The Evidence That Wasn’t And Still Isn’t There

Heather M. Grizzle · August 4, 2026

An investigation into the missing independent evidence behind 2026’s biggest AI sign-language claims, from SignGemma to commercial translation tools.

THE FIFTH PARAMETER

Novara Consulting Group · Novara Press · Journal of Governance and Evidence in Accessibility AI

The Fifth Parameter, Vol. 1 · Issue no 13

AUTHOR Heather M. Grizzle

ORCID 0009-0005-2605-8256

EVIDENCE DATE Current to August 2026

KEYWORDS sign language AI; independent evaluation; evidence classes; SignGemma; transparency; trust; Deaf access

The most important fact about AI and sign language in 2026 is not what happened.

It is what did not happen.

Between January and August, companies announced products, demonstrated systems, raised money and spoke about the future of communication. New tools were presented as faster, more capable and closer to everyday use. Some attracted enthusiastic headlines. Others arrived with impressive claims about training data, response times or the number of signs they could recognise.

What did not appear was independent evidence that these systems worked as advertised. No third party published a public measurement of accuracy, error rates or user comprehension for any of the main systems examined by the Index. Searches of the required sources found no independent benchmark for Signapse, Sorenson AST, Rylo Sign, Sign-Speak, ChatSign, SignVrse, Talksign or Google’s SignGemma.

The formal wording of the finding is careful: no Class A evidence was “located as of August 2026”. This leaves

open the possibility that some evidence exists somewhere outside the public record. It may be sitting inside a company, held by a customer or contained in research that has not yet been published.

But evidence that cannot be examined cannot support a public claim. This distinction matters because the industry has not been quiet. It has produced a steady flow of announcements, partnerships, demonstrations and funding stories. The impression is one of rapid progress. Yet the public evidence needed to judge that progress has not kept pace.

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

Three kinds of evidence

The Index separates evidence into three broad classes.

- Class A is independent evidence. It comes from a third party and includes measurements that can be examined: accuracy results, error rates, comprehension studies, comparisons with other

systems or performance on a published benchmark.

- Class B is evidence produced by the company or organisation responsible for the system. It may contain useful technical detail, but the organisation making the claim is also the organisation being assessed.
- Class C consists of assertions: statements in announcements, marketing material, interviews, presentations or press coverage that cannot be independently checked.

This does not mean every Class B or Class C claim is false. A company may describe its product honestly. An engineering team may report a real result. A demonstration may show genuine technical progress. The problem is that none of these things answers the central question: how does the system perform when someone without an interest in its success measures it?

In 2026, almost every public judgement about AI and sign language still depended on Class B or Class C material. The

companies supplied the claims, the language and, often, the measures by which those claims were understood. Independent evaluation was largely absent.

The SignGemma problem

Google’s SignGemma offers the clearest example. When it was announced, the model attracted attention because it appeared to place the resources of a major technology company behind sign-language AI. Claims associated with it included training on 10,000 hours of data and latency below 200 milliseconds. These figures travelled widely because they suggested both scale and speed.

Fifteen months later, the model had still not been officially released. Google staff confirmed this on the company’s AI Developers Forum. The discussion continued through June 2026. Researchers from Pakistan, Egypt and elsewhere asked for access. They did not receive it.

This created an unusual public object: a model that could be discussed but not tested, cited but not inspected, anticipated but not used. Without access to the model, outsiders could not reproduce its reported results. They

could not examine how its training data had been assembled, test its performance across different sign languages or discover how it behaved outside controlled examples. The claims about training scale and latency therefore remained Class C.

Google’s reputation does not alter their status. Authority may make a claim more persuasive, but it does not make the claim independent. Repetition does not turn an assertion into a measurement. The point is not that SignGemma must have failed. The point is that the public was given no reliable way to know whether it had succeeded.

Accuracy is not a single number

Independent testing is particularly important in sign-language technology because the word “accuracy” can conceal as much as it reveals. A system may perform well on a limited vocabulary recorded in stable lighting but struggle with natural conversation. It may recognise signs from people whose appearance resembles those in its training data while performing poorly for others. It may

identify individual signs correctly but lose meaning across a sentence. It may work for one sign language and be promoted as though it works for sign language in general.

