The Coordinates of Trust

Heather Grizzle · August 23, 2026

A governance analysis of Google DeepMind's SL2T sign-language-to-text model, examining its architecture, evidence, known limitations, human verification safeguards, prohibited uses, and unresolved questions through the five trust domains of the SLAT Index.

NOVARA CONSULTING GROUP · NCG SPECIAL ARTICLE

The Coordinates of Trust

A governance reading of Google DeepMind's SL2T

Heather M. Grizzle · Novara Consulting Group · 22 August 2026

On 12 August 2026, Google DeepMind released SL2T, a sign-language-to-text translation model, as a shipping feature of Gboard and Live Transcribe on the Pixel 11. This article examines the technology on the published record and reads it through the five domains of the Sign Language Access Trust (SLAT) Index. It is an analysis, not an

assessment: it applies no scoring instrument, assigns no rating, and states its questions rather than its conclusions. It relies exclusively on the public materials listed in the references.

A release in two documents

Google DeepMind's announcement of SL2T arrived as a pair of artifacts, and the pairing is the most instructive thing about it. The first is a product announcement of a familiar kind: a blog post introducing a massively multilingual sign-language-to-text model, trained on more than 100,000 hours of signing across more than fifty sign languages, roughly a quarter of it in American Sign Language, and shipping as a dictation feature in Gboard and Live Transcribe on the Pixel 11, ASL to English first, at no additional cost. The second is less familiar. Linked from the announcement itself is a Joint Impact Report co-authored by the DeepMind team and named representatives of the National Association of the Deaf, the World Federation of the Deaf, the Deaf Professional Arts

Network, and Rochester Institute of Technology's National Technical Institute for the Deaf, together with an independent subject matter expert. The report describes the model's architecture and benchmark results, records the advisory committee's findings in ranked priority bands, discloses known failure modes with worked examples, and draws an explicit boundary around where the technology may and may not be used. A vendor publishing its capability claims and its governance claims in the same motion is rare in this market, and both documents repay close reading.

The company describes the release as bringing sign language AI out of the lab and into consumer products for the first time. That claim wants a qualifier. Consumer sign language AI has existed for years, mostly in the other direction: avatar applications and website services that render text or speech into signed output have been in consumer and institutional deployment for over a decade, including from Deaf-founded companies. What is genuinely new here is narrower and, stated precisely, still substantial. SL2T appears to be the first sign language understanding

model shipped as a standard input method in a mainstream phone's keyboard and accessibility features, by a company whose mobile platform reaches billions of devices. Precision on this point is not pedantry; the arrival of the field's largest entrant is exactly the moment to be exact about what changed. One further clarification is worth recording, because the two are already being conflated in coverage: SL2T is not SignGemma, the sign language model DeepMind previewed in 2025. The company has confirmed they are separate models.

A note on standing. Novara Consulting Group is independent of Google and of every company named in this article, holds no financial interest in any sign language AI vendor, and had no contact with the developer in preparing this analysis. This article is also not a SLAT evaluation. The Index's written criteria, evidence rules, and rating scale are applied only within formal engagements, and nothing here should be read as a rating of SL2T. What follows is the anatomy of the release, and the questions it puts to any institution that will eventually be asked to answer for a deployment.

What the machine sees

SL2T does not watch video, at least not in the way the phrase suggests. An on-device computer vision model, MediaPipe Holistic, reduces each frame of the camera feed to two-dimensional coordinates for 130 landmark points across the signer's face, body, and hands. The video is then discarded on the device. Only the coordinate sequence travels to Google's servers, where a transformer-based model translates it directly into English text, generated word by word as the signer signs. According to the report, no logs of user inputs or system outputs are retained unless the user explicitly authorizes retention in the context of an evaluation study.

