NVIDIA NeMo

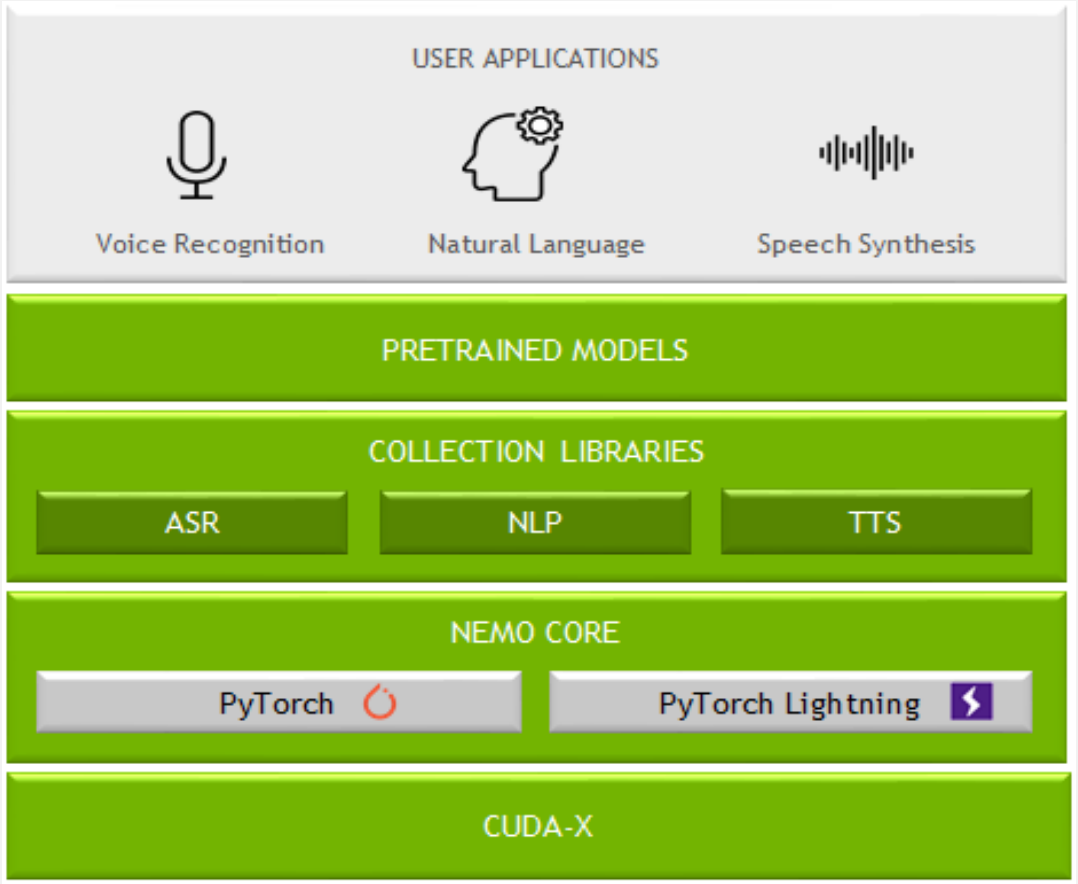
Raman – Solutions Architect

July 2022

NVIDIA NEMO

Toolkit for Building SOTA Conversational Models

- DL-Based Speech & Language Understanding Models
- Include Semantic Checking for Correct-by-Construction Architectures
- Support Expanding Set of Languages:
  - 8 for ASR
  - 5 for NLU
- Open-Sourced
- Integrated with PyTorch & PyTorch Lightning
- Easy-to-Use APIs
- Optimized Training Performance
- 100+ Pre-Trained GPU-Optimized Checkpoints
- Scale to 1000s of NVIDIA GPUs



https://ngc.nvidia.com/catalog/containers/nvidia:nemo

https://github.com/NVIDIA/NeMo

NOTE: ASR – Automatic Speech Recognition | NLU – Natural Language Understanding

HIGHLY ACCURATE PRETRAINED NEMO MODELS AVAILABLE IN NGC

Collection	Task	Accuracy / Performance	Optimized Pre-Trained Models
NLP	Named Entity Recognition (NER)	**74.21%** F1 Score (GMB Test Set)	NER BERT
	Question Answering	**80.22%** Exact Match **83.05%** F1 Score (SQuADv2.0)	BERT-Large Squad 2.0
	Translation	**39.3** WMT13 (SacreBLEU Scores) **35.6** WMT14	NMT En Es Transformer12x2
	Translation	**30.2** WMT14 (SacreBLEU Scores) **46.4** WMT18 **41.1** WMT19 **31.5** WMT20	NMT En De Transformer12x2

