NVIDIA Riva ASR

Agenda:

1. Overview of ASR.
2. Introduction for NVIDIA RIVA ASR.
3. Demo

NVIDIA Riva: Using Speech AI for Transcription, Translation, and Voice

Kindly Refer Below Link:

https://www.youtube.com/watch?v=MuQsYVO4Kpo

AUTOMATIC SPEECH RECOGNITION IS EVERYWHERE!

Hundreds of Billions of Minutes of Speech Generated Daily

Contact Center

500M Calls Daily

Consumer Applications

1.8B Daily Minutes

Online Meetings

200M Daily

CRISP / Service Provider

Education

Energy

Finance

Healthcare

Media & Entertainment

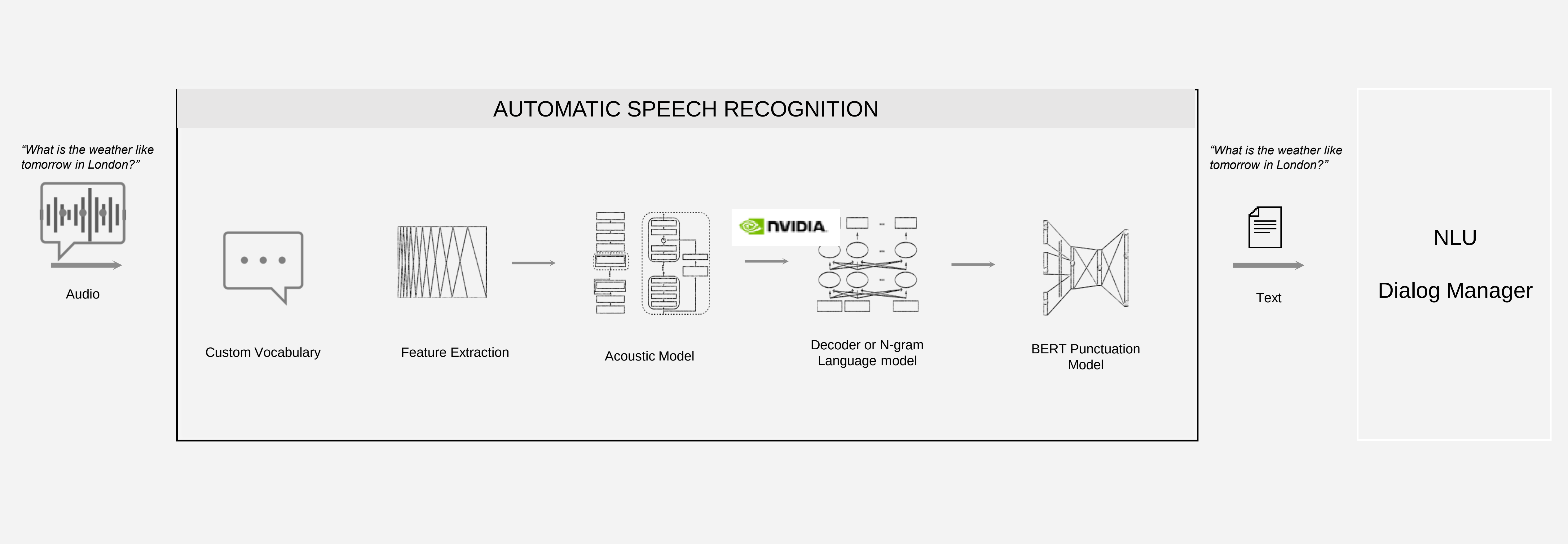
Retail

Telecom

Transportation

ACHIEVING HIGH ACCURACY FOR REAL-WORLD ASR APPLICATIONS IS CHALLENGING

Domain-Specific Language| Contextual | Enterprise-Specific Vocabulary| Noisy Environments | Accents



"What is the weather like tomorrow in London?"

Audio

AUTOMATIC SPEECH RECOGNITION

Custom Vocabulary

Feature Extraction

Acoustic Model

Decoder or N-gram Language model

BERT Punctuation Model

"What is the weather like tomorrow in London?"