The architectural decision with the deepest linguistic significance is what the pipeline omits. Prior sign language translation systems have generally routed through glosses, written labels standing in for individual signs, either as an output or as an intermediate training target. Glossing has been the field's bottleneck and its distortion: it caps vocabulary at what has been annotated, and it flattens

exactly the parts of a signed language that written labels cannot carry, the grammar performed on the face, the use of space, the classifier constructions that depict rather than name. DeepMind's stated position is that glosses fail to capture these features and that translating directly from landmarks removes artificial vocabulary limits and lets quality scale with data. That position is linguistically well grounded, and it is notable to see a major developer state it plainly, because a large share of this market's historical failures trace to treating signed languages as manual codes for spoken ones.

The report is unusually candid about what the landmark representation gives up. MediaPipe's face model does not track the tongue, which the report itself identifies as a critical articulator in ASL. The representation carries nothing of the signer's environment, so signs that point at or depict real-world referents lose their targets. It lacks depth ordering, so the difference between contact and hovering, which can be lexically significant, is difficult to perceive. Input clips are capped at sixty seconds and evaluated independently, with no memory across turns and

no awareness of the conversation partner's side of an exchange. The tracker follows the largest subject in the frame and does not handle multiple people. And because the landmark model expects a standard hand, it may misestimate or hallucinate keypoints for signers with limb differences. Each of these limits is stated in the vendor's own document, which is to the document's credit; each is also a fact about who the system will serve well and who it will not.

The privacy architecture deserves the same double reading. Discarding video at the device and transmitting only coordinates is a genuine safeguard: appearance, environment, and identifiable imagery never reach the server, and nothing is retained to leak or be compelled. But the abstraction should not be mistaken for anonymization of content. The report notes, in passing, that speech produced after signing may still be erroneously transcribed through lipreading, which is to say the coordinate stream retains enough facial articulation to recover spoken words. The stream is stripped of what a signer looks like, not of what the face and body are saying; that is precisely what

makes it useful for translation. A careful procurement reading therefore examines not what the discarded video would have shown but what the retained coordinates can reveal, and to whom, under what terms.

One property of SL2T frames everything else about it: its direction. This is a comprehension system, not a generation system. The machine reads the Deaf person, and someone else, a hearing conversation partner, an application, eventually perhaps an institution, reads the machine. The two directions carry different risks. When an avatar signs at a Deaf viewer, the harm of error is failed access; the viewer does not receive the message. When a model transcribes a Deaf signer, the harm of error is misattribution; words the person did not produce enter the written record over their name. The first direction fails its audience. The second puts statements in a person's mouth. SL2T's release design, as the following sections describe, is built around keeping that second risk in the signer's own hands. The governance question is what happens when it leaves them.

The numbers, read carefully

The report publishes a compact benchmark table. On FLEURS-ASL, a translation benchmark built from complicated, abstract source text rendered into ASL by Certified Deaf Interpreters under studio conditions, SL2T scores 25 BLEU and 70 BLEURT zero-shot, meaning the model was not trained on the set. On a one-handed variant of the same benchmark it scores 32 BLEU and 74 BLEURT. On PRESTO-ASL, a corpus of assistant-style phrases signed to a docked tablet, it reaches 67 BLEU and 85 BLEURT, and on FSboard, a fingerspelling dataset collected on mobile devices, 64 percent exact-match accuracy. The announcement leads with the 70 BLEURT figure and presents it as well above any previously published result.

Two things should be said about these numbers, in order. The first is that they are legitimate in a way accuracy claims in this field often are not. Because SL2T's output is English text, reference-based machine translation metrics genuinely apply here; scoring generated signing with text-

overlap metrics is a category error, but that is not what is happening in this direction. The second is what the numbers cannot carry. BLEU counts n-gram overlap against reference translations; BLEURT is a learned metric trained to predict human quality judgments. The same test set yields 25 on one scale and 70 on the other, and both figures are accurate, which is a useful reminder that they measure different proxies on different scales and that neither is a comprehension guarantee. The report's own worked examples show the residual failure profile: fluent, well-formed English arriving with propositional errors inside it, a fingerspelled word misread as another word, a described feature silently dropped, past tense drifting into present. That is exactly the class of error reference-based metrics under-punish, output that reads well and says something else, and it matters more, not less, in a system whose output will be received as what a person said.