NOTE: WER – Word-Error Rate | GMB – Groningen Meaning Bank | WMT – Workshop on Statistical Machine Translation |

NeMo COLLECTIONS

Domain-Specific Collections to Develop Conversational AI Models in Multiple Languages Easily

- Trained ASR, NLP, NMT, TTS pre-trained models on tens and thousands of GPU hours
- Build & train models with 3 lines of code
- Created using Neural Modules - building blocks of NeMo models
- Train and fine-tune models on your domain to understand jargon
- Tightly integrated with PyTorch and Lightning modules
- Optimizes performance for many ASR, NLP and TTS tasks

ASR Support in Multiple Languages

- Catalan
- French
- German
- Italian
- Mandarin
- Polish
- Russian
- Spanish

NLP Support in Multiple Languages

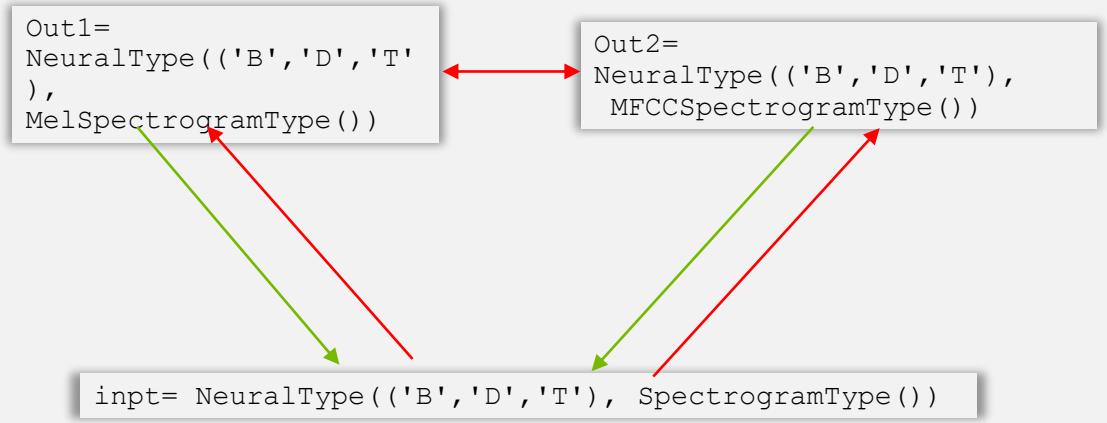
- Spanish
- Russian
- Mandarin
- German
- French

Speech models | NLP models | Translation Models | TTS models

NEURAL TYPES

Strongly Typed Tensors

- Ensure module compatibility with semantic checks
- Simplifies module connections to develop models
- Early mismatches catch - semantic, rank and dimensionality
- Minimize runtime and compile time exceptions
- Easier error debugging



- Green arrow indicates a valid neural type connection
- Red arrow indicates an invalid neural type connection
- Bi-directional red arrows indicate neural types that cannot be interchanged

https://docs.nvidia.com/deeplearning/nemo/user-guide/docs/en/stable/tutorials/neuraltypes.html

MIXED PRECISION & DISTRIBUTED TRAINING

Up to 4.5x Faster Training on Single GPU, Scale to Multiple GPUs Easily

- Tight integration with PyTorch Lightning Trainer to easily invoke training actions.
- Scale to multi-GPU and multi-node to speed-up training while retaining the accuracy
- Speed-up training up to 4.5X on a single GPU with mixed-precision versus FP32 precision
- Ease to use parameters to enable Multi-GPU/node training and mixed-precision

```
trainer = pl.Trainer(**cfg.trainer)
asr_model = EncDecCTCModel(cfg=cfg.model,
trainer=trainer)
trainer.fit(asr_model)
```

