Unlocking Speech AI Technology for Global Language Users

NVIDIA Speech Summit 2022

EM Lewis-Jong and Caroline de Brito Gottlieb

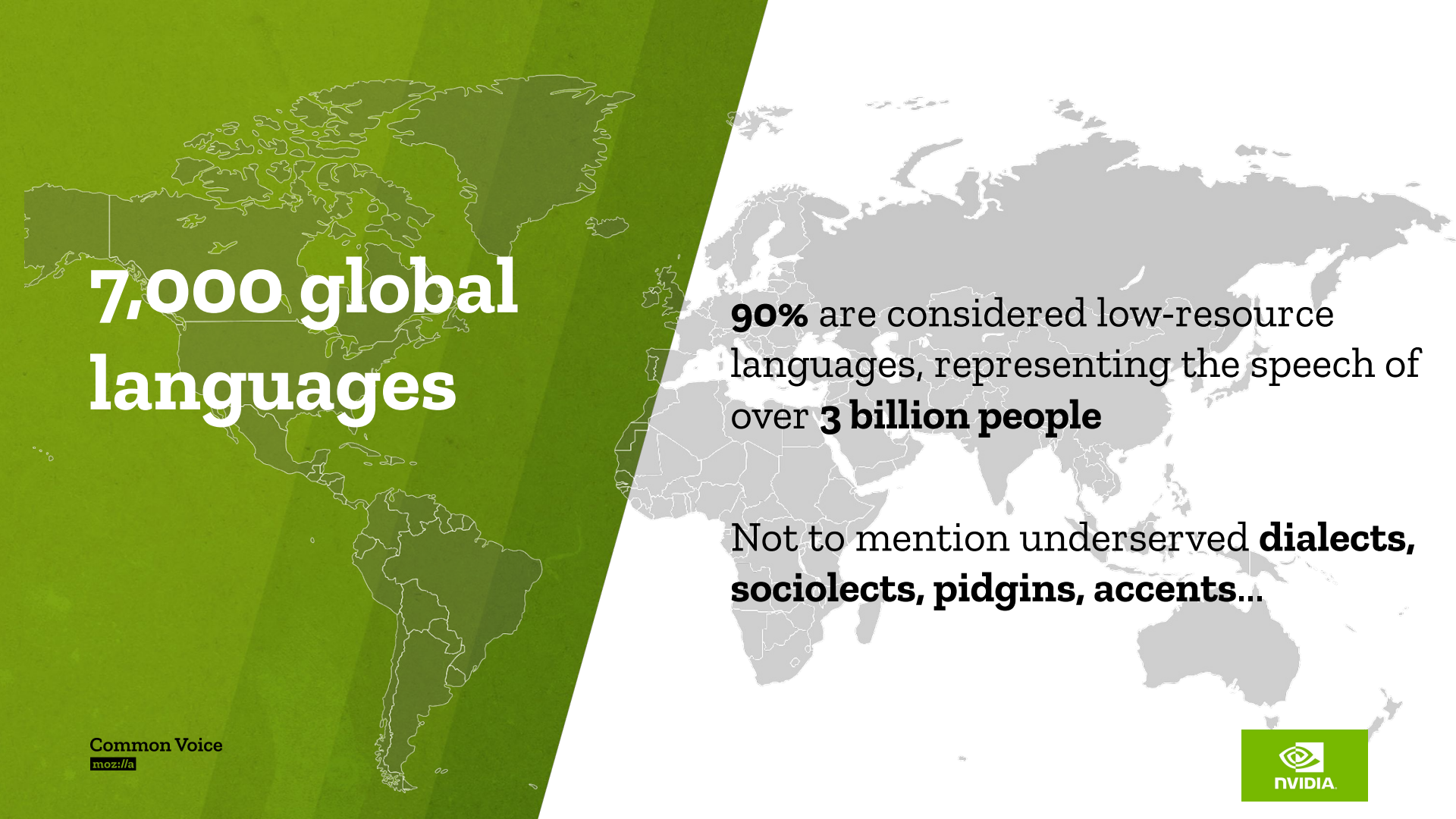
Common Voice
moz://a

THE PROBLEM:

Voice assistants like Alexa and Google Home support fewer than

1% of the world's spoken languages

NVIDIA

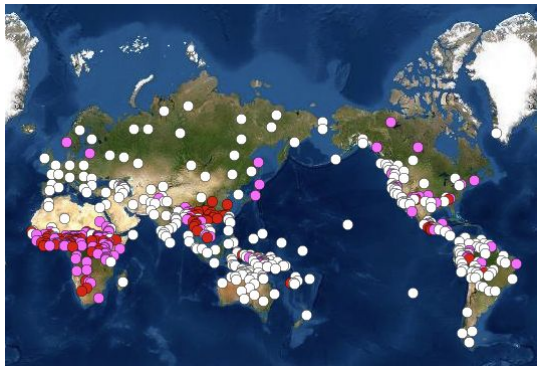

7,000 global languages
Common Voice
moz://a
90% are considered low-resource languages, representing the speech of over **3 billion people**
Not to mention underserved **dialects, sociolects, pidgins, accents...**
NVIDIA

Sentence structure

irino vakhe inagu (**Nias**)

VERB-OBJECT-SUBJECT

Tone in the world's languages



VARIATION is the natural state of language

Writing system

دستت درد نکنه (**Farsi**)

Word structure

khi tôi dên nhà ban tôi, chúng tôi bát dâu làm bài (**Vietnamese**)

Tuntussuqatarniksaitengqiggtuq (**Yup'ik Inuit**)

