
AWS
re:Invent

AIM409

Enhance speech recognition accuracy using custom language models

Paul Zhao
Principal Product Manager
Amazon Transcribe
AWS

Dr. Javier Asensio-Cubero
Lead of Data Platform
Octopus Energy

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agenda

Overview of Amazon Transcribe

Customer benefits

Main use cases

How to improve ASR accuracy with Amazon Transcribe

Case study: Octopus Energy

Demo

Next steps

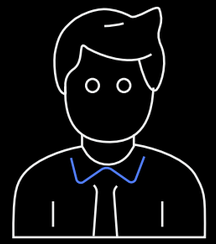


© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Audio and video data is growing exponentially

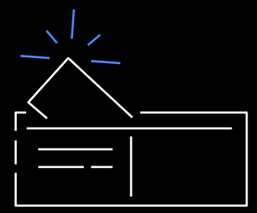
Traditional transcription services



Manual



Slow



Expensive

Other automatic speech recognition (ASR) vendors



Hard to use

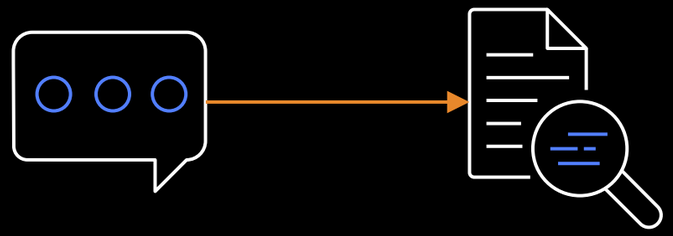


Limited performance



Lacking features

Amazon Transcribe



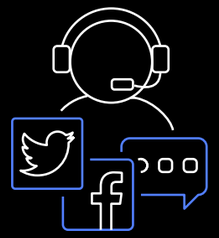
Automatic speech recognition (ASR)

Fully managed and continuously trained

Accurate and fast performance

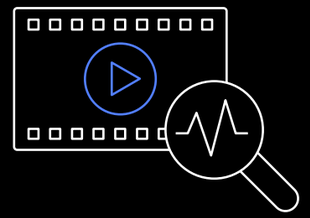
Designed for high availability and scale

Use cases



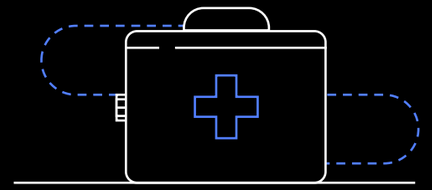
Contact center analytics

- Call driver analysis
- Agent assist
- Customer satisfaction



Media & entertainment

- Metadata indexing
- Video subtitling
- Game chat moderation



Medical transcription

- Clinical documentation
- Pharmacovigilance
- Telehealth

Key features



Punctuation, capitalization



Speaker labeling / channel identification



Word-level time stamps



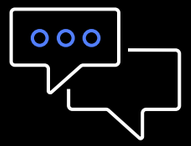
Word-level confidence scores



Language ID



Custom vocabulary



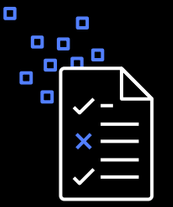
Custom language models (CLM)



Alternative transcriptions

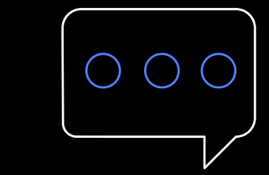


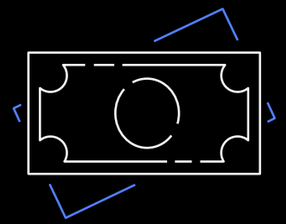
Vocabulary filtering



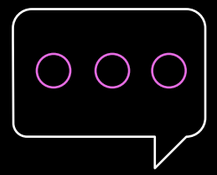
PCI / PII content redaction

Diverse transcription use cases



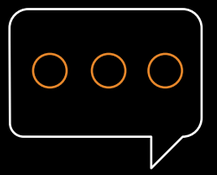


Finance





Law

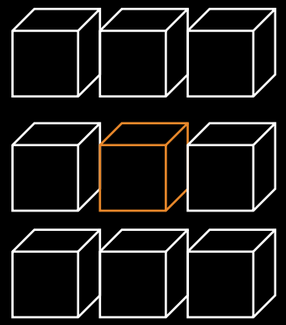




Real Estate

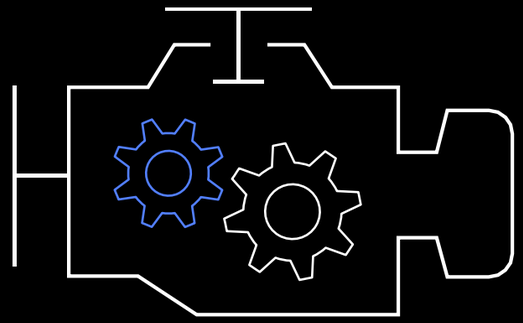
Custom language models (CLM)

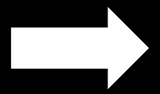
Your data



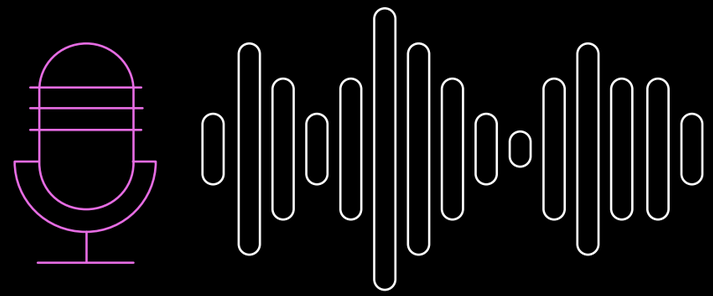
+

Our training engine

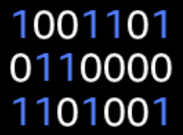




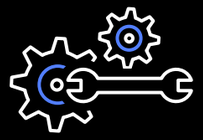
Custom speech model



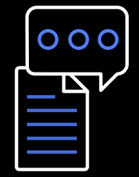
Custom language models (CLM)



Prepare training data



Train a custom language model (CLM)



Transcribe an audio sample using CLM



Evaluate CLM transcription performance

Custom language models (CLM)



Video Game Domain audio sample
https://bit.ly/36kyMfy

Reference transcript (human-annotated)