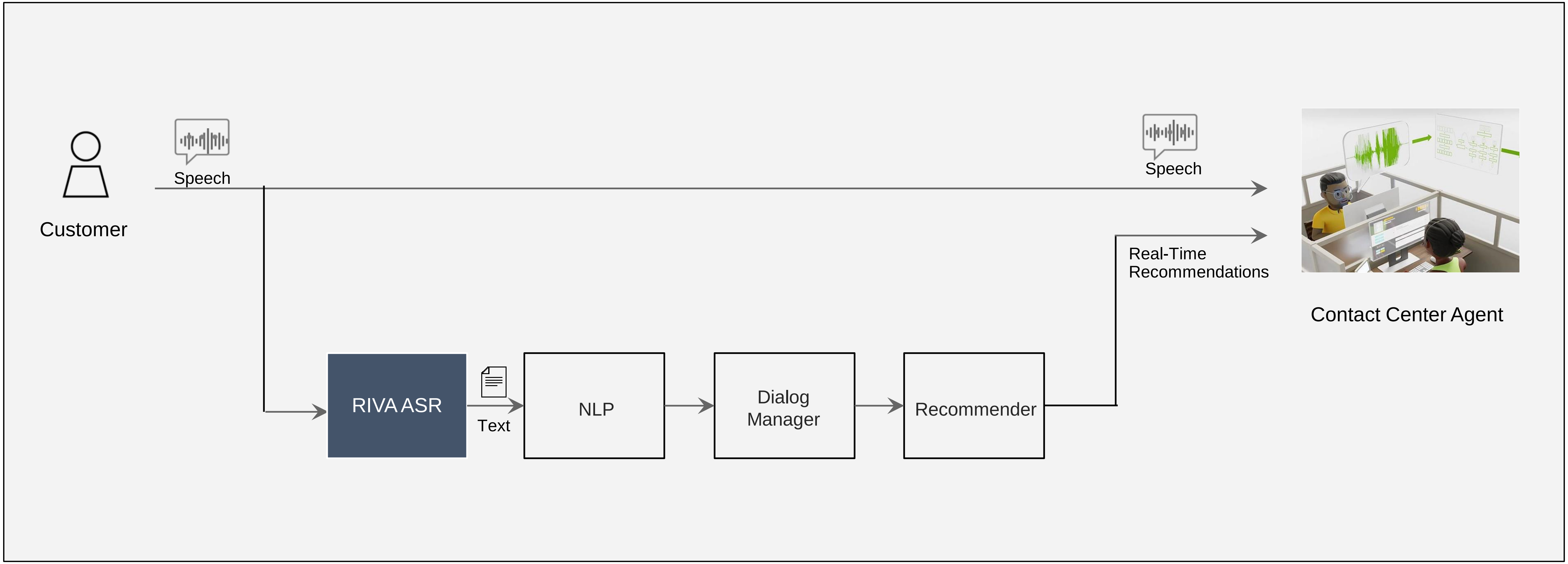
Text

NLU

Dialog Manager

NLU – Natural Language Understanding

CONTACT CENTER AGENT ASSIST REQUIRE HIGHLY ACCURATE REAL-TIME ASR



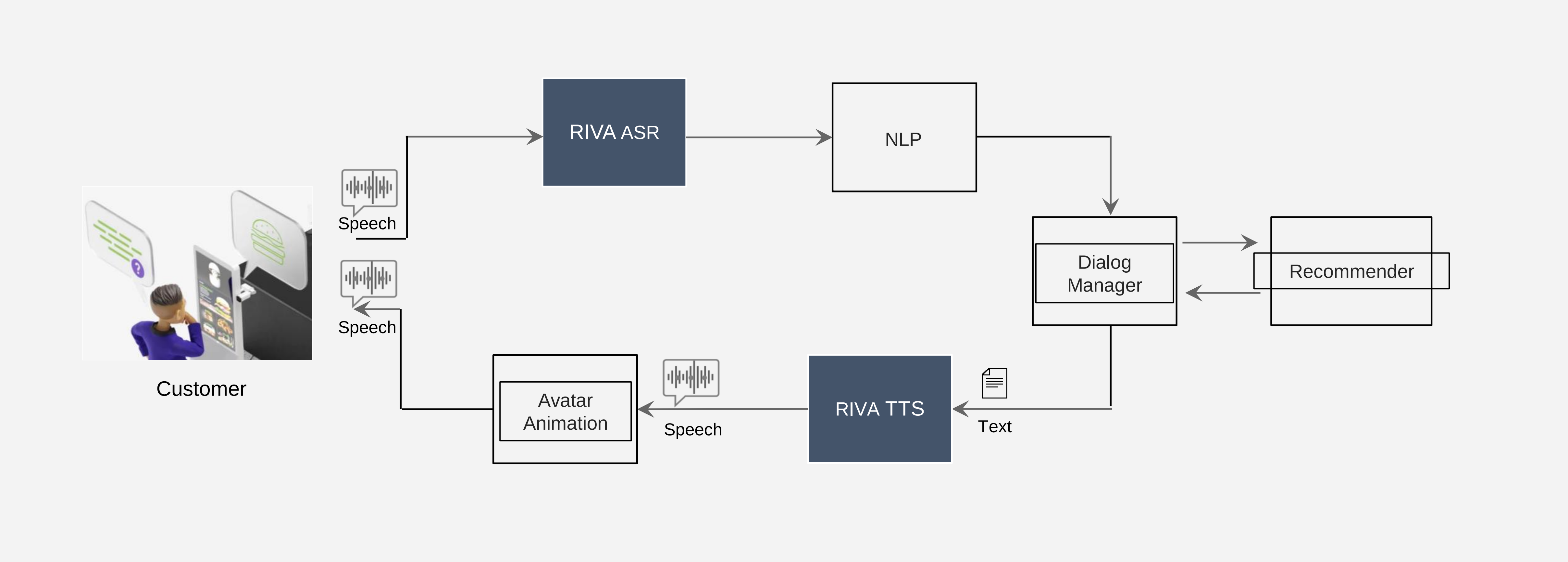
ASR – Automatic Speech Recognition | NLP – Natural Language Processing

Reinvent Contact Center Experiences with NVIDIA Riva Transcription

Kindly Refer Below link:

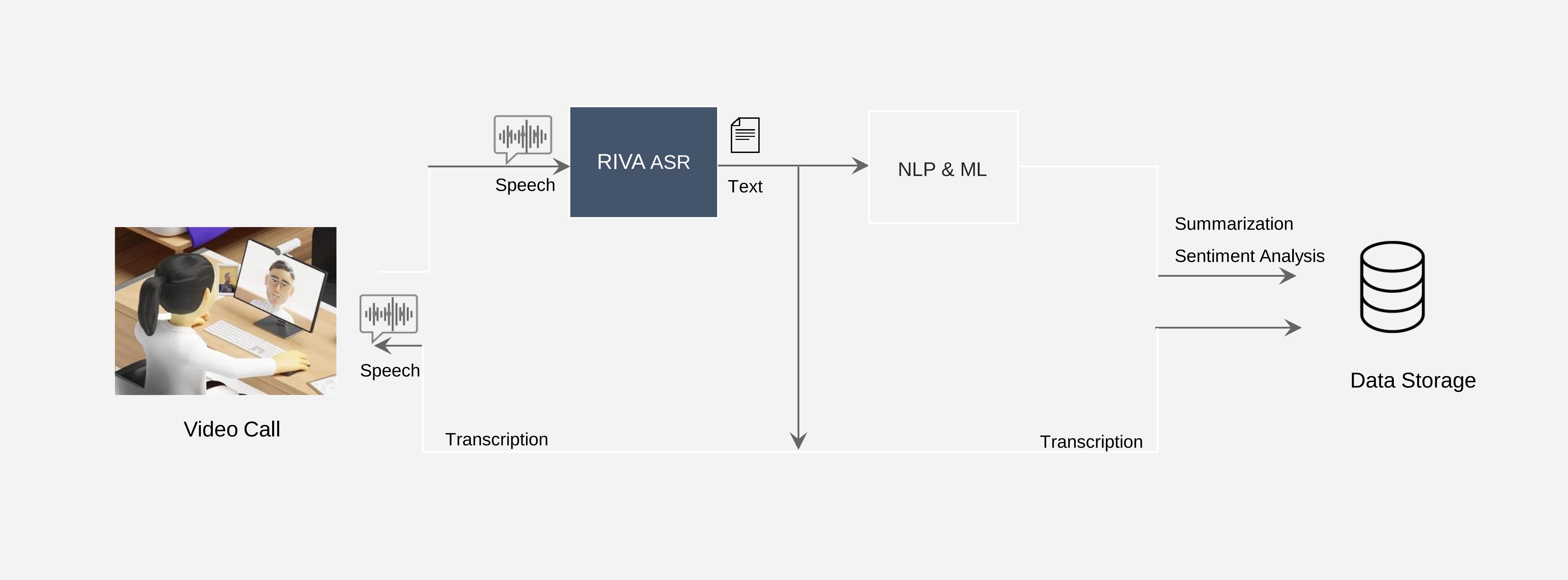
https://www.youtube.com/watch?v=QLi1vljEGNo

CONSUMERS EXPECT REAL-TIME, ACCURATE, AND EXPRESSIVE DIGITAL AVATARS



ASR – Automatic Speech Recognition | NLP – Natural Language Processing | TTS – Text-To-Speech

ONLINE MEETING TRANSCRIPTION ACCURACY INFLUENCES DOWNSTREAM TASKS QUALITY



ASR – Automatic Speech Recognition | NLP – Natural Language Processing | ML – Machine Learning

RIVA ASR SOLVES CUSTOMERS PAIN POINTS

Customer Pain Points

Low Accuracy with No Access to SOTA Speech Models

< 300 ms

Tight Latency for Natural Conversations

Flexible Scalability & Data Privacy Issues

World-Class ASR

✓ Best-In-Class Accuracy with From-Data-To-Model ASR Pipeline Customization

✓ < 300 ms

Real-Time Latency Delivery for Natural Conversations

✓ Large-Scale Deployment on Customer-Choice Infrastructure

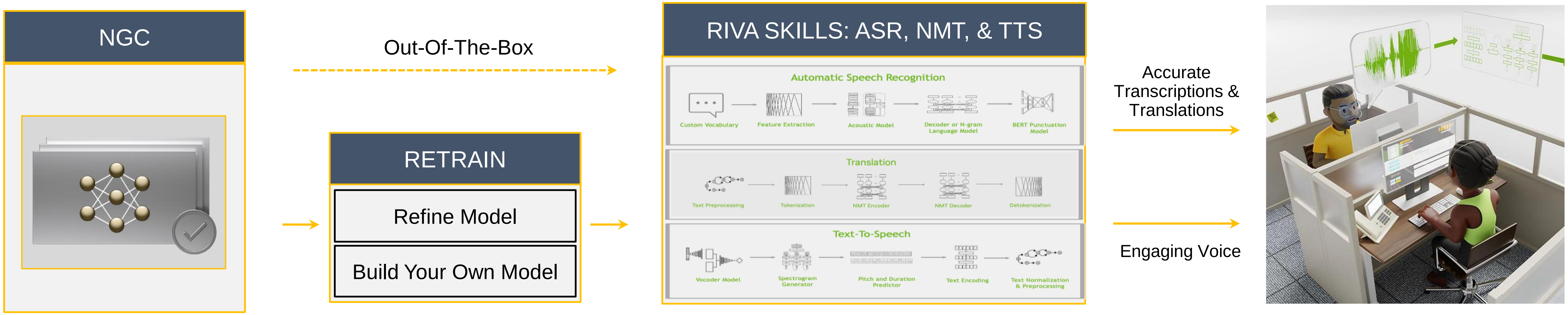
Any Cloud On-Prem Edge Embedded

ASR – Automatic Speech Recognition | SOTA – State-Of-The-Art

NVIDIA Riva ASR

NVIDIA RIVA

Fully Customizable, GPU-Accelerated SDK for Real-Time Speech and Translation AI



NGC

Out-Of-The-Box

RETRAIN

Refine Model

Build Your Own Model

RIVA SKILLS: ASR, NMT, & TTS

Accurate Transcriptions & Translations

Engaging Voice

- SOTA Pre-Trained Models
- AI Workflows:
  - Audio Transcription
  - Intelligent Virtual Assistant