Common Voice

moz://a

NVIDIA

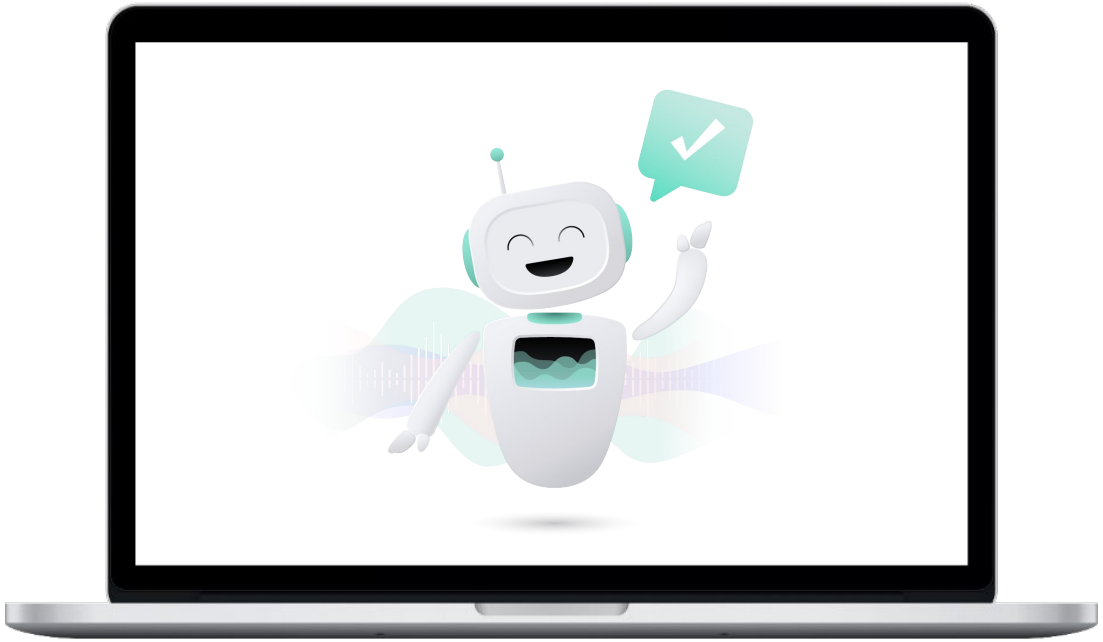
Sources: World Atlas of Language Structures (WALS); SIL International

Improving linguistic inclusion in Speech AI is one part of the solution to this

Mozilla and NVIDIA are partnering to advance this ecosystem for global language users

Common Voice:

- ★ Build, maintain and improve the **Common Voice platform** for linguistic inclusion, usability and experience
- ★ Support **communities** to join Common Voice
- ★ Support dataset health by providing **guidance, tools and resources**
- ★ Help datasets to grow through community **mobilisation**, targeted interventions
- ★ **Release the datasets** publicly so anyone can make use of them



NVIDIA:

- ★ Support **dataset health** by providing feedback
- ★ **Train and open source high quality ASR models** on Common Voice data that anyone can make use of
- ★ **Invest** in the platform and dataset **ecosystem** to support it to grow and thrive
- ★ Sponsor **interventions** like low bandwidth improvements and DEI-oriented model competitions

Common Voice
mozilla

NVIDIA

Open source datasets like Mozilla Common Voice are part of the solution

Languages

100

Hours of data

24,000+

Contributors

500,000

Common Voice

mozilla

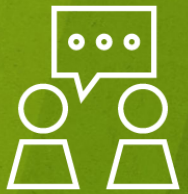
NVIDIA

Linguistic inclusion has wider data health benefits, like speaker diversity and noise profiles

Common Voice
moz://a

NVIDIA

Demographic diversity is key to capturing language diversity



Factors impacting speech variation include:

- Regional origin
- Socioeconomic status/social class
- Race & ethnicity
- Social & personal identities
- Age
- Sex
- Speech-language pathologies
- Context