The 2020 holiday season is right around the corner. And with the way that the year's been going, we can all hope for a little excitement around the next-gen video game consoles coming out soon. So, what's the difference in **hardware specs** between the upcoming Playstation 5 and Xbox Series X? Well, let's take a look under the hoods of each next-gen gaming **console**. The PS5 features **an AMD Zen 2 CPU** with up to 3.5 GHz frequency. It sports **an AMD Radeon** GPU that **touts** 10.3 teraflops, running up to 2.23 GHz. Memory and storage respectively **dial in** at 16 GB and 825 GB. The PS5 supports both PS The PS5 supports both 4K and **8K resolution** screens. Meanwhile, the Xbox Series X also features an **AMD Zen 2 CPU**, but clocks in at **3.8 GHz** instead. **The console boasts a similar AMD custom** GPU with 12 teraflops and 1.825 GHz. Memory is the same as that of the PS5's, **coming in at** 16 GB. But the default storage is where the system has an edge, bringing out a massive 1 TB hard drive. Like the PS5, the **Series X** also supports 4K and 8K resolution screens as well. Those of course, are just the numbers. Therefore, it remains to be seen exactly how the performance plays out in practice. It's worth noting that both systems have incorporated ray-tracing technology, something that's used to make light and shadows look

Machine transcript (generic speech engine)

the 2020 holiday season is right around the corner. And with the way that the years been going, we can all hope for a little excitement around the next Gen video game consoles coming out soon. So what's the difference in **heart respects** between the upcoming PlayStation five and Xbox Series X? Well, let's take a look under the hood of each of these **Consul's**. The PS five features in **a M descend to CPU** with up to 3.5 gigahertz frequency is sports **and AM the radio on** GPU that **tells** 10.3 teraflops running up to 2.23 gigahertz memory and storage, respectively. **Dahlin** at 16 gigabytes in 825 gigabytes. The PS five supports both PS. The PS five supports both four K and **A K resolutions**. Meanwhile, the Xbox Series X also features an **AM descend to CPU**, but clocks in at **three point take. It hurts** instead, **the cost. Almost a similar AMG custom** GPU, with 12 teraflops and 1.825 gigahertz memory, is the same as out of the PS. Fives **come in at** 16 gigabytes, but the default storage is where the system has an edge. Bring out a massive one terabyte hard drive. Like the PS five. The **Siri's X** also supports four K and eight K resolution screens as well. Those, of course, are just the numbers. Therefore, it remains to be seen exactly how the

How does CLM work?

PART 1 – DATA PREP

1. Train CLM using text data that represents your business domain and operation

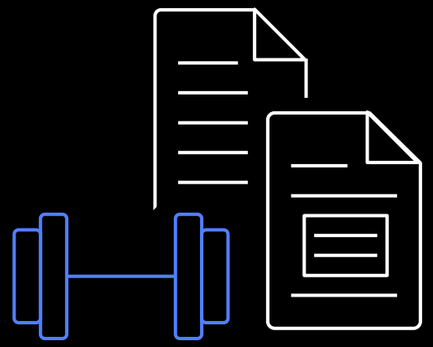
   Examples:
   - Training / instruction manuals
   - Website content
   - Brochures
   - Textbooks
   - Business emails and correspondence
   - Existing human-annotated (reference transcripts) of audio

2. No acoustic data required

How does CLM work?

PART 1 – DATA PREP

Types of data



Training data

- *Domain-specific*
- *Required*



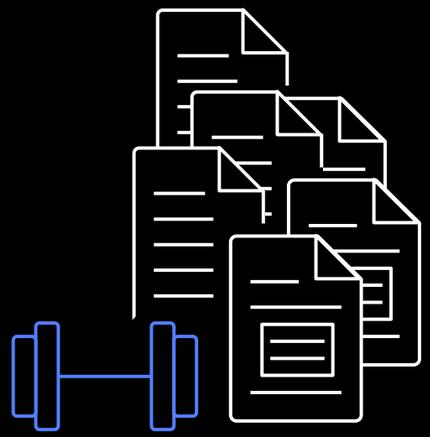
Tuning data

- *Use case–specific*
- *Optional*

How does CLM work?

PART 1 – DATA PREP

Scenario 1



Training data

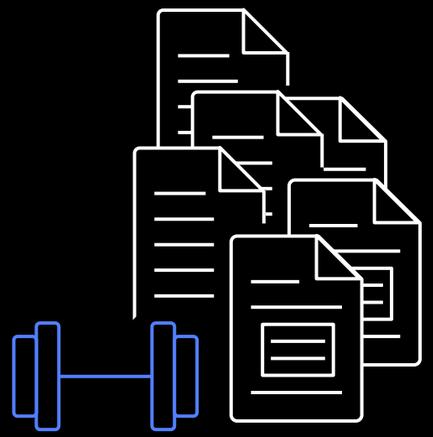


Tuning data

Use them separately

OR

Scenario 2



Training data

+



Tuning data

Combine into one large training set

How does CLM work?

PART 1 – DATA PREP



Amazon S3

Data guidelines

- Plain text
- One sentence per line
- UTF-8 encoding
- No formatting characters
- Training data file size <= 2 GB
- Tuning data file size <=200 MB

The PS5 will feature a custom eight-core AMD Zen 2 CPU clocked at 3.5GHz (variable frequency) and a custom GPU based on AMD's RDNA 2 architecture hardware that promises 10.28 teraflops and 36 compute units clocked at 2.23GHz (also variable frequency).

It'll also have 16GB of GDDR6 RAM and a custom 825GB SSD that Sony has previously promised will offer super-fast loading times in gameplay, via Eurogamer.

One of the biggest technical updates in the PS5 was already announced last year: a switch to SSD storage for the console's main hard drive, which Sony says will result in dramatically faster load times.