Train for Any
Domain and Language

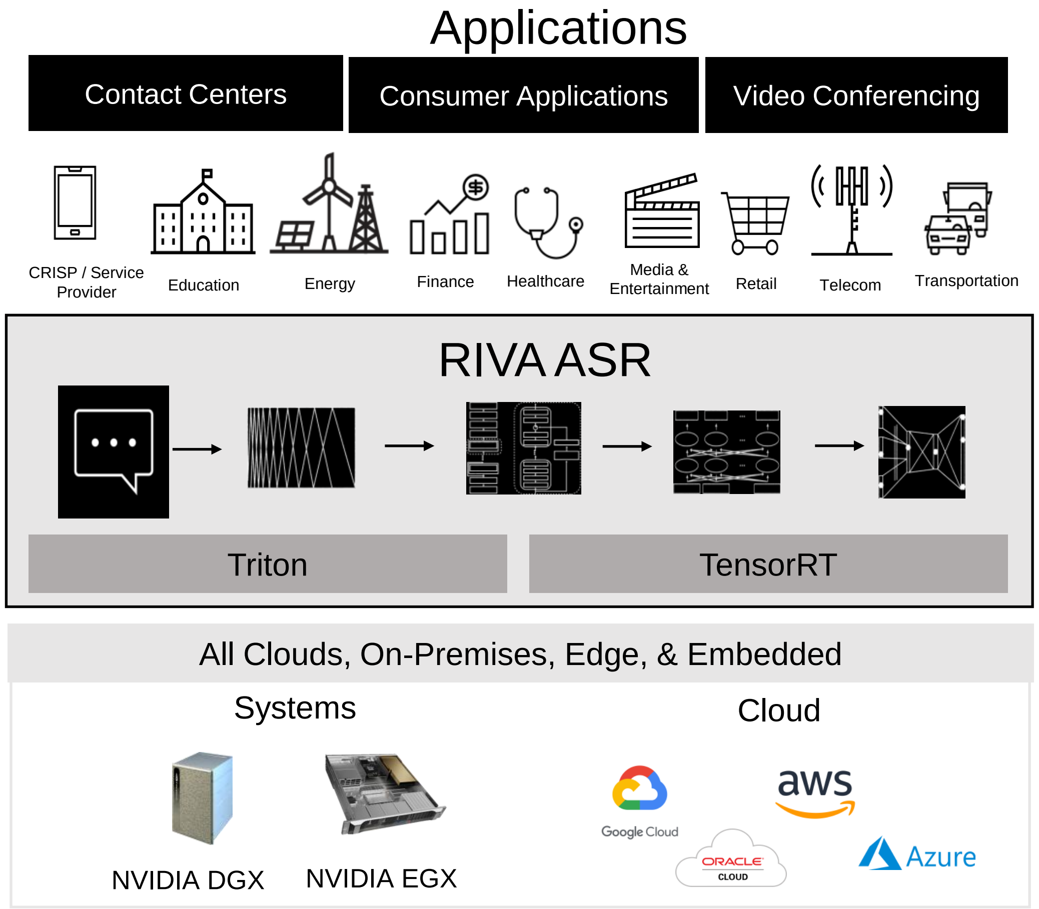
- Fully Customizable for the Best Possible Accuracy
- GPU-Accelerated, Real-Time
- Scale to Hundreds of Thousands of Users
- On-Prem, in Any Cloud, at the Edge, or Embedded

SOTA – State-Of-The-Art | ASR – Automatic Speech Recognition | NMT – Neural Machine Translation | TTS – Text-To-Speech

NVIDIA RIVA AUTOMATIC SPEECH RECOGNITION

The Best Possible ASR with SOTA Deep Learning

- **Accurate ASR with SOTA* models** trained for 1M+ hrs on 70K hrs of speech
- **14 languages support**: Arabic, Chinese (Mandarin), English (US/UK), French, German, Hindi, Italian, Japanese, Korean, Portuguese, Russian, Spanish (LATAM/Spain)
- **2X accuracy improvement** with customizations for:
  - Industry specific jargon
  - Accents & dialects
  - Noisy environments
- **Real-time** performance far below 300ms for interactive speech apps
- **High scale** of 100s thousands of concurrent streams
- **Runs anywhere:** all clouds, on-premises, at the edge, embedded



Applications

Contact Centers | Consumer Applications | Video Conferencing

CRISP / Service Provider, Education, Energy, Finance, Healthcare, Media & Entertainment, Retail, Telecom, Transportation

RIVA ASR

Triton | TensorRT

All Clouds, On-Premises, Edge, & Embedded

Systems: NVIDIA DGX, NVIDIA EGX

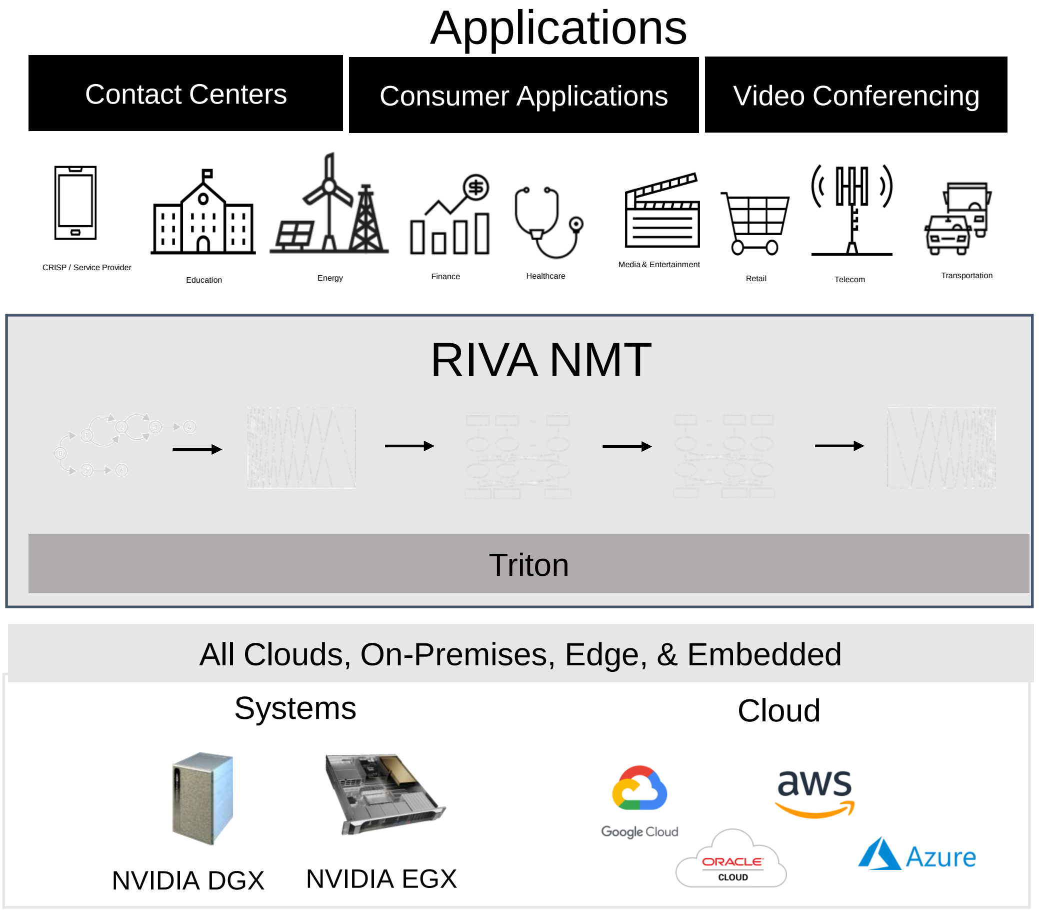
Cloud: Google Cloud, aws, ORACLE CLOUD, Azure

