NOVARA
CONSULTING GROUP

Eighteen Signers on Their Own Phones: What a Real Saudi Sign Language Dataset Shows That a Formula Could Not

Heather Grizzle · November 28, 2026

A month after a paper gave Saudi Sign Language a simulated 91 percent, a real dataset of 18 signers shows what honest numbers look like.

Contents

1. What Isharah is
2. What the numbers say
3. Who signed, and who checked
4. What this means for buyers
5. The number you can use
6. References

IN SHORT

- This week AIhub interviewed Sarah Alyami, whose doctoral work produced Isharah, described as the first large-scale dataset for sentence-level Saudi Sign Language: 30,000 video clips of more than 2,000 sentences, signed by 18 people on their own smartphones.
- A month ago, a paper in Scientific Reports reported 91 percent accuracy for Saudi Sign Language without observing a single signer. Isharah is what a real number looks like: worse, and far more useful.
- On new signers, the best published model gets about 27 percent of glosses wrong. On sentences it has never seen, that rises to about 39 percent. A newer selfie-style dataset from the same group pushes the unseen-sentence figure to about 68 percent.
- The dataset is honest about its limits. What it does not yet say is who checked the language: whether the two experts who annotated it are Deaf, and why hearing translators make up a fifth to more than a third of the signers.

Last month I took apart a paper that ranked twenty sign languages, including Saudi Sign Language, and reported a maximum accuracy of 91 percent. Under the paper's own formula and its own listed attributes, Saudi Sign Language should have scored 71.5. The paper's table gave it 91, the ceiling. No model was run on any Saudi signing to get there. The number came from a points table about language family, fingerspelling hands, population and legal status, and the paper itself called its results a theoretical simulation [4].

This week, AIhub published an interview with a researcher who did the slow, unglamorous opposite. Sarah Alyami, now an assistant professor of computer science at Imam Abdulrahman Bin Faisal University, finished her PhD at King Fahd University of Petroleum and Minerals in 2025, and her dissertation received an Honourable Mention for the 2025 ACM/AAAI Doctoral Dissertation Award. At its center is Isharah, which the interview describes as the first large-scale dataset for sentence-level Saudi Sign Language, built to reflect real signing conditions rather than a studio [1].

I want to put these two pieces of work side by side, because together they show buyers the difference between a number that describes a language and a number that describes a spreadsheet.

What Isharah is

The dataset paper, now published in IEEE Transactions on Multimedia, describes 30,000 video clips covering more than 2,000 unique sentences of Saudi Sign Language [2]. Eighteen people signed them: 13 men and 5 women, aged 21 to 53. Eleven are deaf, three are hard of hearing, and four are hearing sign language translators. Every signer, the paper says, gave informed consent for their recorded clips to be used for research.

What makes Isharah different from most sign language datasets is where it was recorded. The signers filmed themselves on their own smartphones, mostly at home or in offices, without anyone supervising. Fourteen used iPhones and the rest used a mix of other brands, so the resolution and orientation vary. Two Saudi Sign Language experts wrote the glossing guidelines and annotated every clip, with Arabic translations alongside the glosses. Because there was no standard way to gloss Saudi Sign Language, the team had to build one [2].

That is real work. Most sign language AI is trained on signing captured in controlled settings: good lighting, a plain wall, a fixed camera. Real people sign on phones, in kitchens, in bad light, holding the phone in one hand. A dataset that starts there is closer to the world a product will meet.