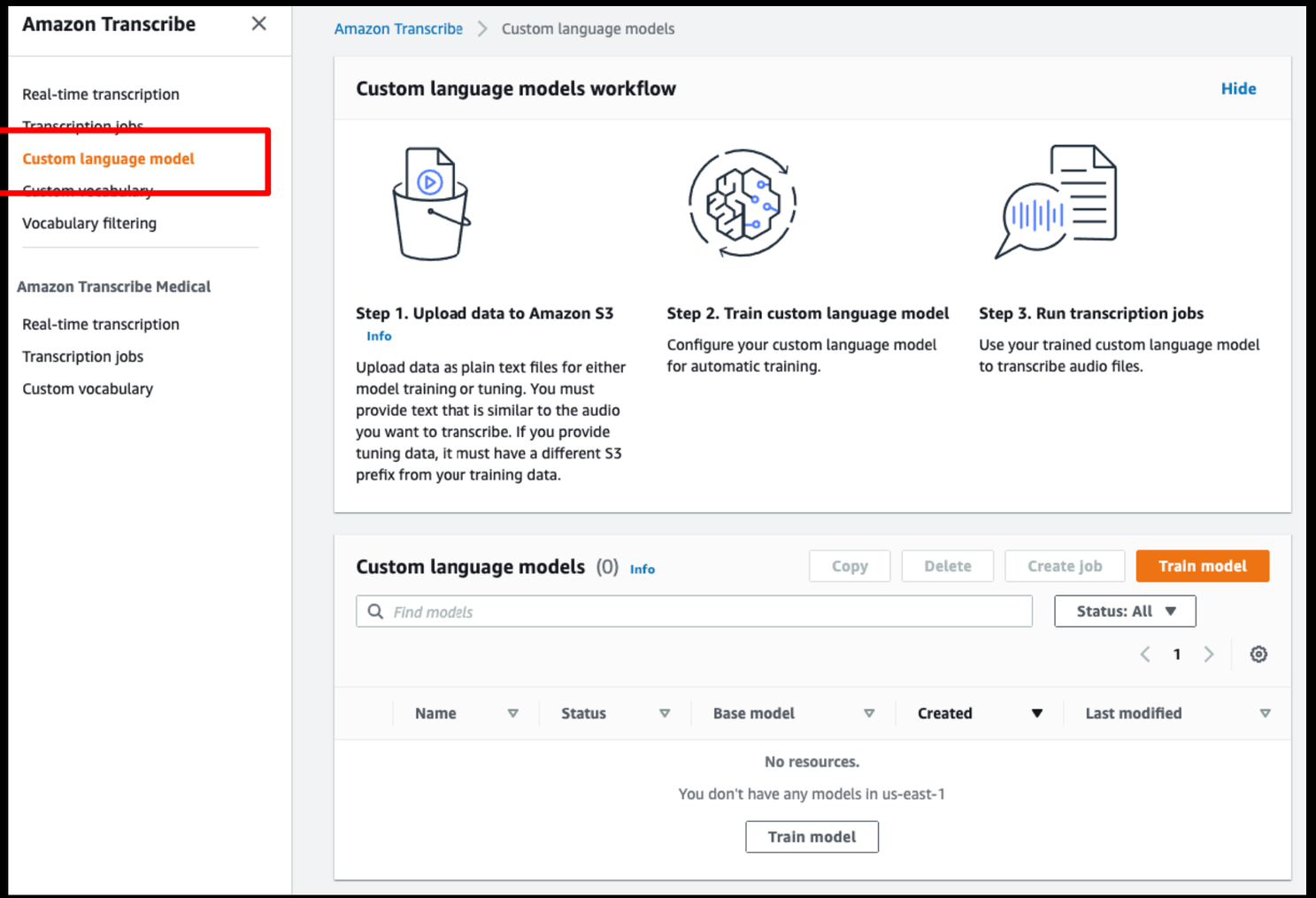
A previous demo showed Spider-Man loading levels in less than a second on the PS5, compared to the roughly eight seconds it took on a PS4.PlayStation hardware lead Mark Cerny dove into some of the details about those SSD goals at the announcement.

Where it took a PS4 around 20 seconds to load a single gigabyte of data, the goal with the PS5's SSD was

How does CLM work?

PART 2 – TRAINING

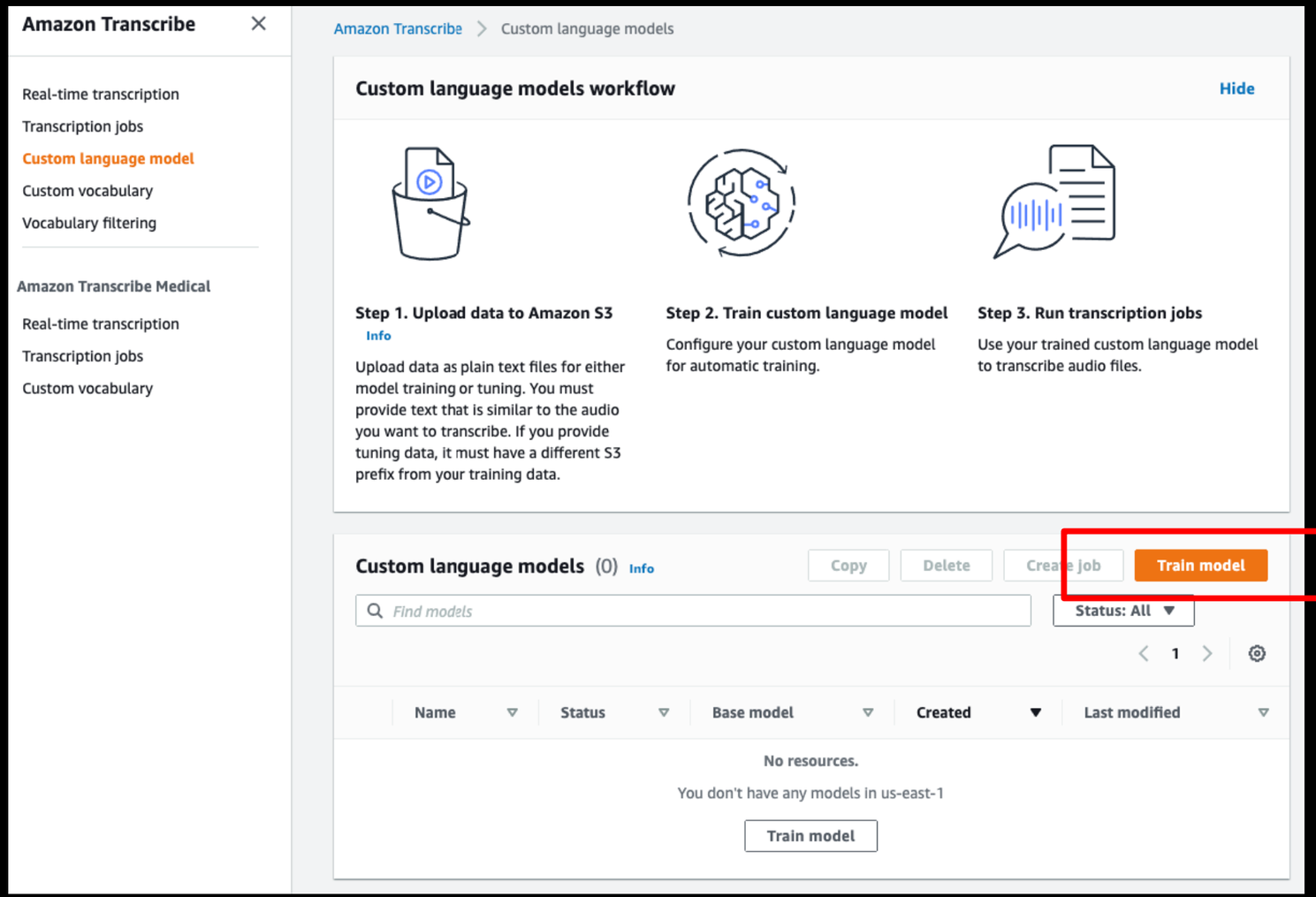
1. Amazon Transcribe service console
2. Custom language models (CLM)



How does CLM work?