SOTA – State-Of-The-Art | ASR – Automatic Speech Recognition | TTS – Text-To-Speech

NVIDIA RIVA NEURAL MACHINE TRANSLATION

High accuracy real-time translations with bilingual and multilingual models

- **High quality OOTB translation models**
- **Bilingual and multilingual models** for:
  - English <-> Spanish, Mandarin, Russian, German, French
- **Customizations** to improve translation accuracy for domain-specific use cases across all industries



Applications

Contact Centers | Consumer Applications | Video Conferencing

RIVA NMT

Triton

All Clouds, On-Premises, Edge, & Embedded

Systems

NVIDIA DGX

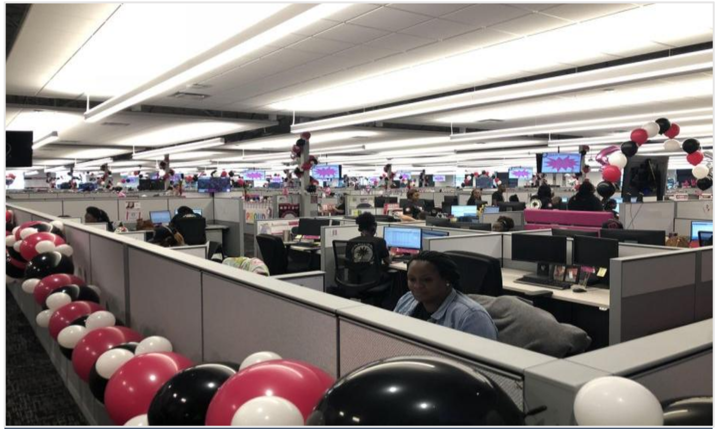
NVIDIA EGX

Cloud

OOTB – Out-Of-The-Box | NMT – Neural Machine Translation

NVIDIA RIVA ASR CUSTOMER WINS

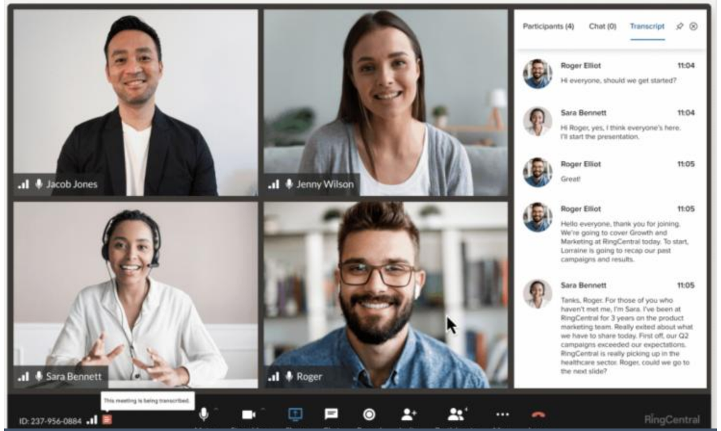
Telecommunications | UCaaS | Healthcare | Consumer Application



T Mobile

~3X Accuracy & 10X Speed Improvement

- ASR for Customer Support
  - Agent Assist
  - Real-Time Analytics



Participants (4) Chat (0) Transcript

Roger Elliot 11:04
Hi everyone, should we get started?

Sara Bennett 11:04
Hi Roger, yes, I think everyone's here. I'll start the presentation.

Roger Elliot 11:05
Great!

Roger Elliot 11:05
Hello everyone, thank you for joining. We're going to cover Growth and Marketing at RingCentral today. To start, Lorraine is going to recap our past campaigns and results.

Sara Bennett 11:05
Tanks, Roger. For those of you who haven't met me, I'm Sara. I've been at RingCentral for 3 years on the product marketing team. Really exited about what we have to share today. First off, our Q2 campaigns exceeded our expectations. RingCentral is really picking up in the healthcare sector. Roger, could we go to the next slide?

RingCentral

2X Better Accuracy &
No Accent & Noisy Environment Issues

- ASR for Video Call Transcripts
  - Real-Time Meeting Transcriptions
  - Summarizations

SOTA WER for Noisy Environments &
Expressive Synthetic Voices

- ASR & TTS for Consumer Application
  - 100M Monthly Active Users
  - Integrated in Voice-Enabling Dev Tool

AI²LABS AVAYA botpress Computacenter DATA MONSTERS Deloitte. Dialoga Group Floatbot Hewlett Packard Enterprise interactions

InstaDeep™ kore.ai LEXISTEMS. MALAMUTE MINERVA moneypenny pendulum plabook quantiphi

Read replicant siminsights SmartCow SVA softserve TARTEEL AI VectorVentures

T-MOBILE AGENTS RESOLVE CUSTOMER QUERIES USING RIVA SPEECH AI

Contact center agents resolve quicker customer queries with agent assists* & real-time analytics

Challenge:

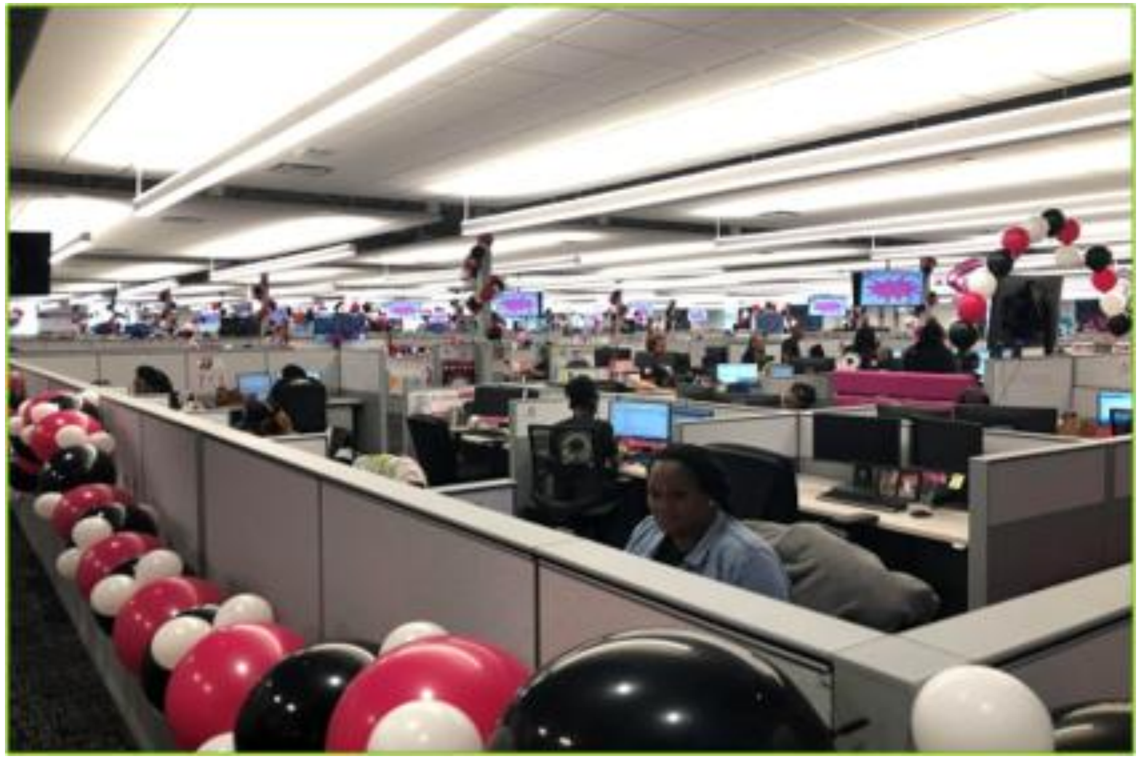
- Low accuracy for product names & domain-specific words
- High latency