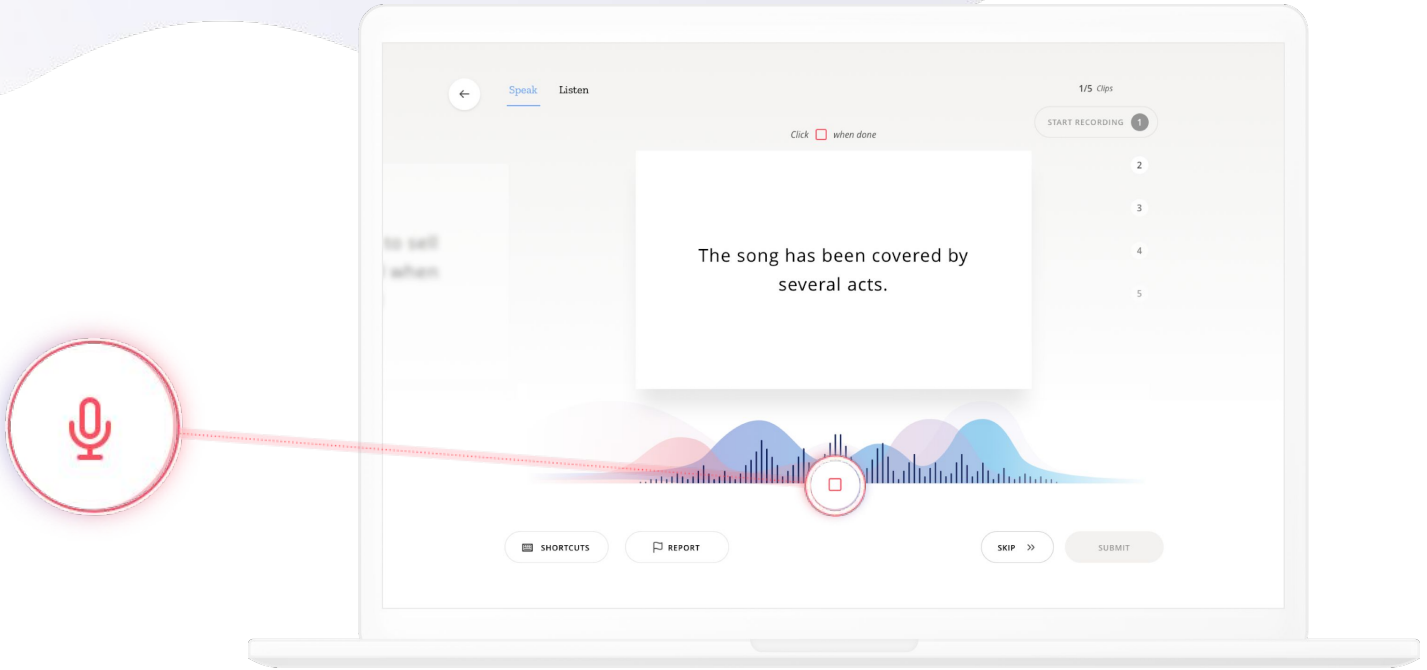
++Interaction Effect

Common Voice
moz://a

NVIDIA

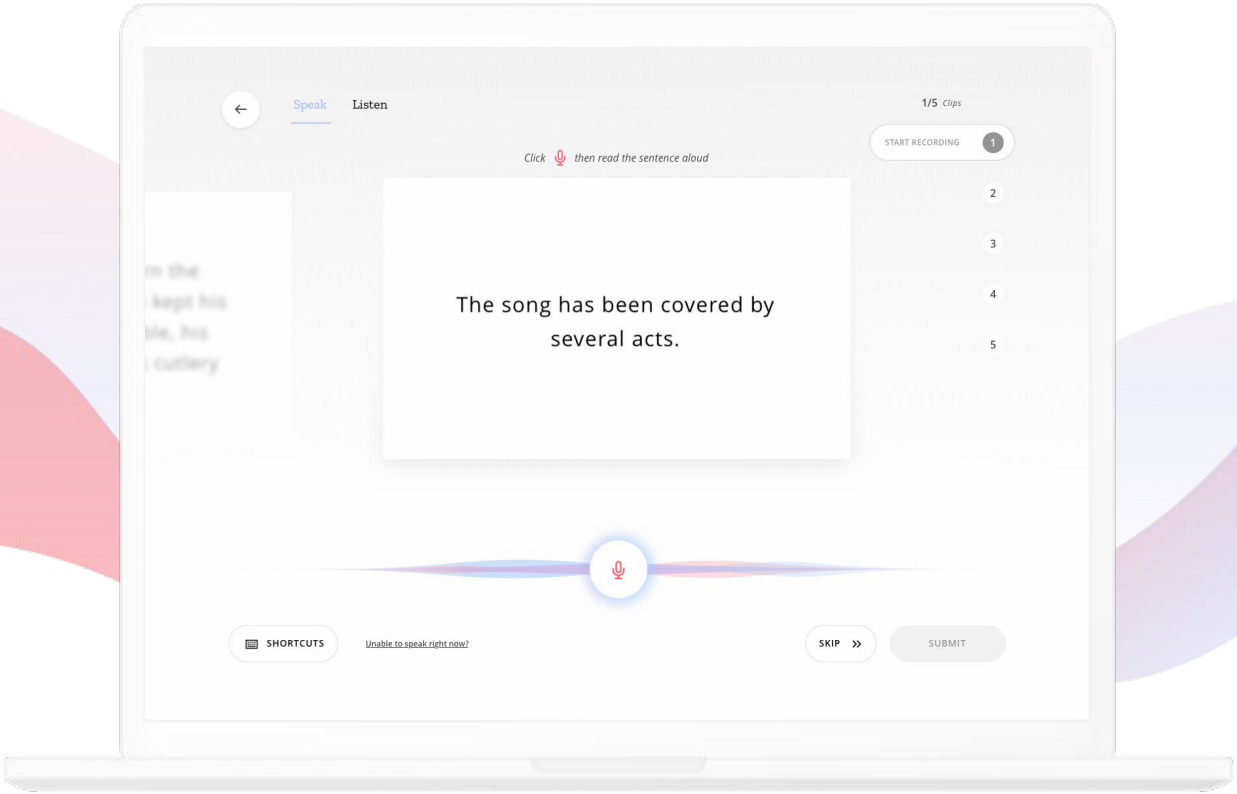

moz://a
Common Voice helps communities build speech datasets for any language or context