PART 2 – TRAINING

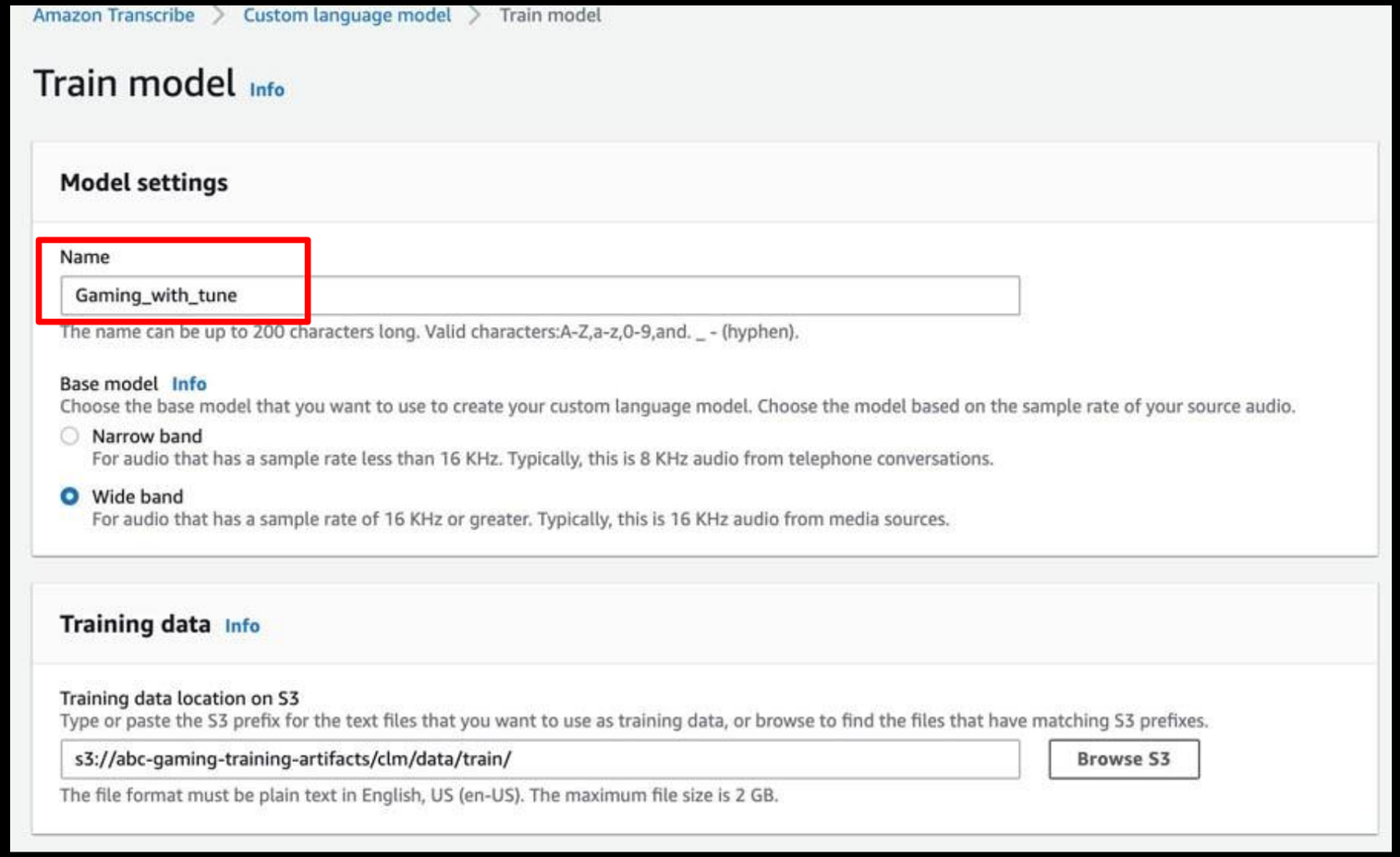
1. Amazon Transcribe service console
2. Custom language models (CLM)
3. Train model



How does CLM work?

PART 2 – TRAINING

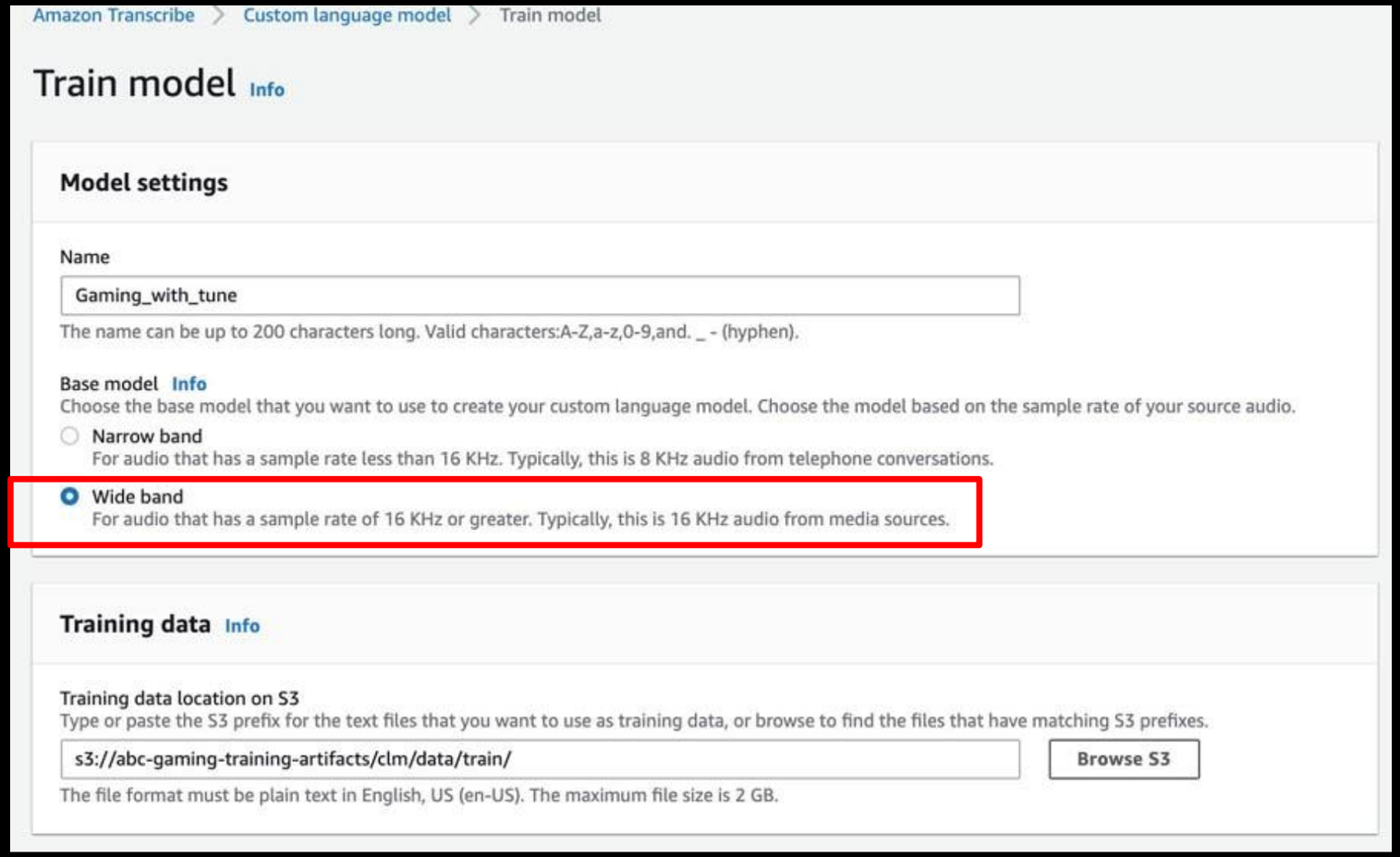
1. Amazon Transcribe service console
2. Custom language models (CLM)
3. Train model
4. Name my model



How does CLM work?