Solution:

- Language model customization for domain-specific words
- Word boosting for product names

Result:

- 3X accuracy improvement with WER drop from 27% to 9%
- 10X speed up with latency drop from 2.5 sec to 250 msec



T Mobile

*Agent Assist with integrated Riva was deployed in the US in 2022 and world-wide by the end of 2023. For more, check out T-Mobile GTC Fall 2022 talk

ENEL X USES RIVA ASR TO IMPROVE CALL CENTER PRODUCTIVITY & REDUCE COST

The global leader in electrification & e-mobility uses Minerva CQ platform & Riva speech AI skills for real-time agent coaching & guidance

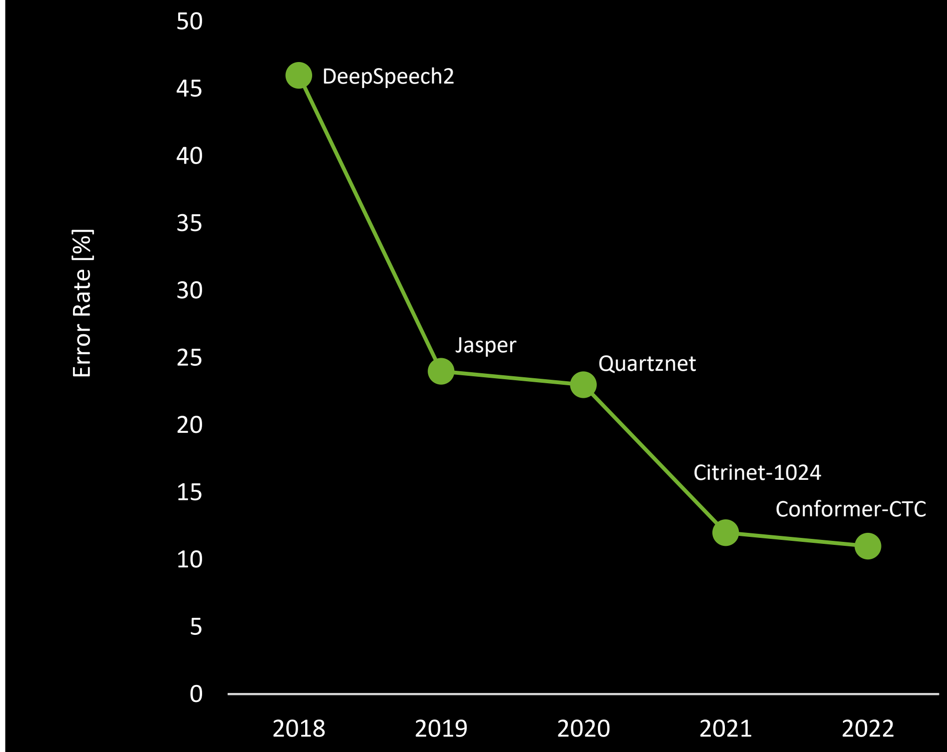
- ✓ Increased First Contact Resolution (Less Follow-Ups) +12.5% | 69.5% to 82%
- ✓ Reduced Handle Times by 44% | 9min to 5min, saving $2.67 per call
- ✓ Decrease Onboarding Time by 75% | 20 days to 5 days
- ✓ NPS and CSAT* Gap Improved by 50%
- ✓ Reduced Cost by Faster agent training: 5 days vs. 20 in training time per year

MINERVA

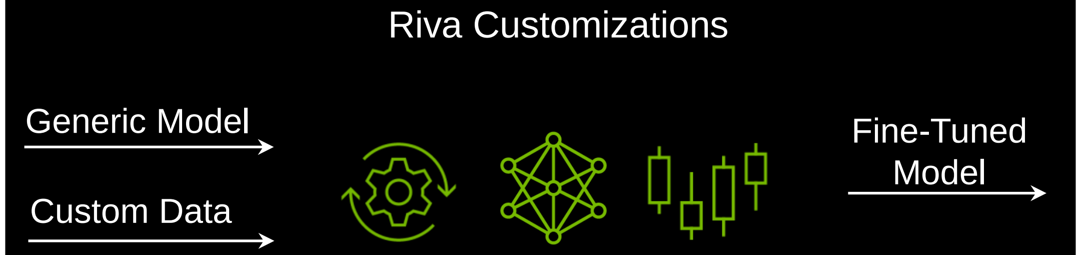
* CSAT – Customer satisfaction score in %