Add sentences in any domain - news, politics, health, education



Common Voice
moz://a

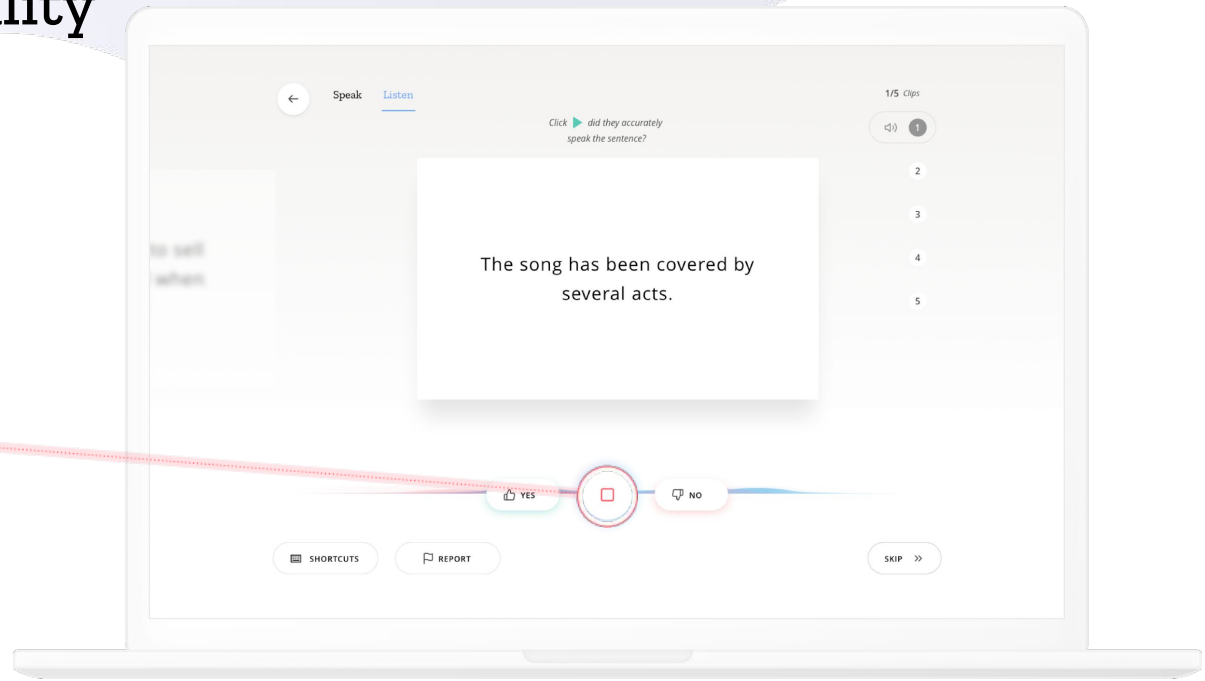
Then, record the sentences as short voice clips