PART 2 – TRAINING

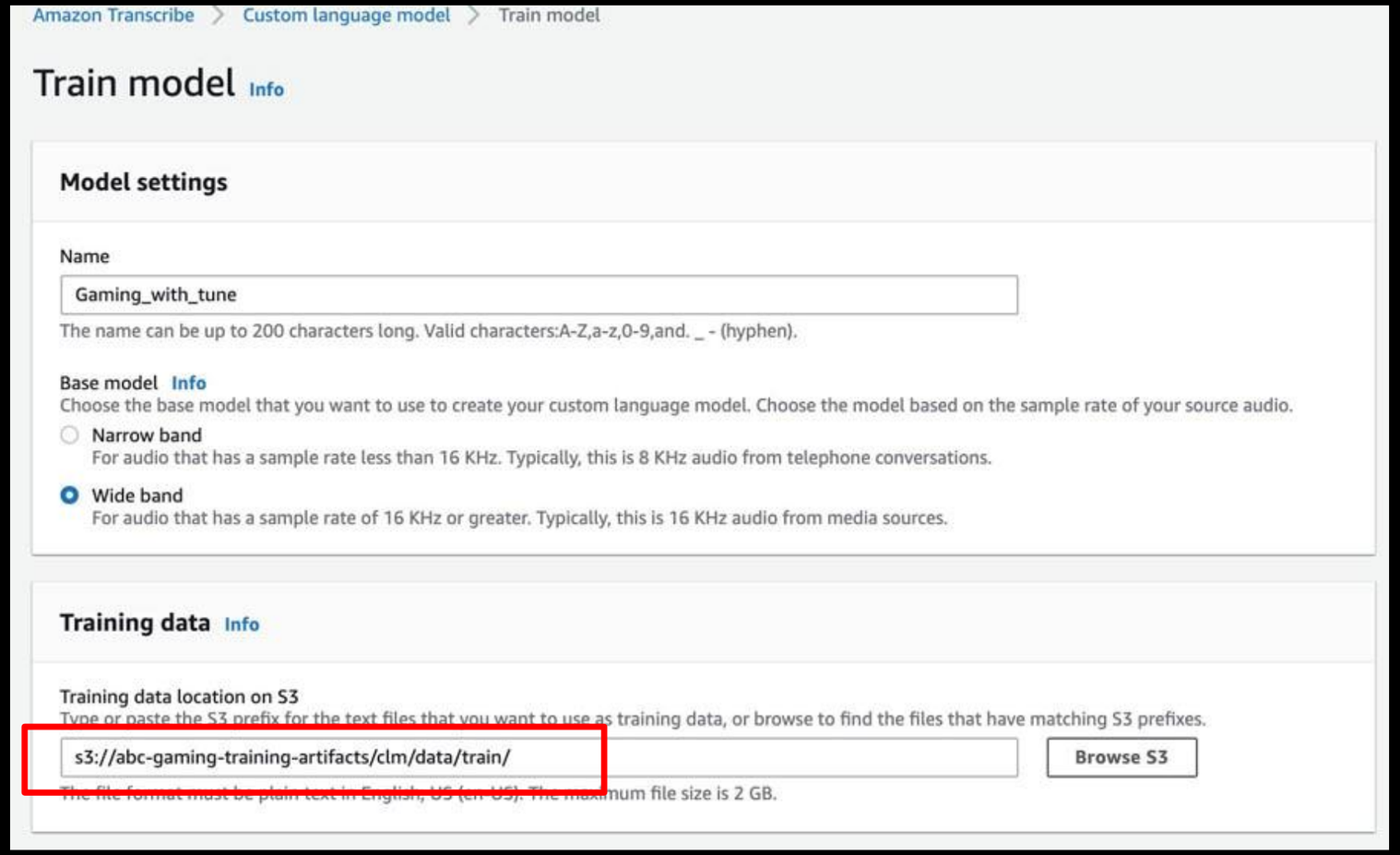
1. Amazon Transcribe service console
2. Custom language models (CLM)
3. Train model
4. Name my model
5. Base model: Wide band



How does CLM work?