Training NeMo model with single line of code

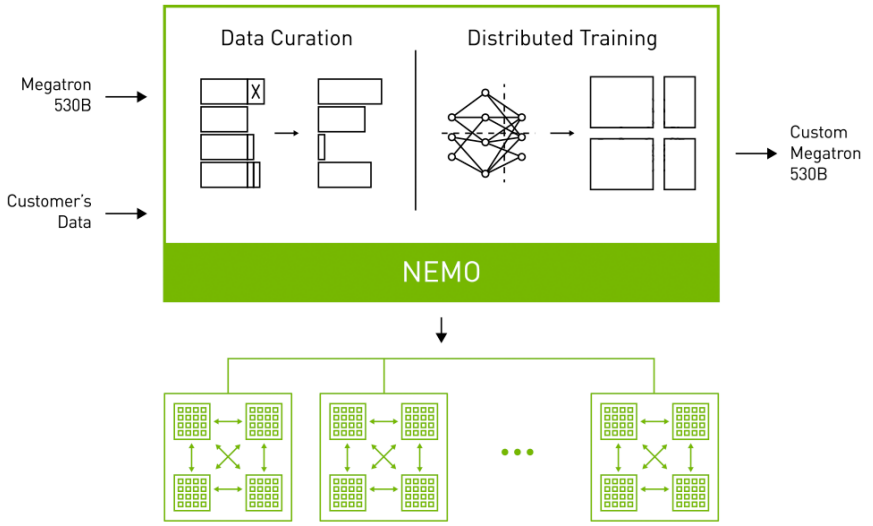
https://docs.nvidia.com/deeplearning/sdk/mixed-precision-training/index.html

NeMo MEGATRON

Accelerated Framework For Training Large Scale NLP Models

- Scale To Models with Trillions of Parameters
- Automated Data Curation for Training
- Pipeline, Tensor & Data Parallelism
- 20B Parameter Model in 1 month on DGX SuperPod
- Optimized for DGX SuperPod

Sign up for Early Access



https://docs.nvidia.com/deeplearning/sdk/mixed-precision-training/index.html

NeMo WITH HYDRA FRAMEWORK

Simplifies Development of Complex Conversational AI Models

- Flexibly configure and customize the model
- Edit end-to-end neural network with single stop solution
- Simple configuration with YAML and CLI

```
#specify name of model
name: &name "QuartzNet15x5"
model:
 sample_rate: &sample_rate 16000
 repeat: &repeat 5
 dropout: &dropout 0.0
 separable: &separable true
 labels: &labels [" ", "a", "b", "c", "d", "e", "f", "g",
"h", "i", "j", "k", "l", "m", "n", "o", "p", "q", "r",
"s", "t", "u", "v", "w", "x", "y", "z", "'"]
#manage training data parameters
 train_ds:
   manifest_filepath: ???
   sample_rate: 16000
   ...
#manage validation data parameters
 validation_ds:
   manifest_filepath: ???
   sample_rate: 16000
   ...

 ...
```

QuartzNet Model Customization with .YAML file
(`quartznet_15x5_aug.yaml`)

DEPLOY TO PRODUCTION

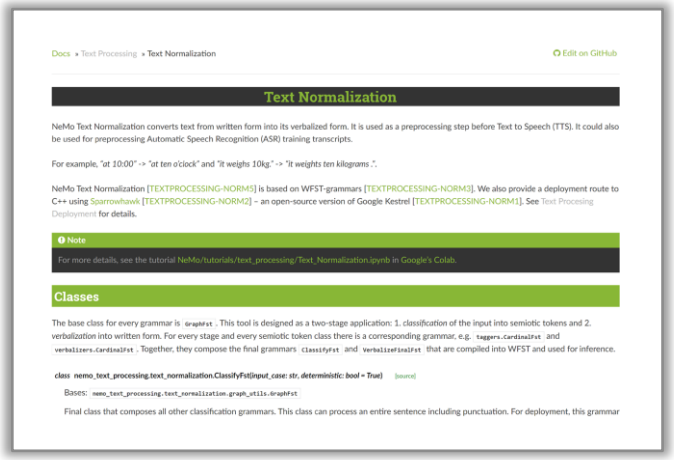
Generate High-Performance Inference

- Quickly export NeMo models to Riva
- Support for speech and language models across multiple languages
- Step-by-step deployment instructions in documentation