Common Voice
moz://a

Clips are validated to ensure dataset quality

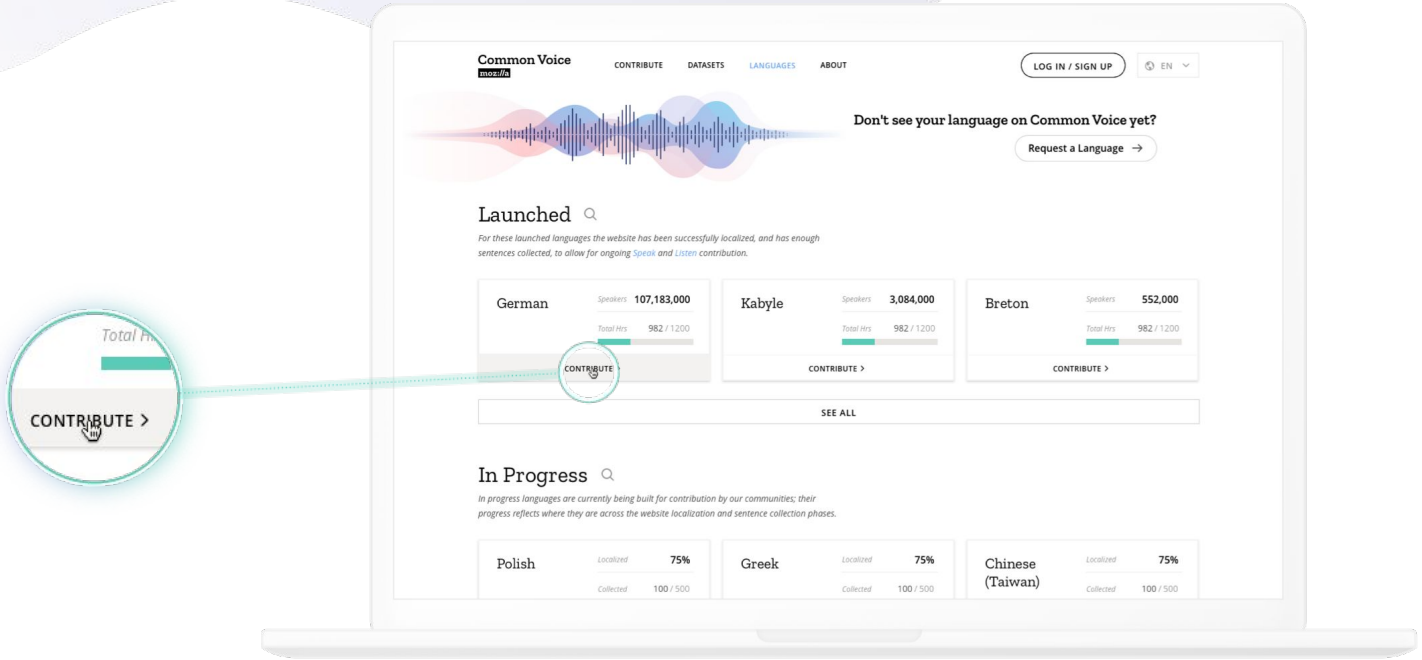
Validate



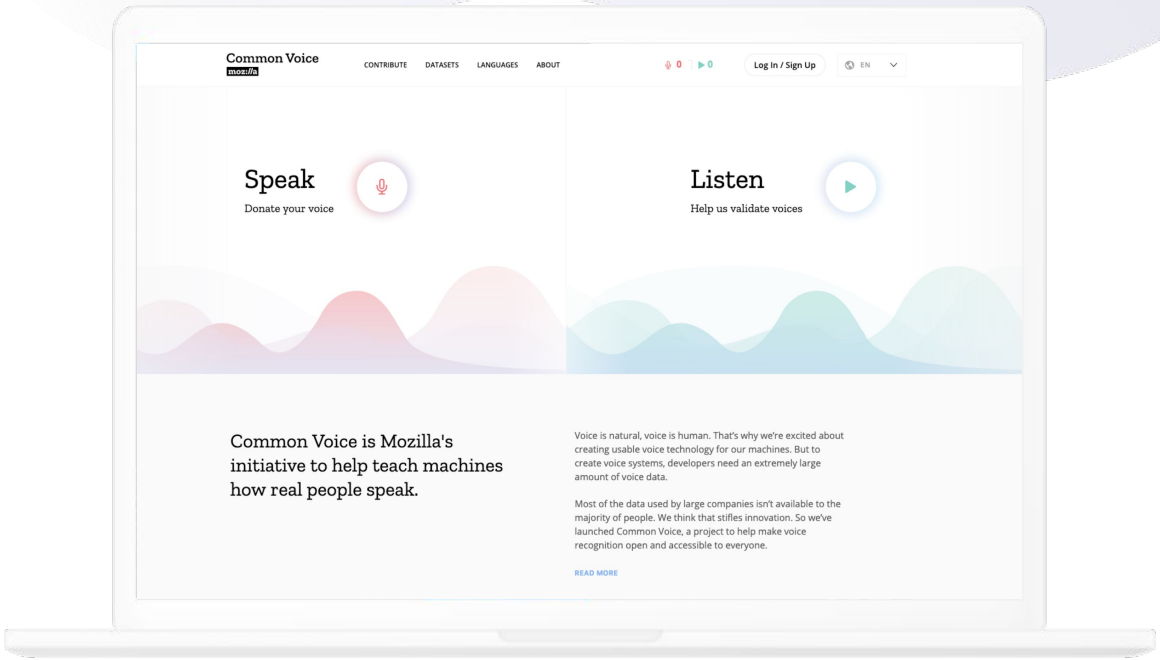
Common Voice
mozilla

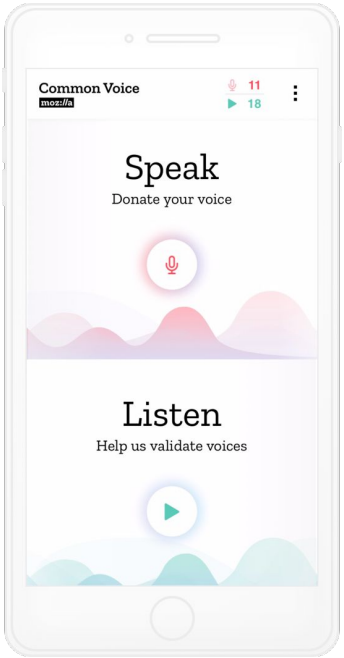
We are adding new languages all the time