PART 2 – TRAINING

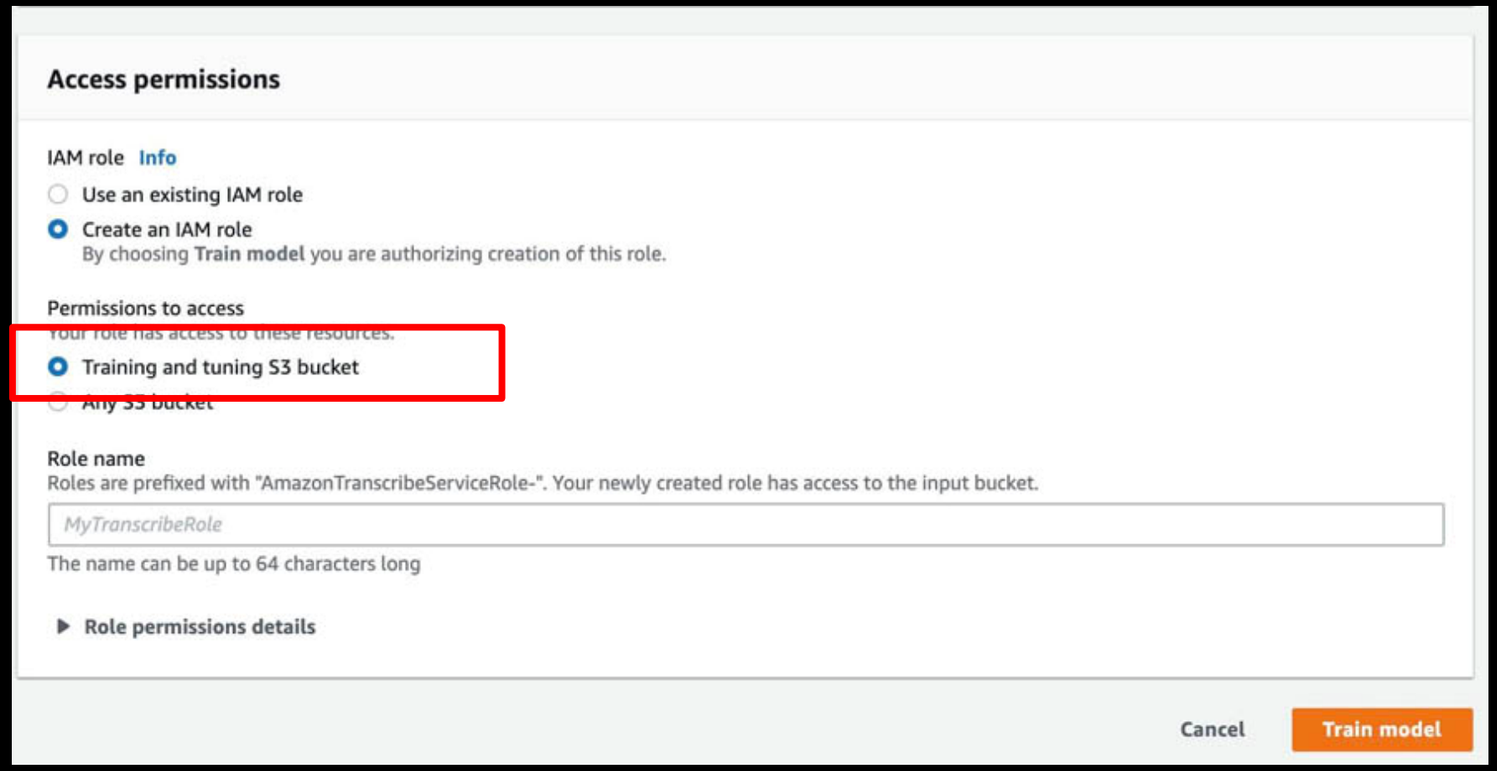
1. Amazon Transcribe service console
2. Custom language models (CLM)
3. Train model
4. Name my model
5. Base model: Wide band
6. Enter training data S3 bucket



How does CLM work?

PART 2 – TRAINING

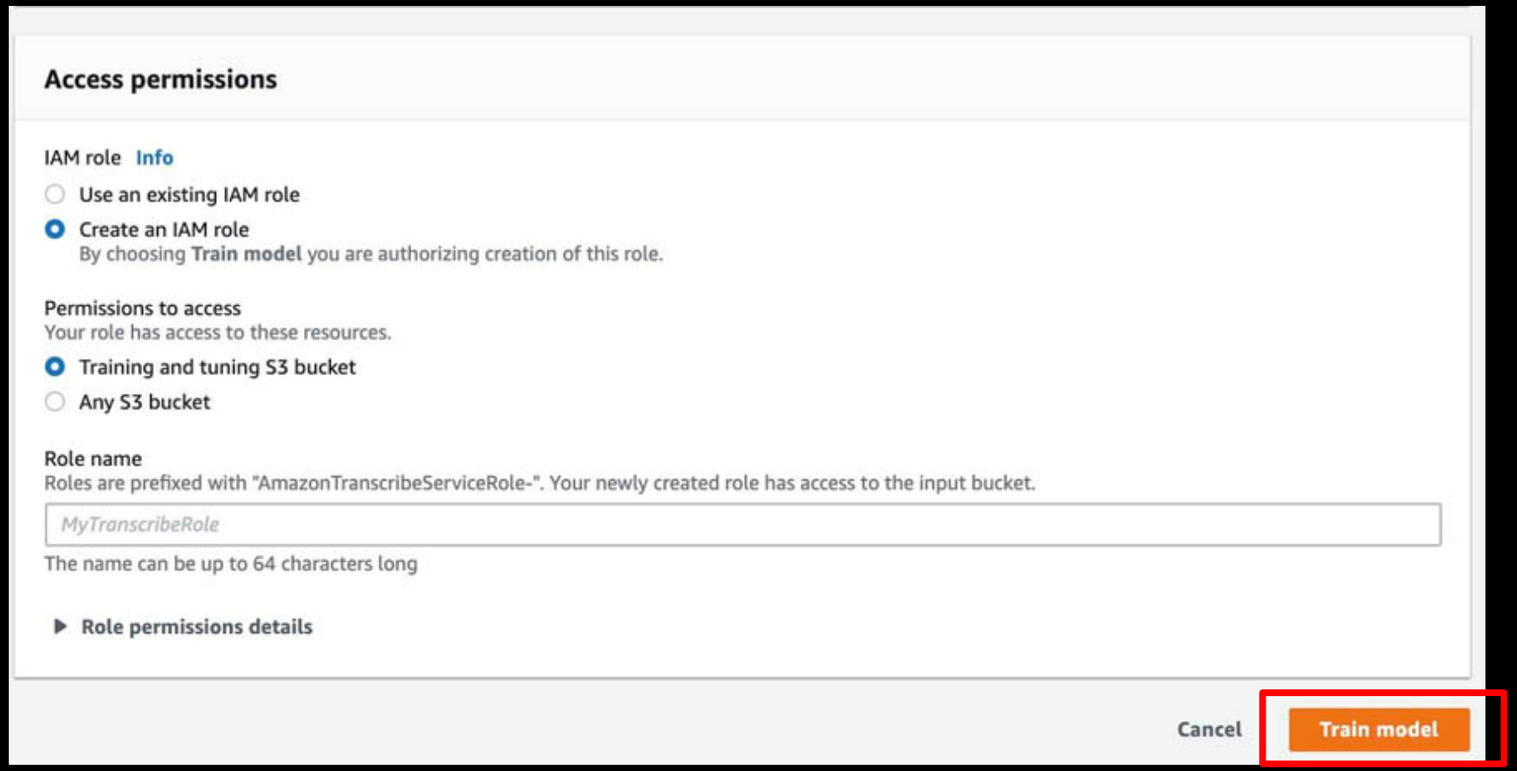
1. Amazon Transcribe service console
2. Custom language models (CLM)
3. Train model
4. Name my model
5. Base model: Wide band
6. Enter training data S3 bucket
7. Permissions: Training and tuning



How does CLM work?