NVIDIA RIVA DELIVERS WORLD-CLASS ASR

4X Accuracy with SOTA Model Architectures



Additional 2X Accuracy with Customizations


Riva Customizations

Generic Model

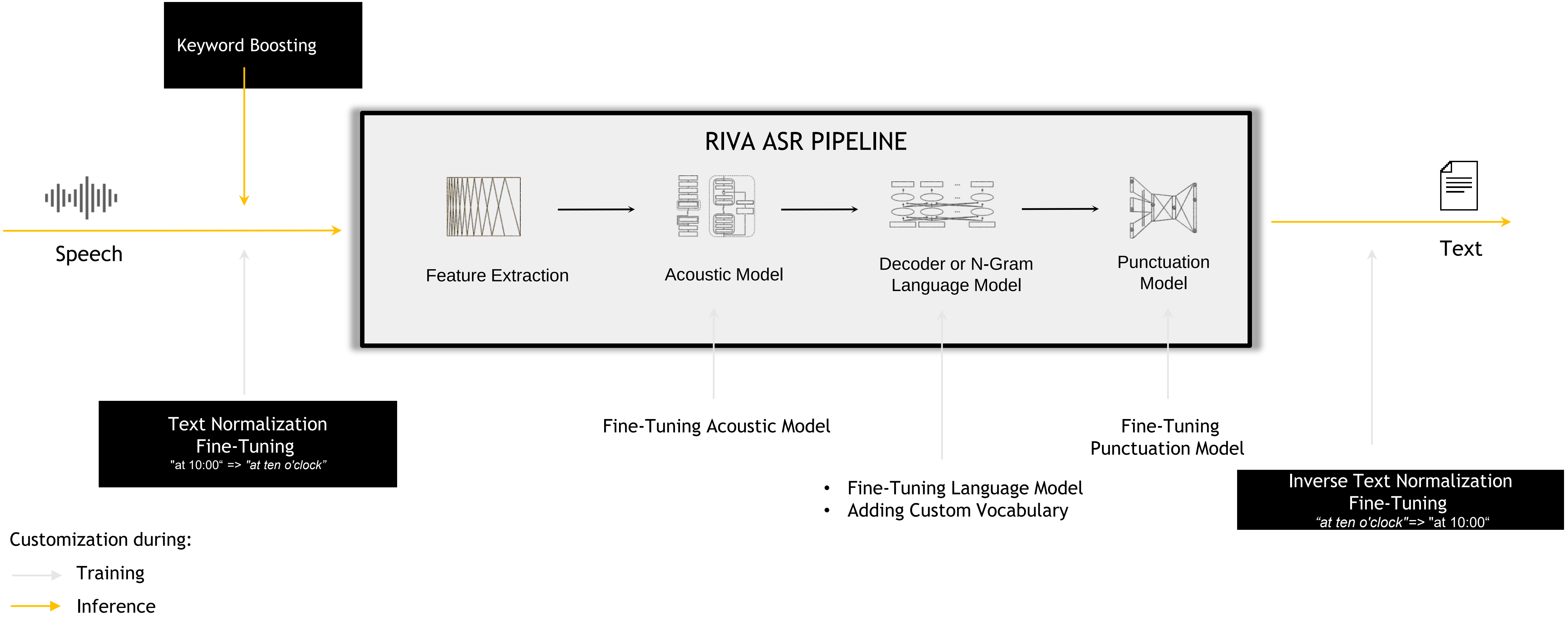
Custom Data

Fine-Tuned Model

ASR – Automatic Speech Recognition | SOTA – State-Of-The-Art

CUSTOMIZATIONS ACROSS RIVA ASR PIPELINE

- Accurate | Customizable | Real Time



Keyword Boosting

RIVA ASR PIPELINE

Text Normalization Fine-Tuning
"at 10:00“ => "at ten o'clock”

Fine-Tuning Acoustic Model

- Fine-Tuning Language Model
- Adding Custom Vocabulary

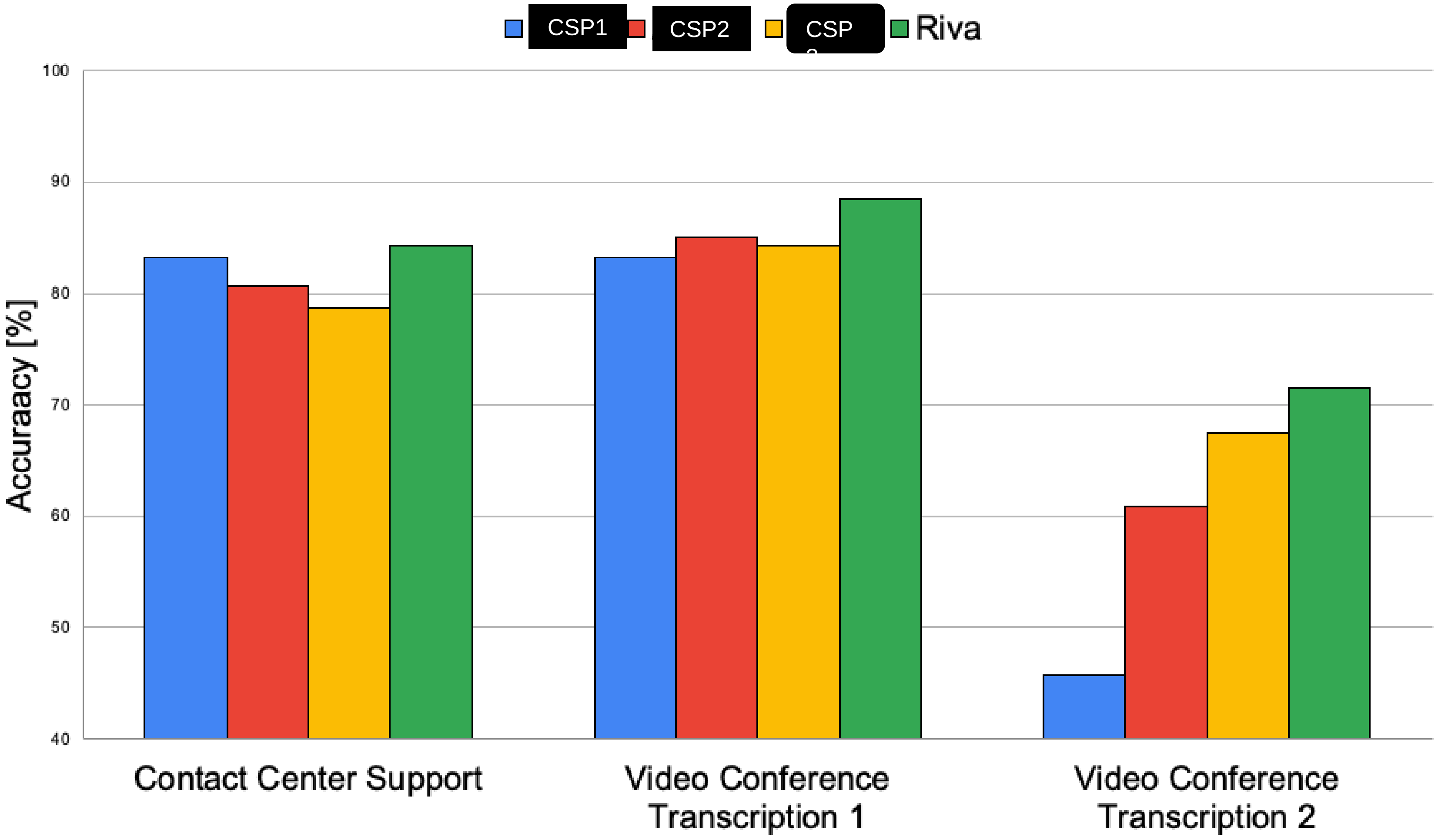
Fine-Tuning Punctuation Model

Inverse Text Normalization Fine-Tuning
“at ten o'clock”=> "at 10:00“

Customization during:

RIVA ASR OUTPERFORMS FOR EVERY TESTED USE CASE

Riva State-Of-The Art Deep Learning Models are Trained on a Variety of Domain-Specific Data