Then there is the question of who is holding the yardstick. FLEURS-ASL, the headline benchmark, was created by a researcher who is listed among the core SL2T development team. FSboard likewise originates with the team and its

collaborators, and every figure in the table was produced by the developer measuring its own system. None of this is improper, and building evaluation infrastructure for a field that lacked it is a contribution in its own right. It is simply not independent, and the distinction becomes material in light of a second disclosure: the builds evaluated hands-on by the advisory committee in July 2026 were early pre-release checkpoints, and the report states that substantial training and post-processing changes were made between those builds and the launch candidate, with the resulting improvements attested by internal metrics. The strongest community-facing evaluation in the record therefore attaches to a version that did not ship. On the published record as of this writing, no evaluation of the deployed feature's output quality has been conducted and published outside the developer's own process. That sentence describes the state of the evidence, not the quality of the model, and the difference between those two things is the subject of this article.

The boundary

The Joint Impact Report is where this release departs most sharply from the field's habits. It is co-authored, with named individuals from named organizations. Committee engagement is managed by two facilitators from Google's Accessibility organization who are described as operating independently of the research and engineering teams. Feedback from testing was triaged into priority bands under an AISLAC Governance Charter, and the top of that triage is not a model defect at all. The committee's highest-ranked critical issue, designated P0-1, is the risk that public agencies, healthcare providers, and employers will substitute automated sign-to-text for qualified human interpreters because software is cheaper and more convenient than people.

Against that risk, the report draws its boundary in unusually concrete terms. The supported envelope is deliberately modest: messaging, dictation, search queries, assistant commands, and brief, elective, one-to-one exchanges in low-stakes settings, ordering at a coffee shop or asking for help at a retail counter. Outside it, the report names prohibited contexts by sector and by scenario:

healthcare, including clinical consultations, triage, emergency medical services, mental health counseling, and informed consent procedures; legal and law enforcement settings, from police interrogations to courtroom proceedings, depositions, attorney-client consultations, and sworn statements; academic evaluation and classroom instruction, including IEP meetings and the replacement of classroom interpreters; employment decisions, including job interviews, disciplinary hearings, performance reviews, and formal grievances; and high-stakes government interactions such as administrative hearings and benefits determinations. The report then states the thing that matters most, jointly and in terms: that SL2T does not satisfy legal obligations for reasonable accommodation under the Americans with Disabilities Act, Sections 504 and 508, or international disability frameworks that are met by human interpreters, and that third parties must not use it to evade those obligations.

Credit should be given precisely here. Ranking institutional misuse above every technical defect, and publishing a sector-by-sector prohibition alongside a plain statement

about accommodation law, is boundary drawing of a kind this market has rarely produced. It aligns the vendor's public position with what Deaf advocacy organizations and accessibility law have maintained throughout: a draft-generation tool is not an interpreter, and procuring it as one is a governance failure before it is a technical one. That the statement is made in the vendor's own voice, jointly with the community's own organizations, gives it a standing that outside commentary cannot supply.

The governance reading begins where the credit ends, with the mitigation column. The recorded mitigation for P0-1 is the publication of the joint policy guidance itself. A published boundary is an act of description, not of control. Within the release materials, nothing describes a technical, contractual, or monitoring mechanism by which a clinic, school district, or employer that points a phone at a Deaf person in a prohibited setting would be prevented, detected, or held answerable, and nothing states who would learn of such a use or what would follow from it. Mechanisms of that kind may exist, or come to exist, in product terms and enforcement policy; the released

documents do not say. The report itself points at the pressure this boundary will come under. It names programmatic access through Google Cloud APIs, and possible open-weights releases in the vein of SignGemma, as directions under exploration, and it acknowledges that wider availability raises governance problems it intends to work through with the committee rather than decide unilaterally. That acknowledgment is the right one. It is also an admission of where things stand: once the same capability is available to integrators through an API, the audience for the boundary changes from end users to institutions, and a statement in a PDF does even less of the work.