PART 2 – TRAINING

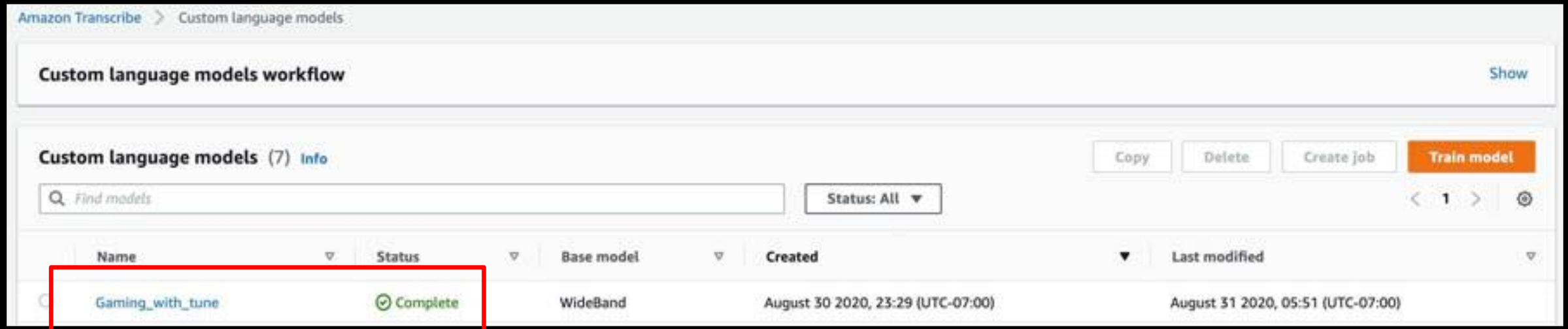
1. Amazon Transcribe service console
2. Custom language models (CLM)
3. Train model
4. Name my model
5. Base model: Wide band
6. Enter training data S3 bucket
7. Permissions: Training and tuning
8. Click “Train model”



How does CLM work?

PART 2 – TRAINING

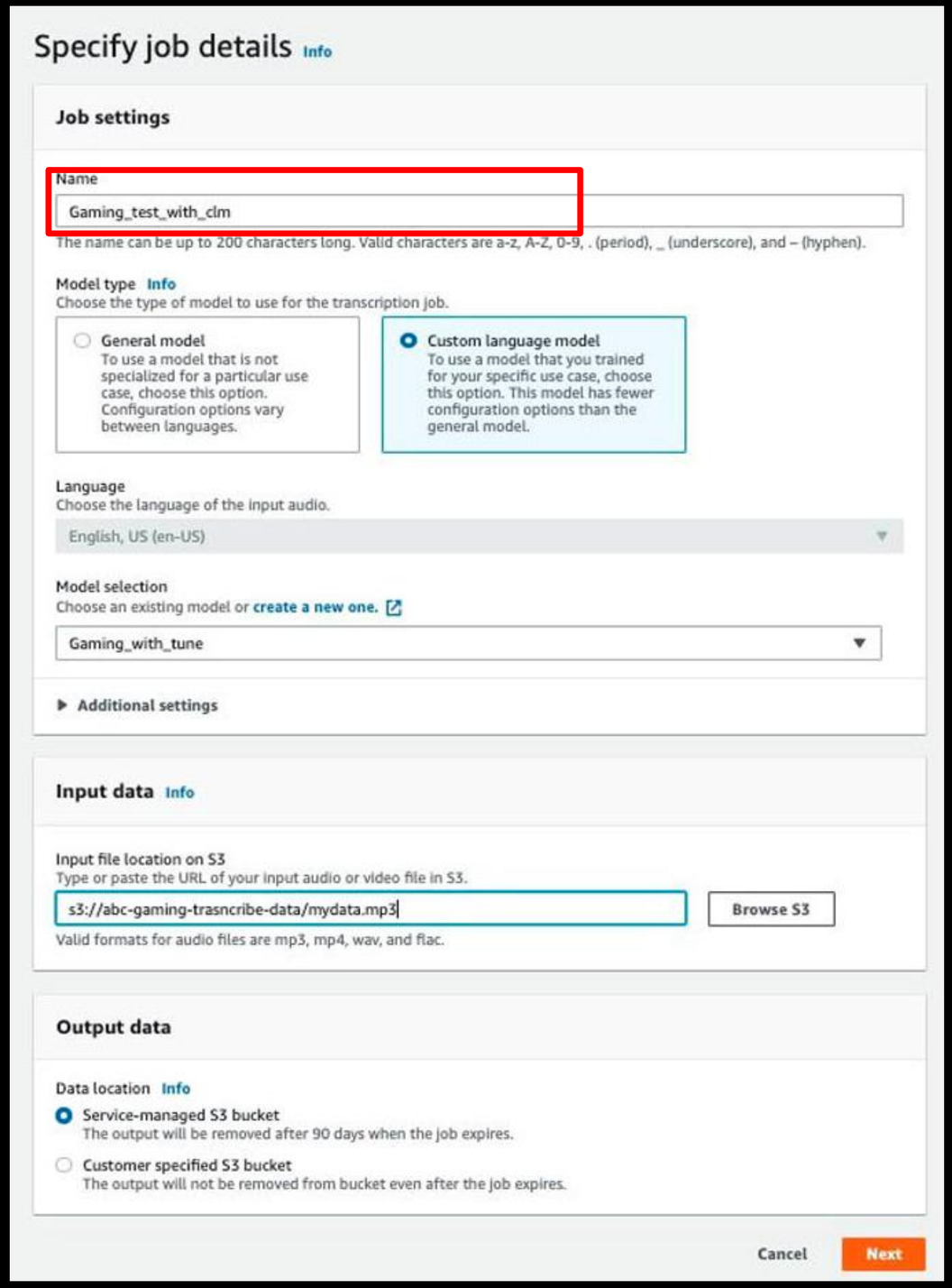
Back to Amazon Transcribe console > Custom language models



How does CLM work?