Exporting NeMo Models to Riva

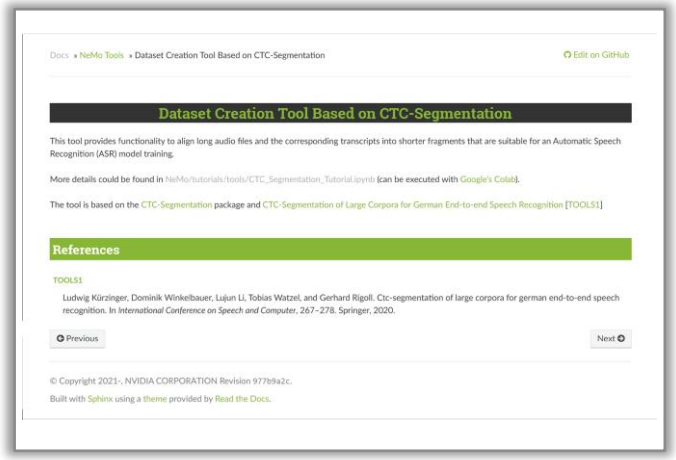
DATA PREPARATION AND EXPLORATION

High Quality Data For Speech Models



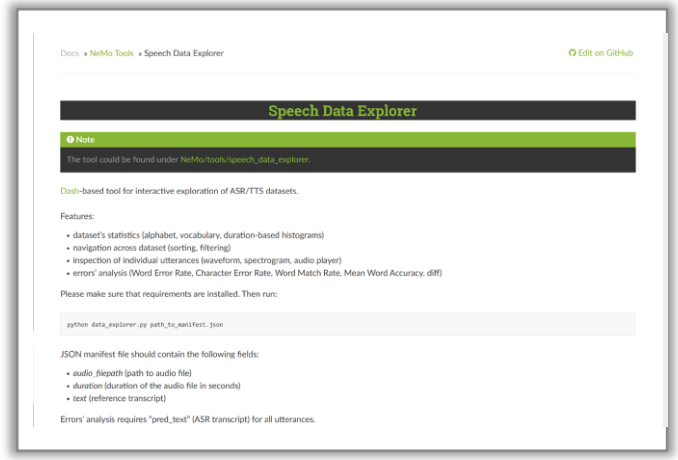
Text Processing

Support for text normalization and denormalization for better readability and accurate ASR and TTS output



Dataset Creation

Chunk long audio files into shorter fragments and align text transcriptions



Data Explorer

Interactive exploration of ASR and TTS datasets (visualizations, error analysis, statistics, sorting, filtering)

IMPROVING CONVERSATIONAL AI IMPROVES CUSTOMER SERVICE

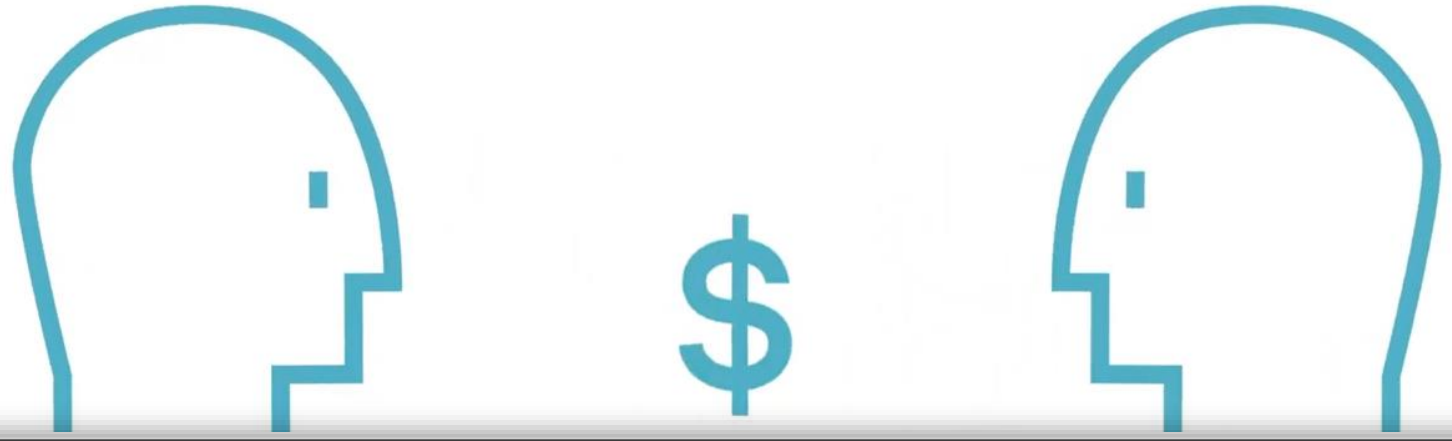
Ping An is one of the largest financial services company in the world, providing service to 200M customers.