Sign languages are not encoded versions of spoken languages. They have their own grammar, regional variation and cultural context. Meaning can depend on movement, location, timing, facial expression and the relationship between signs. A system that recognises hands but misses the face may produce words while losing the sentence. Even a high headline accuracy rate would tell us little unless we knew what had been tested, who had taken part and what counted as a correct result.

Was the system tested on isolated signs or continuous conversation? Were the participants familiar to the model? Did Deaf signers judge whether the output made sense? Were errors merely counted, or were they examined for possible harm? Did the test include different ages, skin tones, signing styles and regional varieties? These are not secondary details. They determine what the number means.

That is why comprehension studies matter alongside technical benchmarks. A system can achieve an impressive score while still producing output that users find confusing, unnatural or misleading. The final test is not whether a model detects movement. It is whether people can understand and rely on what it communicates.

The cost of anticipation

The repeated promise that an effective system is about to arrive has consequences. An unreleased model can occupy space that working services might otherwise fill. It can shape funding decisions, influence policy discussions and encourage institutions to postpone investment in human provision. It can make current shortcomings appear temporary, even when no timetable or public evidence supports that belief.

This is the politics of anticipation. Help is always coming, so the absence of help in the present is treated as less urgent. For Deaf users, the cost is practical. A hospital, university, employer or public authority may point to emerging technology as evidence that accessibility is

improving. But a promising announcement cannot interpret a medical consultation. A demonstration cannot guarantee access to a classroom. A model unavailable to researchers is also unavailable to the people whose lives are used to justify its importance.

There is another cost. When public attention gathers around systems that have not been independently tested, smaller organisations offering less dramatic but more reliable forms of support can disappear from view. Human interpreters, community-led services and established accessibility practices rarely receive the same excitement as a new AI model. They are judged as present expenses, while technology is valued as future possibility. The comparison is unequal. Existing provision must answer for its limitations now. The promised system is allowed to remain perfect because it has not yet arrived.

What transparency would require

The absence of independent evidence is not difficult to correct in principle. Developers could provide qualified researchers with access before making broad performance

claims. Evaluations could be designed with Deaf signers rather than merely conducted on them. Test sets, methods and definitions could be published. Results could be broken down by language, context and user group instead of compressed into a single headline figure.

Researchers should also be able to report failures without depending on the permission of the company being assessed. None of this would remove uncertainty. Independent studies can be badly designed. Benchmarks can reward narrow forms of performance. The published results can become outdated as systems change, yet imperfect scrutiny is better than managed visibility. A public benchmark gives others something to question, repeat or improve. An unsupported claim gives them only something to circulate.

Transparency would also mean being clear about what a system cannot do. A product designed for a restricted setting should not be described as a general solution. A model tested on one sign language should not be allowed to borrow the apparent universality of the word “sign”. A prototype should be identified as a prototype. An

unreleased model should not be treated as public infrastructure.

The Fifth Parameter

Technical systems are often judged by familiar measures: speed, scale, cost and accuracy. The missing measure is trust. Trust is not another performance claim. It is the result of making claims testable. It grows when outside researchers can inspect a system, when users can describe its failures and when evidence remains available after the announcement cycle has moved on. This is ***the fifth parameter.***

By that measure, the AI sign-language field entered August 2026 with a serious deficit. It had products, promises and publicity. What it did not have was publicly available independent evidence for the systems attracting the most attention.

That absence should not be mistaken for a temporary gap in paperwork. It is part of the technology’s present condition. Until independent evaluators can measure these systems, the honest answer to many questions about their

performance is not that they work, or that they fail.

It is that the public has not been given enough evidence to know.

Declaration of interests The author is founder of Novara Consulting Group LLC, which builds and licenses the Sign Language Access Trust Index referenced in this article, and edits The Fifth Parameter.

Funding None. Produced internally by Novara Consulting Group.

Data and materials The article draws on the public record for the systems named and on the Google AI Developers Forum thread concerning SignGemma.

AI use disclosure The text of this article is the author’s own. AI assistance was used only for typesetting into the branded master; no argument, figure, quotation or citation was generated by the model.

Suggested citation Grizzle, H. M. (2026). The evidence that wasn’t there. The Fifth Parameter. Vol 1 Issue 13.

Novara Press.