung ist besser

als gesteuerte Sichtbarkeit. Ein öffentlicher Benchmark gibt anderen etwas, das sie hinterfragen, wiederholen oder verbessern können. Eine unbelegte Behauptung gibt ihnen nur etwas, das sie verbreiten können.

Transparenz würde auch bedeuten, klar darüber zu sein, was ein System nicht leisten kann. Ein Produkt, das für ein eingeschränktes Umfeld entwickelt wurde, sollte nicht als allgemeine Lösung bezeichnet werden. Ein Modell, das an einer einzigen Gebärdensprache getestet wurde, sollte sich nicht die scheinbare Universalität des Wortes „Gebärde“ zunutze machen dürfen. Ein Prototyp sollte als Prototyp gekennzeichnet werden. Ein nicht veröffentlichtes Modell sollte nicht als öffentliche Infrastruktur behandelt werden.

Der Fünfte Parameter

Technische Systeme werden oft nach vertrauten Maßstäben beurteilt: Geschwindigkeit, Skalierung, Kosten und Genauigkeit. Der fehlende Maßstab ist Vertrauen. Vertrauen ist kein weiterer Leistungsanspruch. Es ist das Ergebnis, Behauptungen überprüfbar zu machen. Es wächst, wenn externe Forscher ein System untersuchen können, wenn Nutzer dessen Fehler beschreiben können und wenn Evidenz auch dann noch verfügbar bleibt, wenn der Ankündigungszyklus längst weitergezogen ist. Das ist ***der fünfte Parameter*.**

Nach diesem Maßstab trat das Feld der KI-Gebärdensprache im August 2026 mit einem erheblichen Defizit an. Es hatte Produkte, Versprechen und Publicity. Was ihm fehlte, war öffentlich verfügbare, unabhängige Evidenz für die Systeme, die die meiste Aufmerksamkeit erregten.

Dieses Fehlen sollte nicht mit einer vorübergehenden Lücke in der Dokumentation verwechselt werden. Es ist Teil des gegenwärtigen Zustands der Technologie. Solange unabhängige Prüfer diese Systeme nicht messen können, lautet die ehrliche Antwort auf viele Fragen zu ihrer Leistung nicht, dass sie funktionieren, oder dass sie versagen.

Sie lautet, dass der Öffentlichkeit nicht genug Evidenz gegeben wurde, um es zu wissen.

Interessenerklärung Die Autorin ist Gründerin der Novara Consulting Group LLC, die den in diesem Artikel erwähnten Sign Language Access Trust Index entwickelt und lizenziert, und ist Herausgeberin von Der Fünfte Parameter.

Finanzierung Keine. Intern erstellt von Novara Consulting Group.

Daten und Materialien Der Artikel stützt sich auf die öffentlich zugänglichen Aufzeichnungen zu den genannten Systemen sowie auf den Thread des Google AI Developers Forum zu SignGemma.

Offenlegung der KI-Nutzung Der Text dieses Artikels stammt von der Autorin selbst. KI-Unterstützung wurde nur für den Satz in die Markenvorlage verwendet; kein Argument, keine Abbildung, kein Zitat und keine Quellenangabe wurde vom Modell erzeugt.

Empfohlene Zitierweise Grizzle, H. M. (2026). The evidence that wasn't there. Der Fünfte Parameter. Bd. 1, Ausgabe 13. Novara Press.

Zitierweise

Grizzle, H. M. (2026, August 4). Der Beweis, der es nie gab und immer noch nicht gibt. Novara Consulting Group. https://www.novaracg.com/de/2026/08/04/the-evidence-that-wasnt-and-still-isnt-there/