Contribute

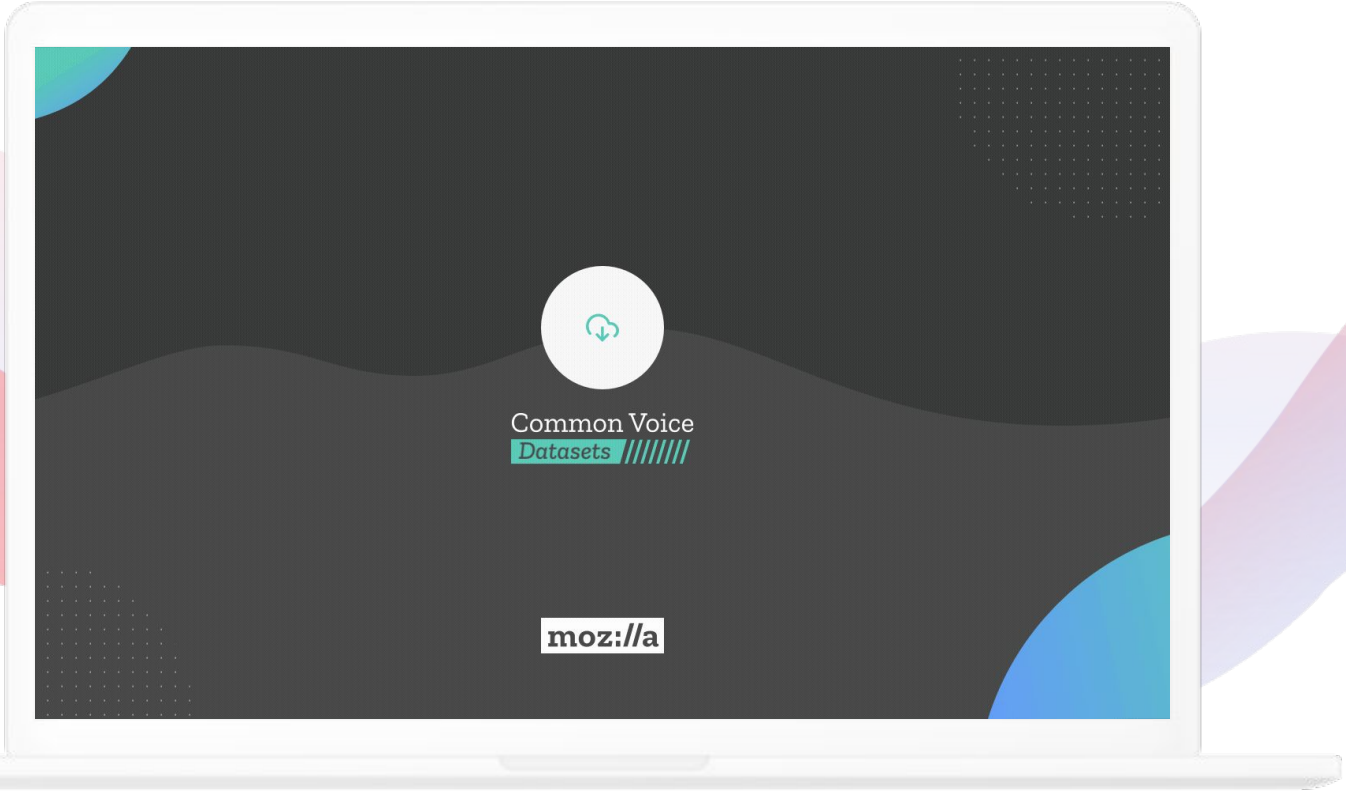


Works on both desktop and mobile

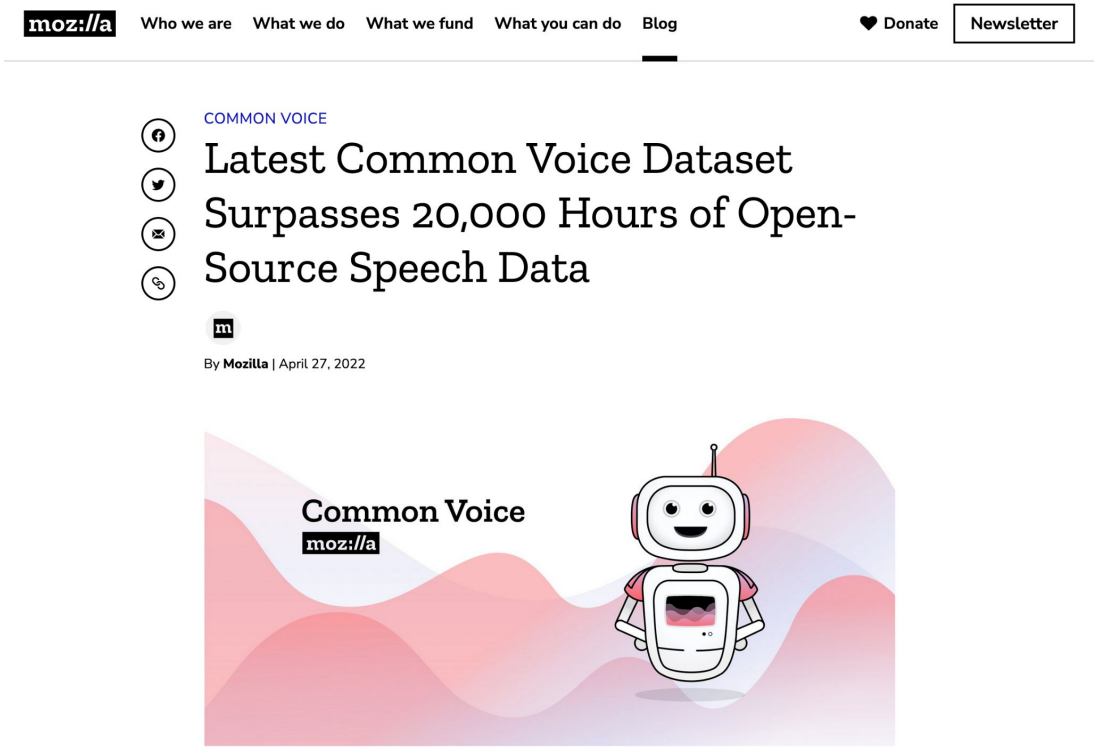




We publish pseudonymised metadata - on variant, accent, gender and age



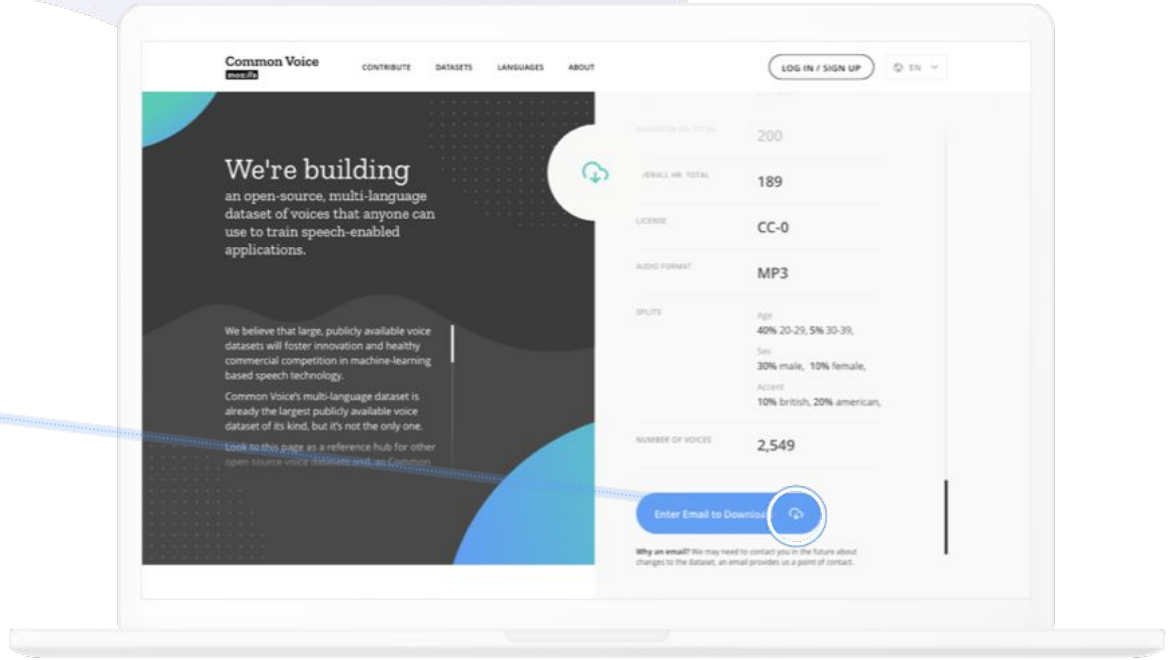
Common Voice Release 9.0 - Case Study



- **The new release features six new languages:** Tigre, Taiwanese (Minnan), Meadow Mari, Bengali, Toki Pona and Cantonese.
- **Twenty seven languages now have at least 100 hours of speech data.** They include Bengali, Thai, Basque, and Frisian.
- **Nine languages now have at least 500 hours of speech data.** They include Kinyarwanda (2,383 hours), Catalan (2,045 hours), and Swahili (719 hours).