The company’s chat-bot agents handle millions of customer queries each day.

Looking to improve the customer experience, Ping An leveraged the NVIDIA NeMo toolkit to improve accuracy and NVIDIA Riva to deliver high throughput and low latency for real-time streaming applications.

With NeMo's pre-trained models and the ASR pipeline optimized using Riva, Ping An’s system achieved a 5% improvement on accuracy out of the gate.

NVIDIA 平安科技 PING AN TECHNOLOGY



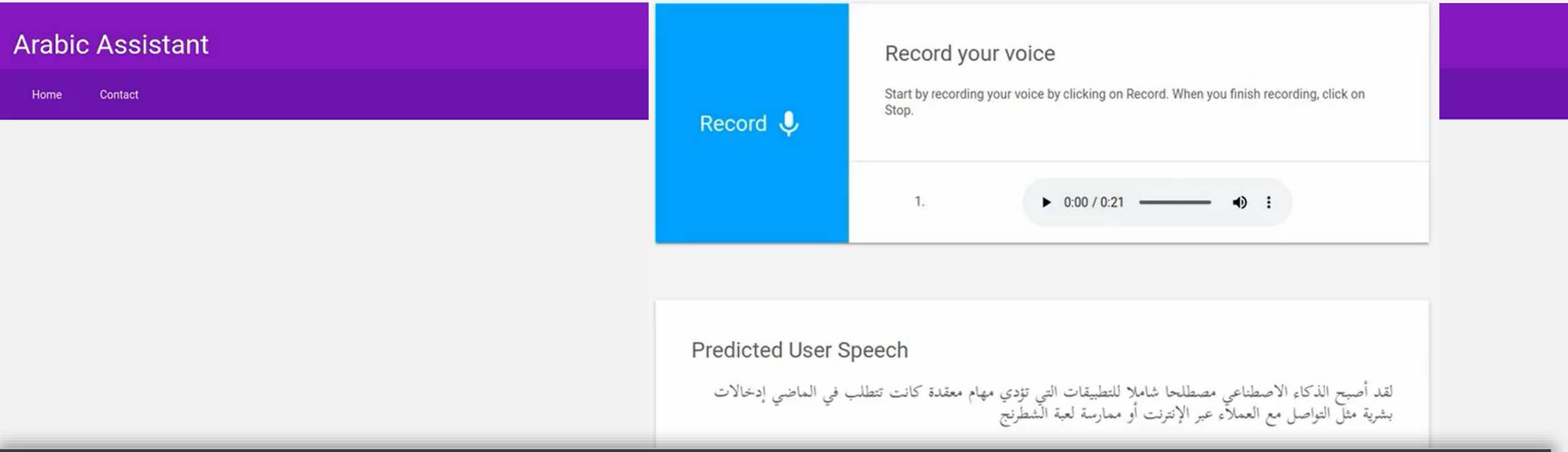
STATE-OF-THE-ART SPEECH RECOGNITION FOR FINANCIAL AUDIO

When commercial ASR solutions weren't meeting the quality needs of S&P Global, the finance industry's premier provider of data and solutions, the company turned to its innovation hub, Kensho.

Kensho developed Scribe, an end-to-end speech recognition solution specifically optimized for the finance industry using NVIDIA NeMo on NVIDIA V100 Tensor Core GPUs. Scribe processes more audio in less time, and with better accuracy.

Kensho trained Scribe to systematically transcribe tens of thousands of earnings calls, management presentations, and acquisition calls each year — improving accuracy and enabling S&P Global to increase earnings call coverage by more than 25%.

NVIDIA. KENSHO



BUILDING VIRTUAL ASSISTANTS FOR UNDERREPRESENTED LANGUAGES

Arabic is the 5th most spoken language worldwide, offering an enormous market opportunity for voice assistants but there are many challenges with the Arabic language — such as the lack of capitalization, letters take multiple shapes according to where they occur in a word, and Arabic is written from right to left.