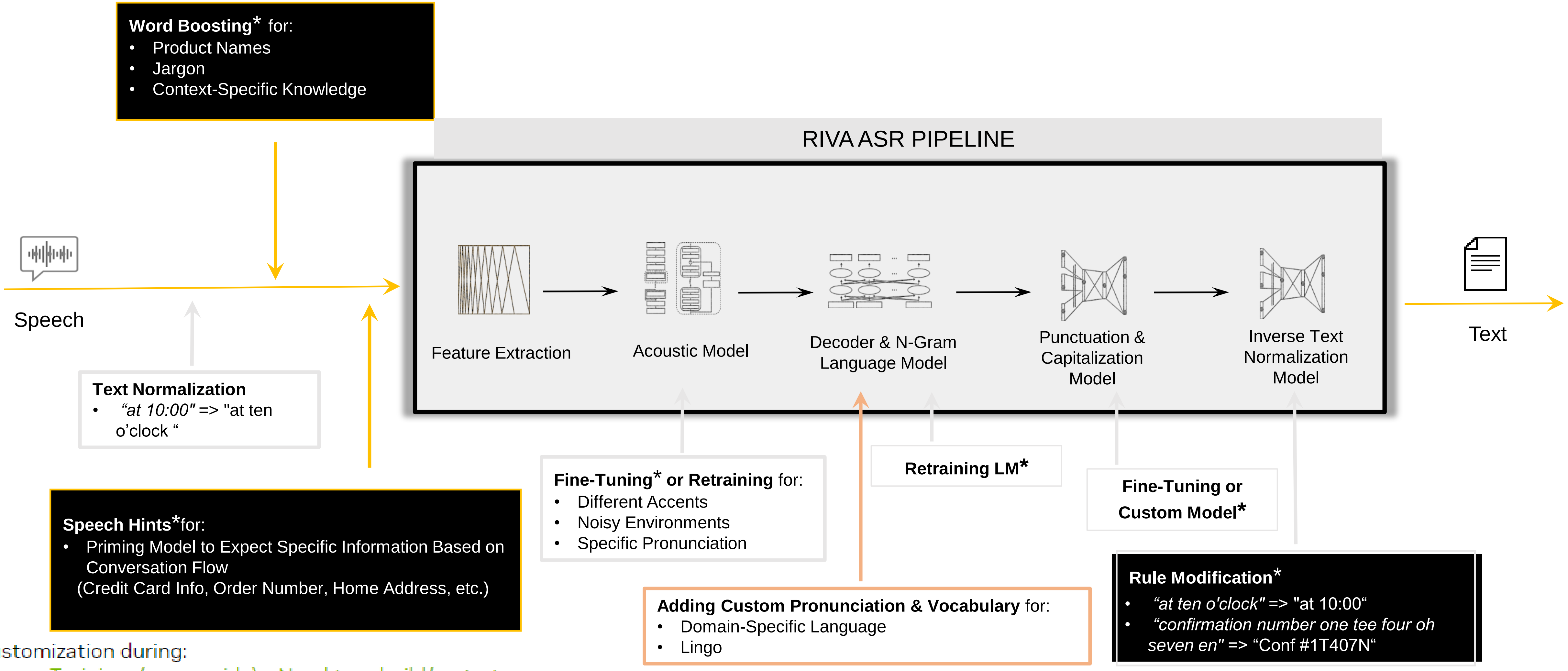


Note: Tests performed in April, 2022

One Vision. One Goal... Advanced Computing for Human Advancement...

CUSTOMIZATIONS ACROSS RIVA ASR PIPELINE

Accurate | Customizable | Real-Time



Word Boosting* for:

- Product Names
- Jargon
- Context-Specific Knowledge

RIVA ASR PIPELINE

Speech → Feature Extraction → Acoustic Model → Decoder & N-Gram Language Model → Punctuation & Capitalization Model → Inverse Text Normalization Model → Text

Text Normalization

- *“at 10:00"* => "at ten o’clock “

Speech Hints*for:

- Priming Model to Expect Specific Information Based on Conversation Flow

(Credit Card Info, Order Number, Home Address, etc.)

Fine-Tuning* or Retraining for:

- Different Accents
- Noisy Environments
- Specific Pronunciation

Retraining LM*

Fine-Tuning or Custom Model*

Adding Custom Pronunciation & Vocabulary for:

- Domain-Specific Language
- Lingo

Rule Modification*

- *“at ten o'clock”* => "at 10:00“
- *“confirmation number one tee four oh seven en"* => “Conf #1T407N“

Customization during:

→ Training (server-side) – Need to rebuild/restart server

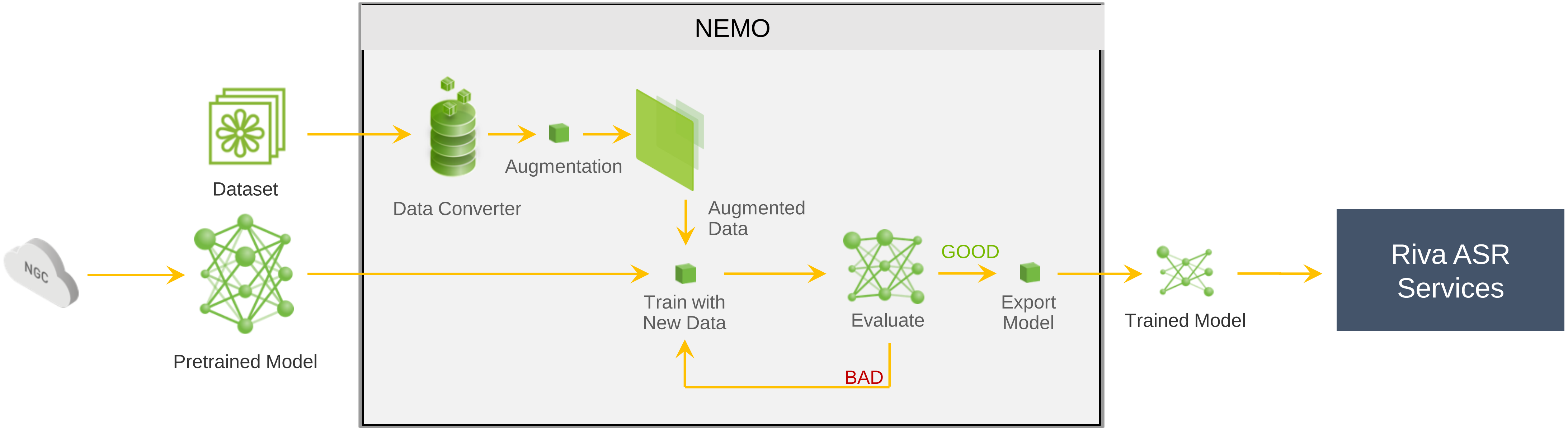
→ Deployment

→ Inference (client-side) – No need to rebuild/start Riva server

* Generic ASR Services offer black box with rigid & selectively limited customization

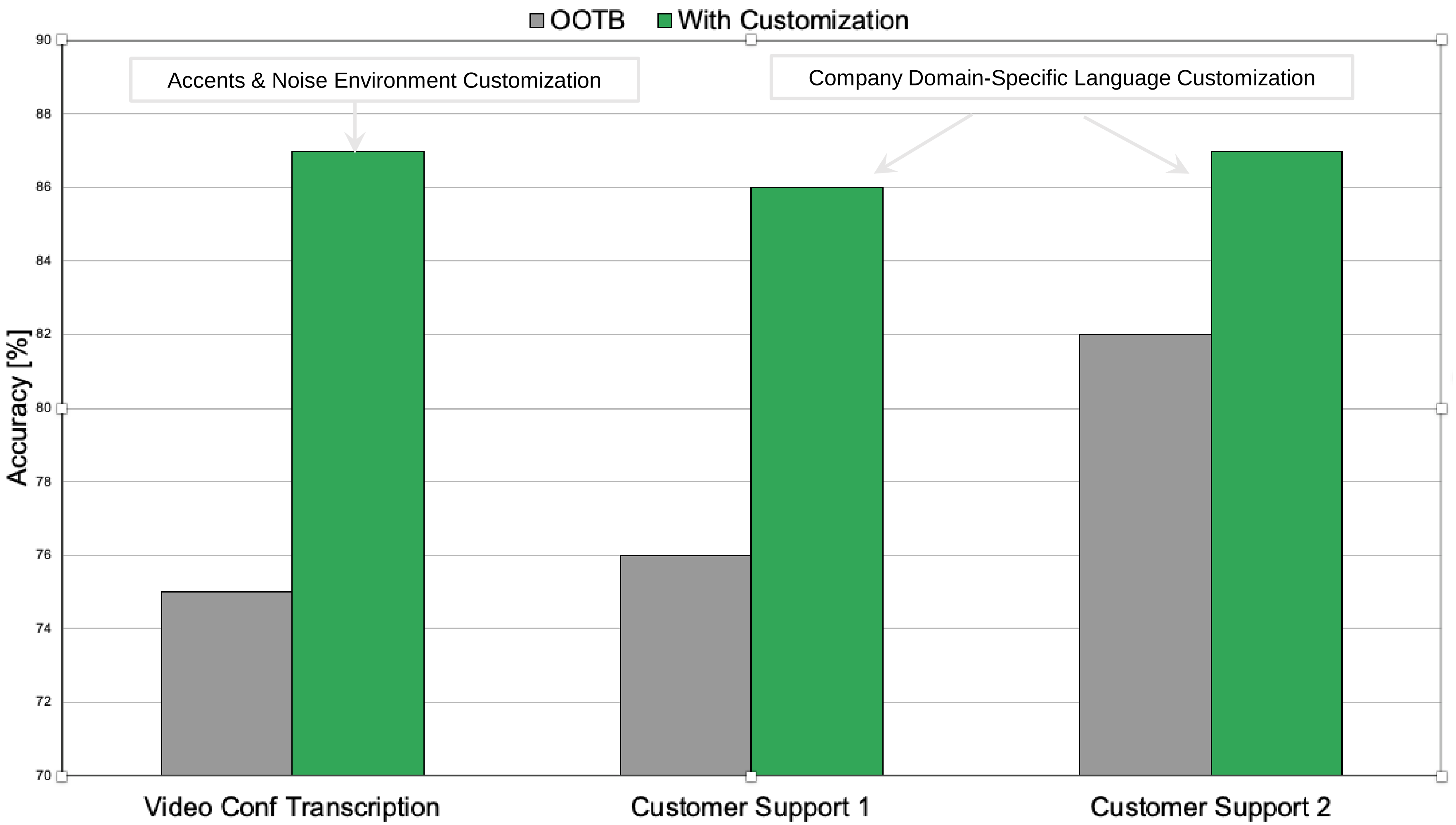
INCREASE ASR ACCURACY BY FINE-TUNING WITH NEMO