- **Nine languages now all have at least 45% of their speaker sex tags as female.** They include Marathi, Dhivehi, and Luganda.

You can download and test the dataset

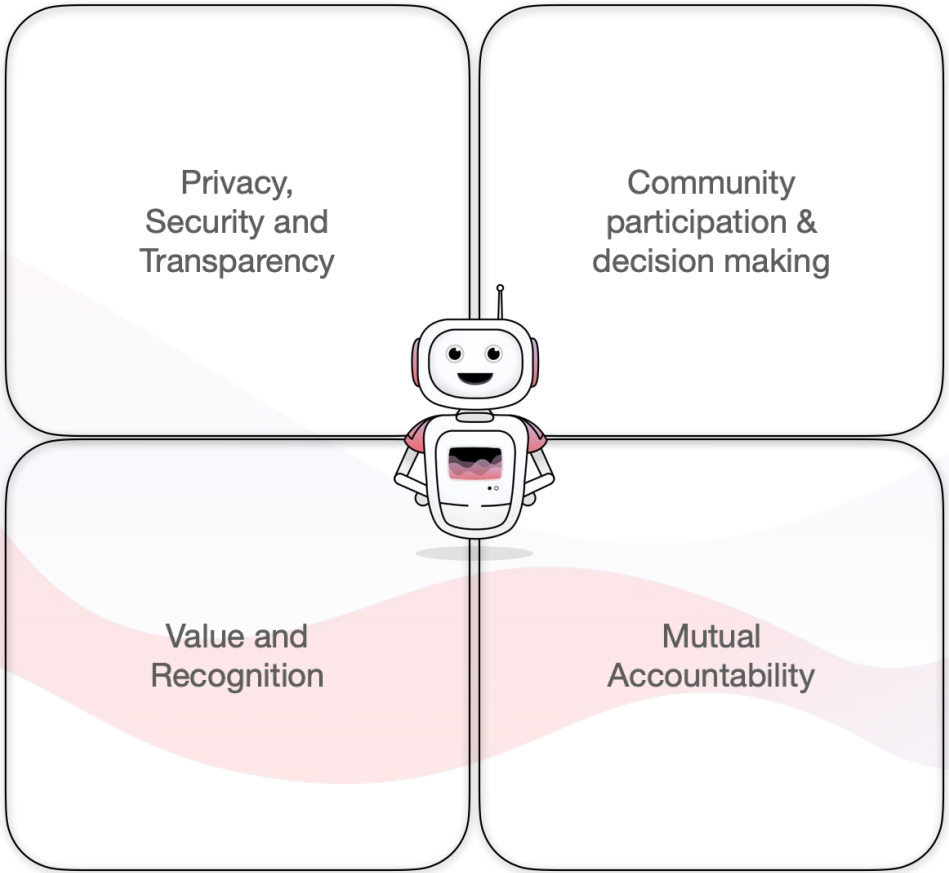
Download



Common Voice

moz://a

Governance



Partnering on Community-driven Speech AI models and methods

The 2022 ‘Our Voices’ Competition

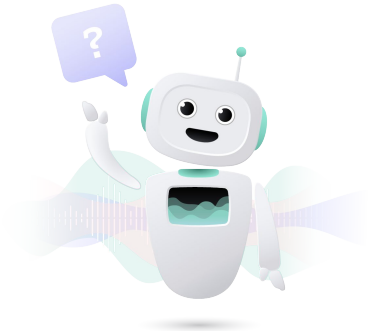
What were we looking for?

We were **deliberately open ended**, but...

Gender

For example...

An STT model that performs equally well for speakers of a gender-marginalised community



Variant and Accent

For example...

Accent classifiers by, and for, a community - with an appropriate use case & usage licence in mind