Where the safeguard lives

Inside the supported envelope, the release's load-bearing safeguard is a design principle the report states explicitly: the signer stays in the loop as verifier. Model output is a draft. The user reviews and edits the generated English before sending it or showing it, and in Live Transcribe nothing is displayed to the conversation partner until the

user elects to release it. As a control for a self-directed input tool, this is well chosen. The person with the most context, and the most at stake, holds the veto.

The report is equally explicit about the safeguard's precondition, and the passage may be the most important one in it. The human-in-the-loop design presumes that the user has functional English literacy sufficient to verify the generated text and enough device proficiency to operate the editing controls, and the report states that where those conditions are not met, translation errors may pass unnoticed. The same document elsewhere notes the wide diversity among deaf people in signing, speaking, reading, and writing proficiency. Put together: the verification burden falls on the user, in the user's second language, and it falls hardest on precisely the users for whom a sign-first input method matters most. The safeguard, in other words, is a user capability rather than a system property. Stating that precondition in the release document is honest, and the honesty should be acknowledged. It is also the beginning of the accessibility analysis, not the end of it.

The deeper structural point follows from direction. The verifier design is coherent because this release confines the tool to use by the signer, on the signer's own words, at the signer's election. Every prohibited scenario in the report shares one feature: there, the tool would be used on someone rather than by them, its output read by an institution as the record of what the person said. In that configuration the verifier and the verified trade places. The party relying on the text is the one least able to check it against the signing, and the party able to check it is the one whose words are at stake and whose objection may or may not be credited. The report's scope boundary and its design principle are the same commitment stated twice, and both depend on the confinement holding.

Even inside the envelope, the disclosed failure modes reward attention. The committee's second critical finding is numeric distortion; the worked example is a Maine area code, 207, rendered as the year 2007, in a tool whose core uses include addresses, times, and phone numbers. Hallucinated text can appear when a second person enters the frame or when the signer pauses mid-thought to

compose. Meanings carried primarily on the face, negation, questioning, the grammar linguists call non-manual markers, often fail to translate at all when the hands do not reinforce them. Polysemous signs and regional variants misread without disambiguating context. None of this is disqualifying for drafting a message in a coffee shop, which is the point of bounding the release as the report does. All of it becomes material the moment the output is treated as a record of what a person said.

Five questions

The Sign Language Access Trust (SLAT) Index evaluates sign language AI systems across five domains set out in its published Evaluation Standard: Linguistic Trust, Transparency Trust, Governance Trust, Accessibility Trust, and Deployment Trust. A full evaluation applies written criteria and evidence rules to a defined public record; those instruments are not reproduced here, ratings are issued only through that process, and none is issued now. What can be usefully stated is the shape of the inquiry each domain brings to this release, as questions the

published materials raise and do not yet close.

Linguistic Trust asks who has established, independently of the developer, that the output means what the signer said, and for which signers. The record shows serious internal measurement, a genuine zero-shot result, and structured community testing. It does not yet show an evaluation of the shipped model's output conducted and published outside the developer's own process, and the developer's own demographic disaggregation, across region, age, gender, and Black ASL, is described in the report as an ongoing research priority rather than a reported result.

Transparency Trust asks what the hundred thousand hours are. Neither document describes where the training data came from, under what consent and licensing terms it was gathered, whether the people recorded in it were compensated, or what share of it originates with the Deaf partners the announcement credits. The domain also asks which model version each published claim attaches to, a question the report itself makes live by distinguishing the evaluated checkpoint from the launch candidate.

Governance Trust asks where authority sits. The committee's membership is named, its findings are ranked, and its influence on the engineering roadmap is documented in the report itself, which is considerably more than assertion. What the record does not establish is the charter under which the committee operates, the terms of its independence given that it is convened, facilitated, and equipped by the company it advises, and whether its role is anywhere decisional rather than advisory. The domain also reaches the people inside the data: what those recorded were told, what they agreed to, and what they received.