InstaDeep built an Arabic out-of-the-box model using NVIDIA NeMo, NVIDIA Riva, and the NVIDIA DGX system.

NeMo helped InstaDeep quickly finetune state-of-the-art models on an Arabic dataset using simple APIs.

InstaDeep gained WER as low as 7.84% for ASR models on Arabic dataset, reduced the WER by 4x by fine-tuning the NeMo model vs training from scratch, and reduced training time from days to hours using NVIDIA DGX.

NVIDIA. InstaDeep™

EASY-TO-USE APIS LOWER DEVELOPMENT COSTS

MeetKai builds virtual assistants that make peoples' lives easier.

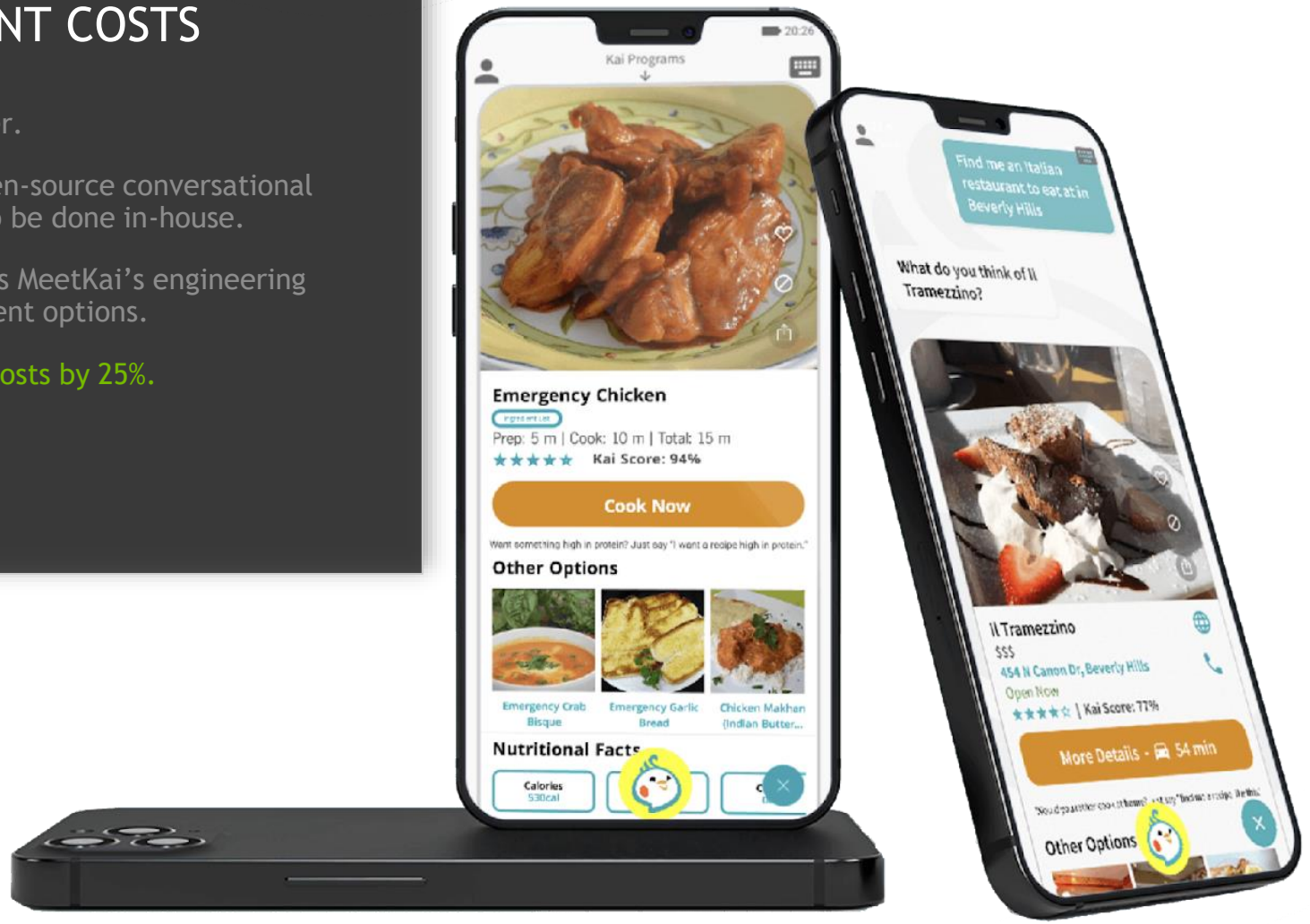
In 2018 when the company began, the lack of high-quality, open-source conversational AI toolkits was a challenge. All machine learning tooling had to be done in-house.