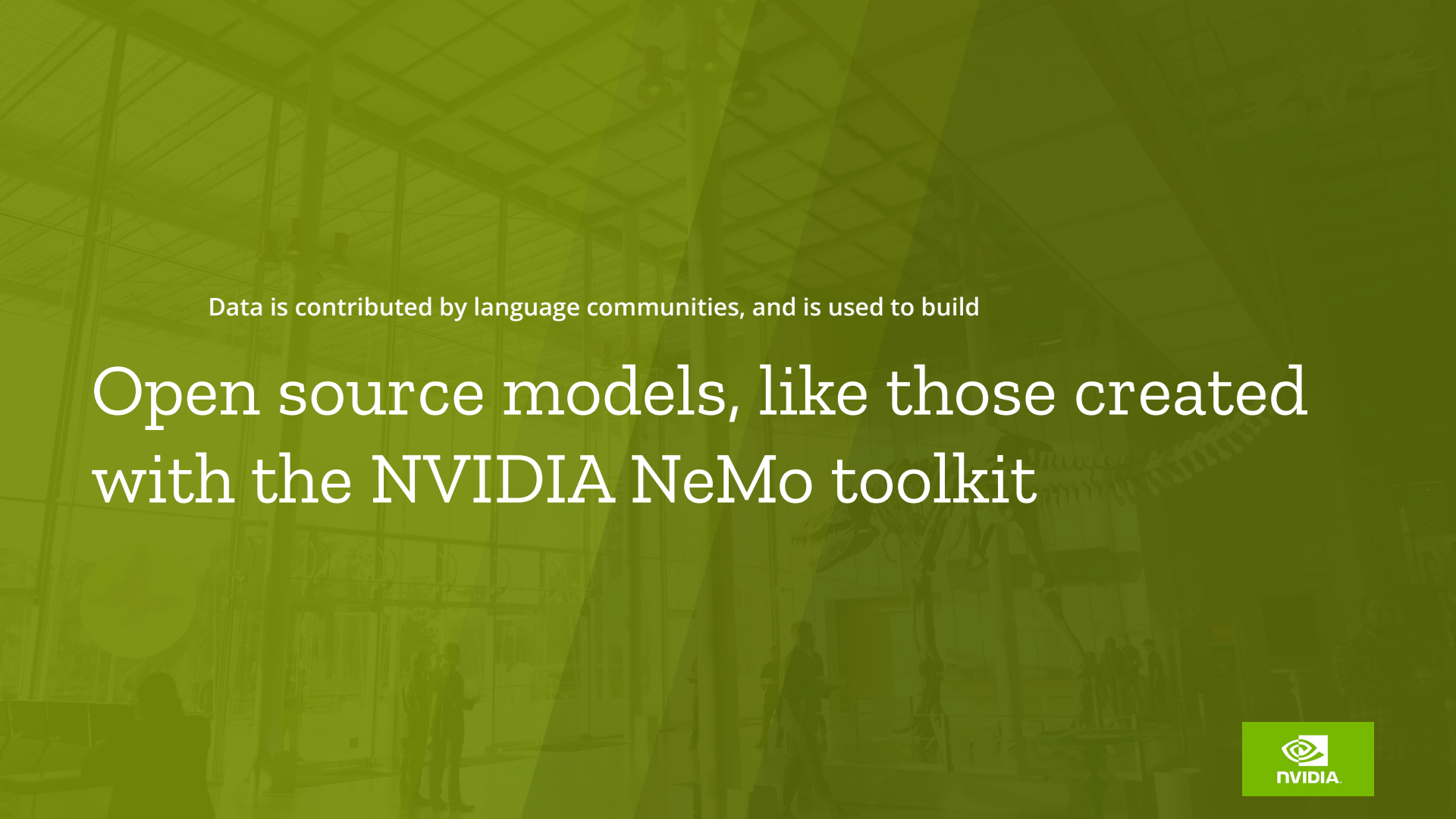
An STT model optimised for an under-served variant of a language family - eg a proof of concept delivered with a small ‘toy’ corpus

Bias methods and measures

For example...

A new benchmark bias corpus

A dataset audit methodology



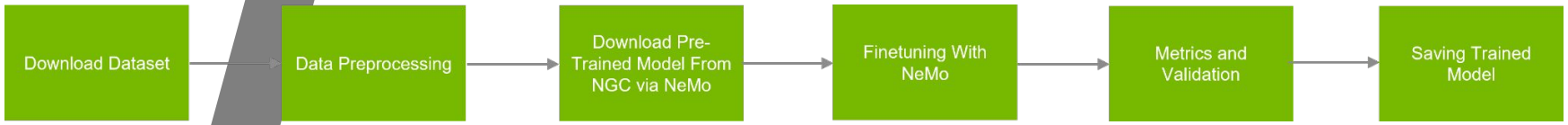
Data is contributed by language communities, and is used to build

Open source models, like those created with the NVIDIA NeMo toolkit

Just some of the NeMo models available to use today...

- ★ *Catalan*
- ★ *English*
- ★ *French*
- ★ *German*
- ★ *Kabyle*
- ★ *Croatian*
- ★ *Ukrainian*
- ★ *Belarusian*
- ★ *Hindi*
- ★ *Marathi*
- ★ *Spanish*
- ★ *Mandarin*
- ★ *Italian*
- ★ *Polish*
- ★ *Russian*
- ★ *Kinyarwanda*

Anyone can use NeMo to obtain state of the art results for a new language, dialect, variant, or accent by simply finetuning NeMo base models for English



Check out these models created by community members:

- Ukrainian lightweight Citrinet model using Mozilla Common Voice Data
- Korean ASR, performed stronger than the latest research model!
- Multilingual Indic SSL model
- Conformer-based ASR for Hindi, Indian English, Kannada, Punjabi, & Tamil

NVIDIA

Together we are opening up Voice Technology for everyone

Speech Synthesis and Text to Speech
Eg. Listen to an audio version of this article

Speech to Text
Eg. See auto generated video captions

Conversational AI
Eg. Automated question and answer chatbots

Voice Assistants
Eg. Operate my mobile device in my own language

IVR Telephone Systems
Eg. Access automated services by calling on my mobile

Common Voice
moz://a

NVIDIA

What you can do...

- ➔ Download the Common Voice dataset!
- ➔ Check out open source NVIDIA models on Hugging Face!
- ➔ Contribute your voice to a dataset on Common Voice today!