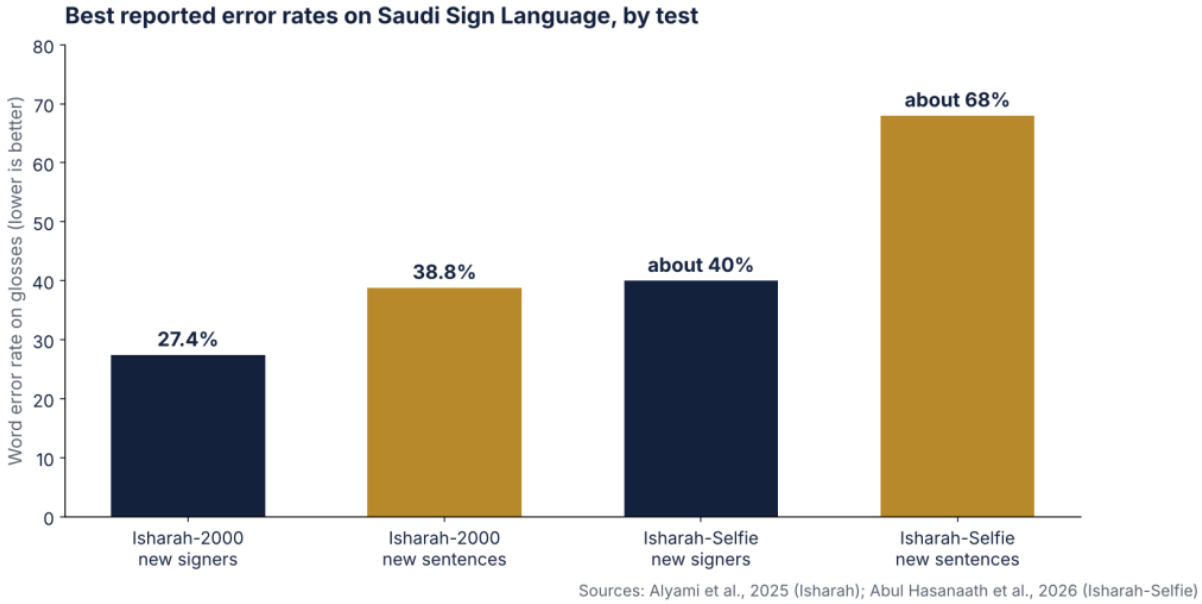
What the numbers say

The paper reports results on two kinds of test, and the difference between them is the most useful thing in it for a buyer.

The first test holds out signers. The model trains on some people and is tested on people it has never seen. On the full 2,000-sentence set, the best model in the paper has a word error rate of 27.4 percent on that test [2]. Word error rate here counts how many glosses the model gets wrong, adds or drops compared with the correct sequence. Roughly, the model gets about one gloss in four wrong for a new signer.

The second test holds out sentences. The model is tested on sentences it never saw in training, which is what happens every day in real use, since people do not limit themselves to a script. On that test, the best result rises to 38.8 percent [2]. That is closer to two glosses wrong in every five.

A newer dataset from the same research group, Isharah-Selfie, pushes further toward real life. It captures one-handed signing, the way people sign with a front-facing camera while holding the phone in the other hand: 8,997 videos, 12.5 hours, 16 signers and 1,493 sentences. On new signers, the best video-based model there has a word error rate of about 40 percent. On unseen sentences it is about 68 percent [3].



Error rises as the test gets closer to real use. Publishing the harder test is what makes these numbers trustworthy.
Sources: Alyami et al. (Isharah); Abul Hasanaath et al. (Isharah-Selfie)

None of this is a criticism. These numbers are the point. A team that publishes a harder test, on which its own models look worse, is telling you the truth about where the field is. Compare that with a 91 percent produced from a formula with no signer in it. The honest number is lower, and it is the only one you can use.

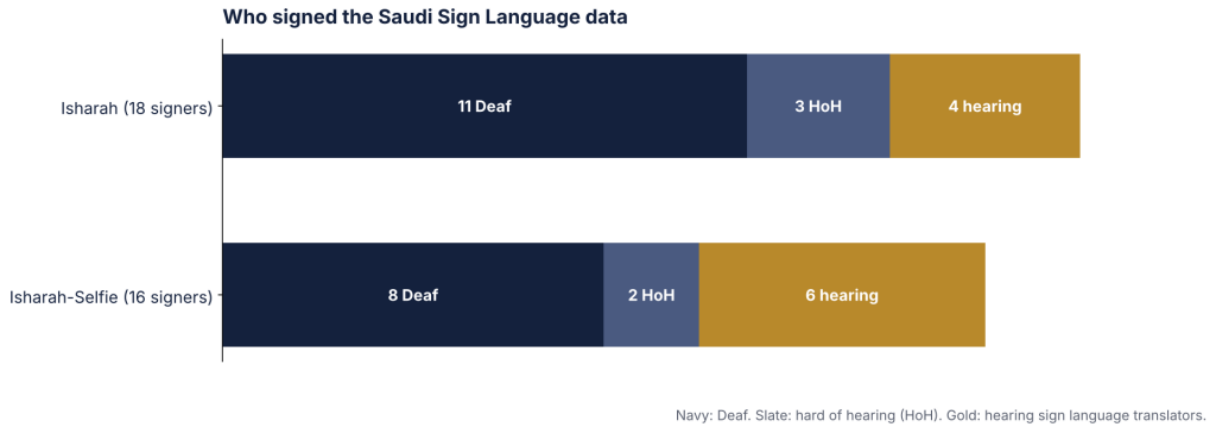
The paper is also candid about its limits. There were no child or elderly signers. Some signers departed from the prescribed sentences, creating glosses the model had never seen, up to 9.9 percent of one test set. Around a quarter of the glosses appear only

once. And the authors note that less fluent interpreters may introduce variations that hurt recognition [2]. That last point deserves more attention than it gets.

Who signed, and who checked

Here is what I would want to know before anyone builds a product on this.

Four of Isharah′s 18 signers are hearing translators, which is a little over a fifth. In Isharah-Selfie, six of 16 are hearing, which is more than a third, with eight Deaf signers and two hard of hearing [3]. Interpreted signing is not the same as Deaf people signing to each other. It tends to follow the structure of the spoken or written sentence it came from, and it can carry the habits of second-language learners. A University of Surrey project funded this year makes that its starting point, with the stated aim of training sign language AI on real conversations between Deaf people, not interpreted signing [5].