NVIDIA NeMo integrated with Riva is the first toolkit that meets MeetKai's engineering requirements for full life cycle — from R&D to hybrid deployment options.

With NeMo's easy-to-use API's, MeetKai reduced development costs by 25%.

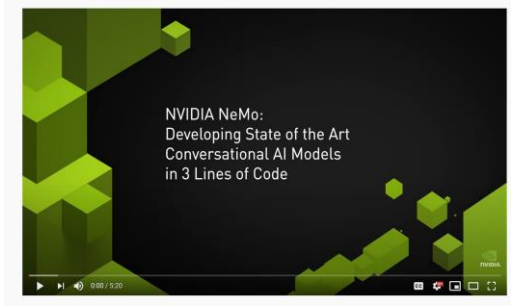
NVIDIA

meetKai



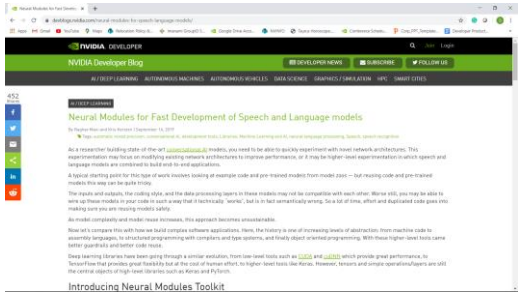
RESOURCES TO GET STARTED WITH NVIDIA NEMO

Download NeMo Today

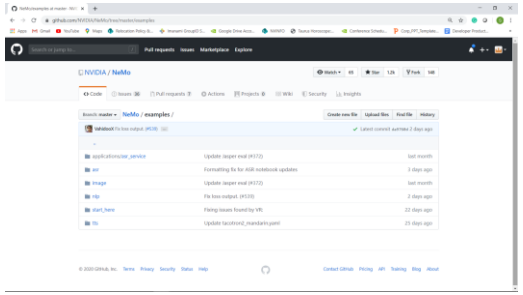


NeMo Introductory Video

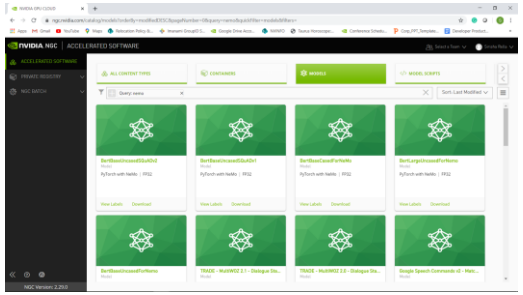
NeMo overview



Introductory Blog

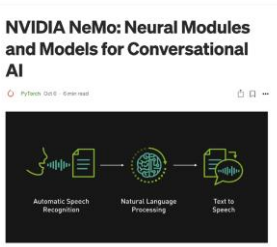


Examples

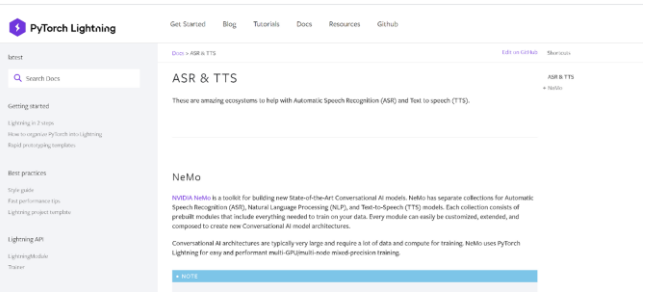


Pre-Trained Models

NVIDIA NeMo product page



NeMo + PyTorch joint blog



NeMo + PyTorch Lightning Joint Documentation

Download NeMo container from NGC: https://ngc.nvidia.com/catalog/containers/nvidia:nemo
Download NeMo with pip install pip install nemo_toolkit['all']