Accessibility Trust asks who carries the residual risk, and the record itself identifies the candidates: signers without confident English literacy, on whom the verification burden falls; signers under eighteen, excluded from training and evaluation on legal grounds but not, on anything published, from use; signers with limb differences or atypical movement, for whom the system may fail at the perception layer; and any user harmed by an error, for whom the release materials describe no incident route and

no named point of accountability.

Deployment Trust asks what happens at the boundary. What binds a third party to the prohibited-use list; what terms will govern programmatic access if the API direction proceeds; what an institution that deploys the tool in a prohibited setting owes the people it is used on; and who verifies, over time and in public, that the boundary is holding.

These are questions, not findings. Several may have good answers that have not yet been published, and the distance between an absent practice and an undocumented one is precisely what a structured evaluation exists to measure.

What this release changes

Two judgments, offered as judgments. First, SL2T is significant on both of the axes that matter. Technically, direct landmark-to-text translation trained at this scale, shipped with attention to streaming latency, one-handed signing, and left-handed signers, moves the field's center

of gravity, and the decision to publish failure modes with worked examples sets a disclosure floor other developers should now be expected to meet. Institutionally, the Joint Impact Report is among the most substantive governance documents any sign language AI release has yet carried, and its central statement, that this technology does not satisfy interpreter-based accommodation obligations and must not be procured as though it did, is the correct statement, made where it counts.

Second, the release changes the market without changing the field's central evidence problem. Capability claims still originate with the developer and are still measured against instruments of the developer's own construction. The most community-facing evaluation on the record attaches to a build that did not ship. The boundary that makes the release defensible is, on the published record, a statement rather than a mechanism, and the roadmap the report describes, toward more languages, more devices, programmatic access, and eventually generation, points directly at the settings where statements will be least sufficient. The entry of the largest technology company yet

to work in this market makes these questions more consequential, not less, because the next institution tempted to propose an automated substitute for a qualified interpreter will cite this release as precedent, whatever the report beside it says.

Novara Consulting Group evaluates sign language AI systems under the SLAT Index; the Index's published Evaluation Standard is available at DOI 10.5281/zenodo.21537271. A full SLAT review of SL2T, applying the Index's written criteria to the complete public record, is in preparation. NCG is independent of every company named in this article and holds no financial interest in any sign language AI vendor.

References

1. Google DeepMind Sign Language Team, “Putting sign language AI into users' hands,” Google DeepMind, 12 August 2026. https://deepmind.google/blog/putting-sign-language-ai-into-users-hands/

2. Google DeepMind Sign Language Team; S. Forbes (DPAN); N. Kiego (NAD); G. Behm and J. Riggio (RIT/NTID); Y. Koraneu and J. Moore (WFD); R. Thibodeau (independent subject matter expert), “AISLAC Joint Impact Report for SL2T 1.0,” 12 August 2026. https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/putting-sign-language-ai-into-users-hands/aislac-joint-impact-report-for-sl2t-1-0.pdf
3. G. Tanzer, “FLEURS-ASL: Including American Sign Language in Massively Multilingual Multitask Evaluation,” Proceedings of NAACL 2025.
4. M. Georg, G. Tanzer, E. Uboweja, S. Hassan, M. Shengelia, S. Sepah, S. Forbes, and T. Starner, “FSboard: Over 3 Million Characters of ASL Fingerspelling Collected via Smartphones,” Proceedings of CVPR 2025.
5. L. O'Dell, “Google DeepMind unveils sign-language-to-text feature for Pixel 11,” 13 August 2026, updated with Google DeepMind's

confirmation that SignGemma and SL2T are separate models. https://liamodell.com/2026/08/13/google-deepmind-artificial-intelligence-ai-sign-language-to-text-sl2t-american-sign-language-asl-live-transcribe-gboard/

6. Novara Consulting Group, “SLAT Index Evaluation Standard,” Version 2.0, July 2026. DOI 10.5281/zenodo.21537271.
7. Google Research, “MediaPipe Holistic: Simultaneous Face, Hand and Pose Prediction, On-Device.” https://research.google/blog/mediapipe-holistic-simultaneous-face-hand-and-pose-prediction-on-device/