Hearing translators are a fifth of Isharah′s signers and more than a third of Isharah-Selfie′s. The papers do not say why.

The question is not whether hearing translators should ever appear in sign language data. It is why they were included, what proportion of the clips they signed, and whether the models perform differently on their signing than on Deaf signers′ signing. The papers I read do not break results down that way. For a product, that breakdown matters more than the average.

The second question is about the annotators. Two Saudi Sign Language experts wrote the glossing rules and labeled all 30,000 clips. The paper does not say whether either

of them is Deaf [2]. In a dataset, the annotation is the answer key. Every error rate in the paper is measured against it. If the people who decided what each sentence means are fluent Deaf signers, that is a strength worth stating. If they are not, it is a limitation worth stating. Either way, it should be on the page.

The AIhub interview describes the work as aiming to reduce communication barriers for Deaf and hard of hearing communities, but it does not describe how Deaf people shaped the project beyond appearing in it [1]. I do not take that as evidence that they did not. I take it as a question that the next paper could answer in a paragraph.

And there is scale. Isharah-Selfie cites a Saudi census figure of more than 84,025 Deaf people in the country [3]. Eighteen signers is a large number for a sign language dataset and a very small number for a language community. A model trained on them has seen a sliver of the regional, generational and personal variation in how Saudi Deaf people sign. The authors say they are expanding the dataset, and the interview says the same [1]. That is the right direction.

What this means for buyers

Isharah is research infrastructure, not a product, and nobody involved claims otherwise. But research datasets end up behind products, and their numbers end up in sales decks. If a vendor tells a ministry, an airline or a hospital in the Gulf that its system has been tested on Isharah, here is what to ask.

1. Which test? A result on new signers and a result on unseen sentences are very different numbers. Ask for both, and treat the unseen-sentence result as the closer guide to real use.
2. What does the error rate mean in your setting? A quarter of glosses wrong may be fine for search or practice tools. It is not fine for a medical consent conversation or an emergency announcement.
3. How does the system perform on Deaf signers alone, compared with hearing translators? If the vendor does not know, the average tells you less than it seems to.
4. Were fluent Deaf signers involved in checking the annotations the system is scored against?

5. What licence covers the data, and does it allow commercial use? The Isharah paper on arXiv carries a non-commercial, no-derivatives licence, and the dataset terms are not separately stated there [ncg_citep key="alyami2025"]. Signers consented to research use. A product is not research.

The number you can use

There is a reason I keep returning to the 91 percent. It was published in a peer-reviewed journal, it has a DOI, and a vendor could cite it tomorrow. It looks better than anything in Isharah. It is also worthless, because no one signed for it.

Isharah's numbers look worse because real Saudi signers, on real phones, in real rooms, signed for them, and because the team chose to publish the hard test as well as the easy one. That is what evidence looks like in this field: a smaller number, a clear method, a list of limits, and signers you can count. The next step, for this team and everyone else building on sign language data, is to make the Deaf people in the dataset as visible in its governance as they are in its videos.

Heather Grizzle-Odland is the founder of Novara Consulting Group, which evaluates sign language AI and other AI accessibility tools before institutions buy them. She is DeafBlind and lives in Minnesota.

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.

References

1. Scallan, E. (2026, November 25). Interview with Sarah Alyami: Continuous sign language recognition, dataset development and novel frameworks. AIhub. https://aihub.org/2026/11/25/interview-with-sarah-alyami-continuous-sign-language-recognition-dataset-development-and-novel-frameworks/
2. Alyami, S., Luqman, H., Al-Azani, S., Alowaifeer, M., Alharbi, Y., & Alonaizan, Y. (2026). Isharah: A large-scale multi-scene dataset for continuous sign language recognition. IEEE Transactions on Multimedia. https://doi.org/10.1109/TMM.2026.3664959 (preprint: https://arxiv.org/abs/2506.03615)
3. Abul Hasanaath, A., Aljubran, M., Alyami, S., Khan, M. H., & Luqman, H. (2026). Isharah-Selfie: Continuous sign language recognition dataset for one-handed signing [Preprint]. arXiv. https://arxiv.org/abs/2610.33200
4. Harish, M. N., Rashmi, B., & Srilakshmi, M. (2026). A global taxonomy and computational characterization of major sign languages for scalable AI recognition systems. Scientific Reports. Article in Press. https://doi.org/10.1038/s41598-026-65810-z
5. University of Surrey. (2026). Sign language AI to focus on real Deaf conversations, not just interpreter data [News release]. EurekAlert. https://www.eurekalert.org/news-releases/1113454

How to cite

Grizzle, H. M. (2026, November 28). Eighteen Signers on Their Own Phones: What a Real Saudi Sign Language Dataset Shows That a Formula Could Not. Novara Consulting Group. https://www.novaracg.com/2026/11/28/eighteen-signers-on-their-own-phones-what-a-real-saudi-sign-language-dataset-shows-that-a-formula-could-not/