Low-Code | GPU-Optimized | Deploy to Riva



ASR – Automatic Speech Recognition | NGC – NVIDIA GPU Cloud

INNOVATE & DIFFERENTIATE WITH RIVA ASR CUSTOMIZATION ON YOUR DATASETS

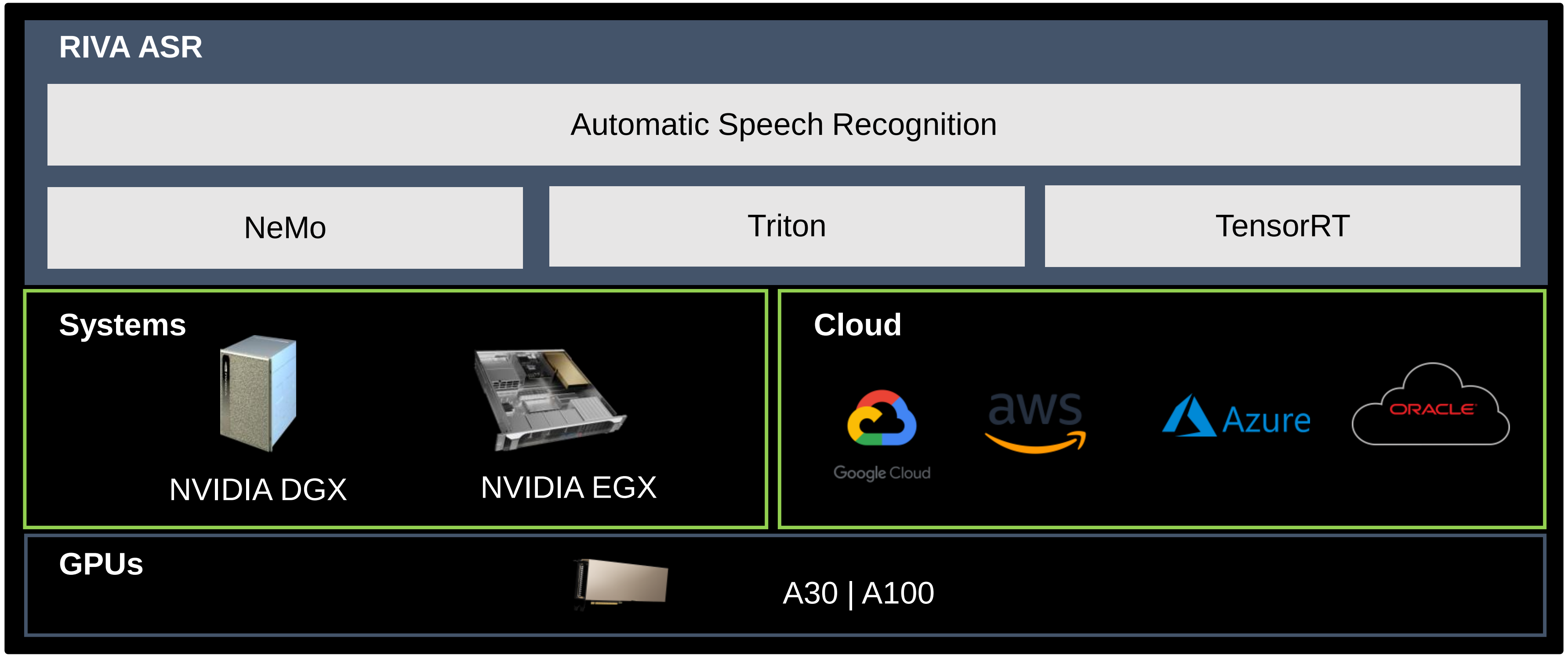


Note: Tests performed in December, 2021.

One Vision. One Goal... Advanced Computing for Human Advancement...

RIVA ASR RUNS ON ANY CLOUD, ON-PREMISES, AT THE EDGE, AND EMBEDDED

GPU-Optimized for Minimum Latency and Maximum Throughput



RIVA ASR

Automatic Speech Recognition

NeMo

Triton

TensorRT

Systems

NVIDIA DGX

NVIDIA EGX

Cloud

GPUs

A30 | A100

NVIDIA RIVA ASR DEMOS

Hello **Hola** **Привет** **Hallo** **Salut** **नमस्ते**

أهلا **안녕하세요** **Olá** **Ciao**

ENGLISH ASR

Kindly refer the link below:

https://www.youtube.com/watch?v=HOCxJqYLf3k

SPANISH, GERMAN, RUSSIAN ASR

Kindly refer the link below:

https://www.youtube.com/watch?v=GljT2YbKodg

https://www.nvidia.com/en-us/ai-data-science/products/riva#demos

REAL-TIME & RELIABLE PERFORMANCE ON DEVICES WITH RIVA EMBEDDED

Intelligent Touchless Kiosks | Medical, Shopping, & Smart Home Assistants | Delivery Robots

High Accuracy with from-data-to-model ASR pipeline customization off device

14 languages support: Arabic, Chinese (Mandarin), English (US/UK), French,
German, Hindi, Italian, Japanese, Korean, Portuguese,
Russian, Spanish (LATAM/Spain)

Female & male **OOTB professional voices** for English deployable immediately on the device & the ability to **create brand voice**

Low latency below 100ms performance for natural conversations

Flexible integration into customer deployments

Privacy with speech data processing on the device

Deployable on NVIDIA Jetson NX & AGX Orin, NX & AGX Xavier

Touchless Kiosk

Shopping Assistant

Theme Park Assistant

NVIDIA RIVA EMBEDDED DEMO

Robot Dog Fetches Snacks Across Town! – Spot

RESOURCES & TIPS FOR GETTING STARTED WITH NVIDIA RIVA ASR

RIVA ASR DEV STARTER KIT

- ASR Docs
- Quick Start Guide
- Using ASR with OOTB Models
- ASR Customization Tutorial
- Contact Center Sample App
- Virtual Assistant with Rasa Sample App

Speech AI technical blogs

- Solving Automatic Speech Recognition Deployment Challenges
  Building a Speech-Enabled AI Virtual Assistant with NVIDIA Riva on Amazon EC2
  Developing the Next Generation of Extended Reality Applications with Speech AI

THANK YOU!