PART 3 – TRANSCRIBING AUDIO

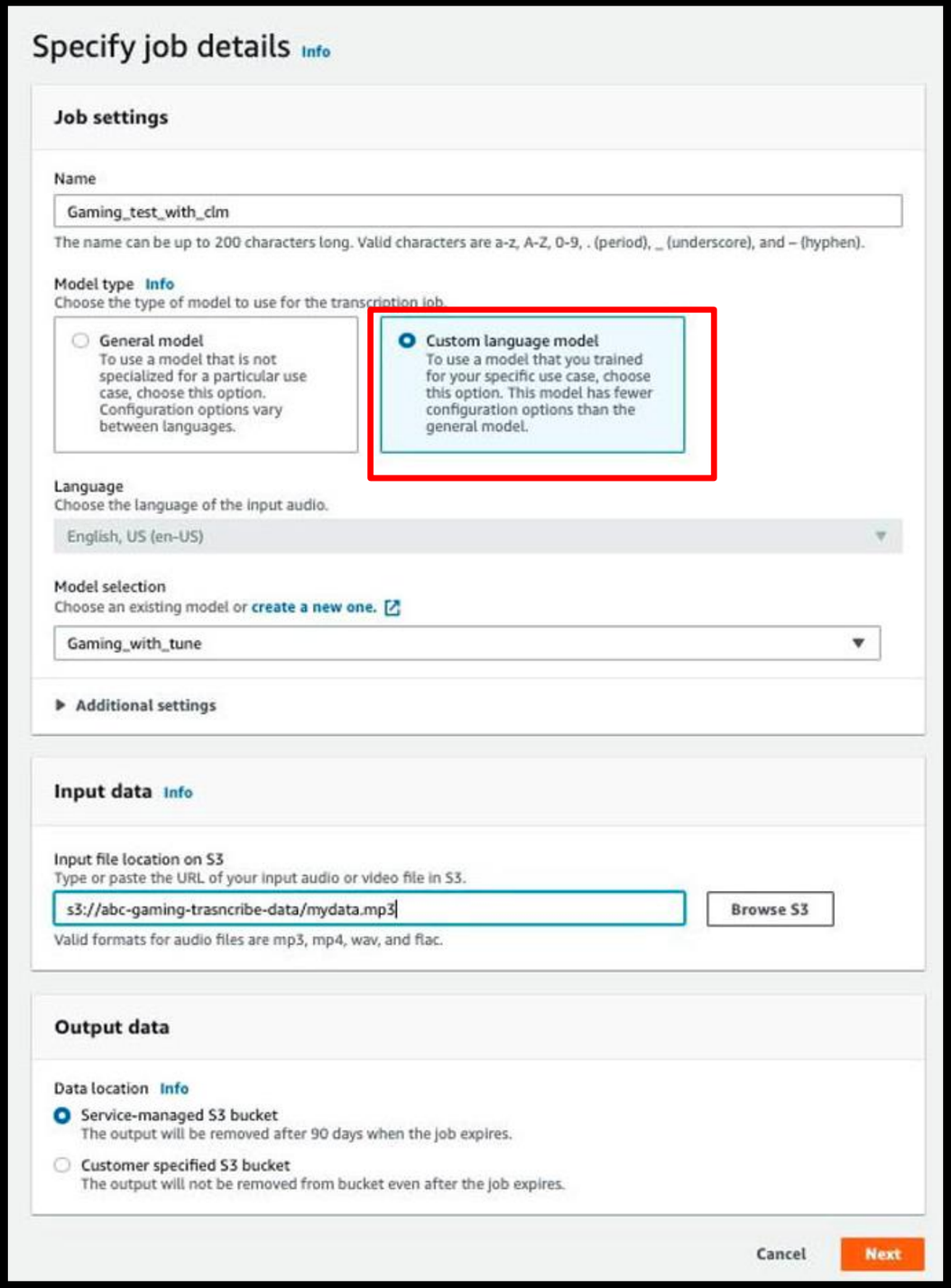
1. Start and name a new transcription job
2. Select the CLM instead of general model
3. If I have tuning data, I can select that
4. Point to the S3 location of audio file
5. Select my transcript output location
6. Then select “Next” to start transcribing



How does CLM work?

PART 3 – TRANSCRIBING AUDIO

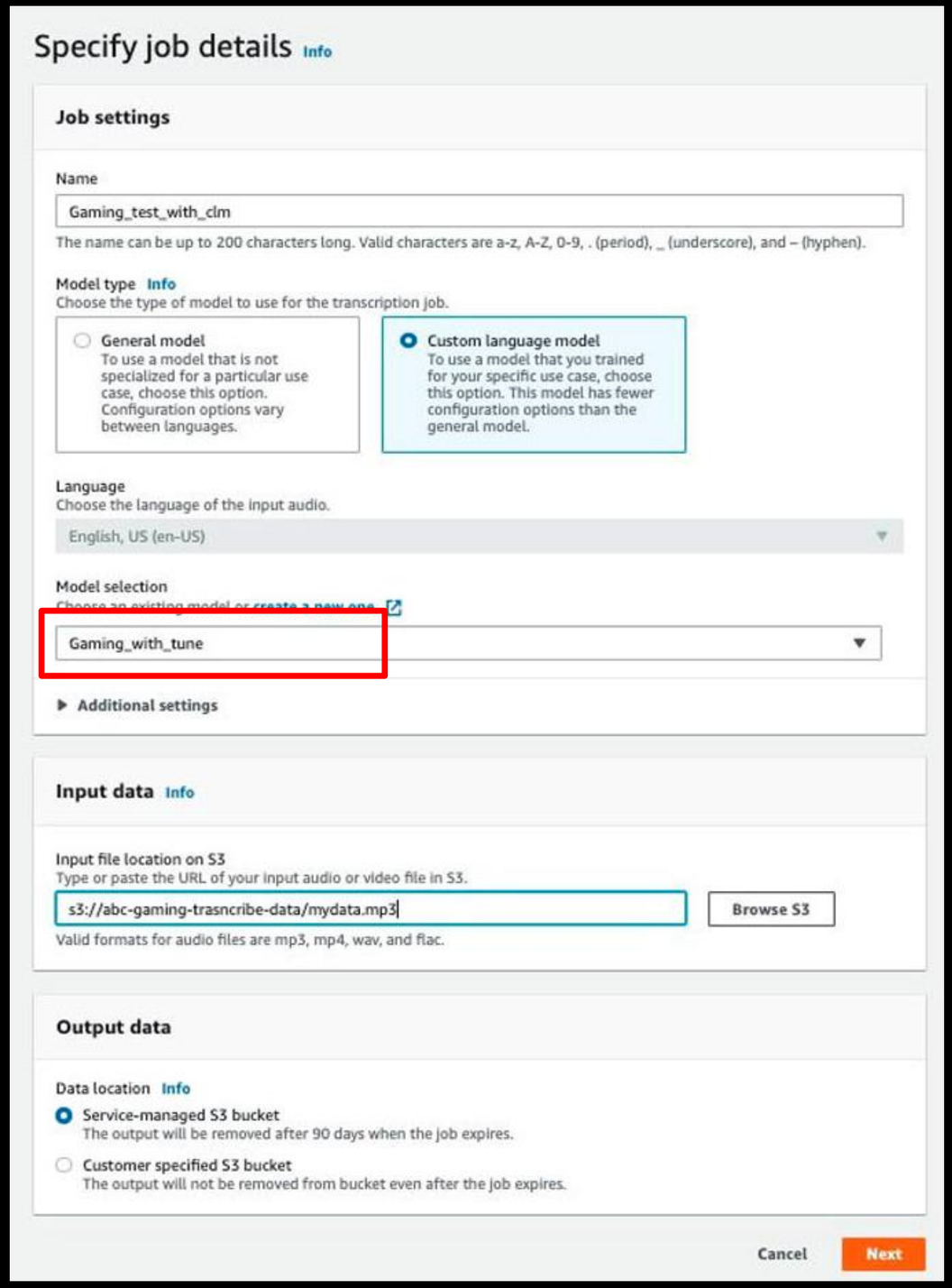
1. Start and name a new transcription job
2. Select the CLM instead of general model
3. If I have tuning data, I can select that
4. Point to the S3 location of audio file
5. Select my transcript output location
6. Then select “Next” to start transcribing



How does CLM work?

PART 3 – TRANSCRIBING AUDIO

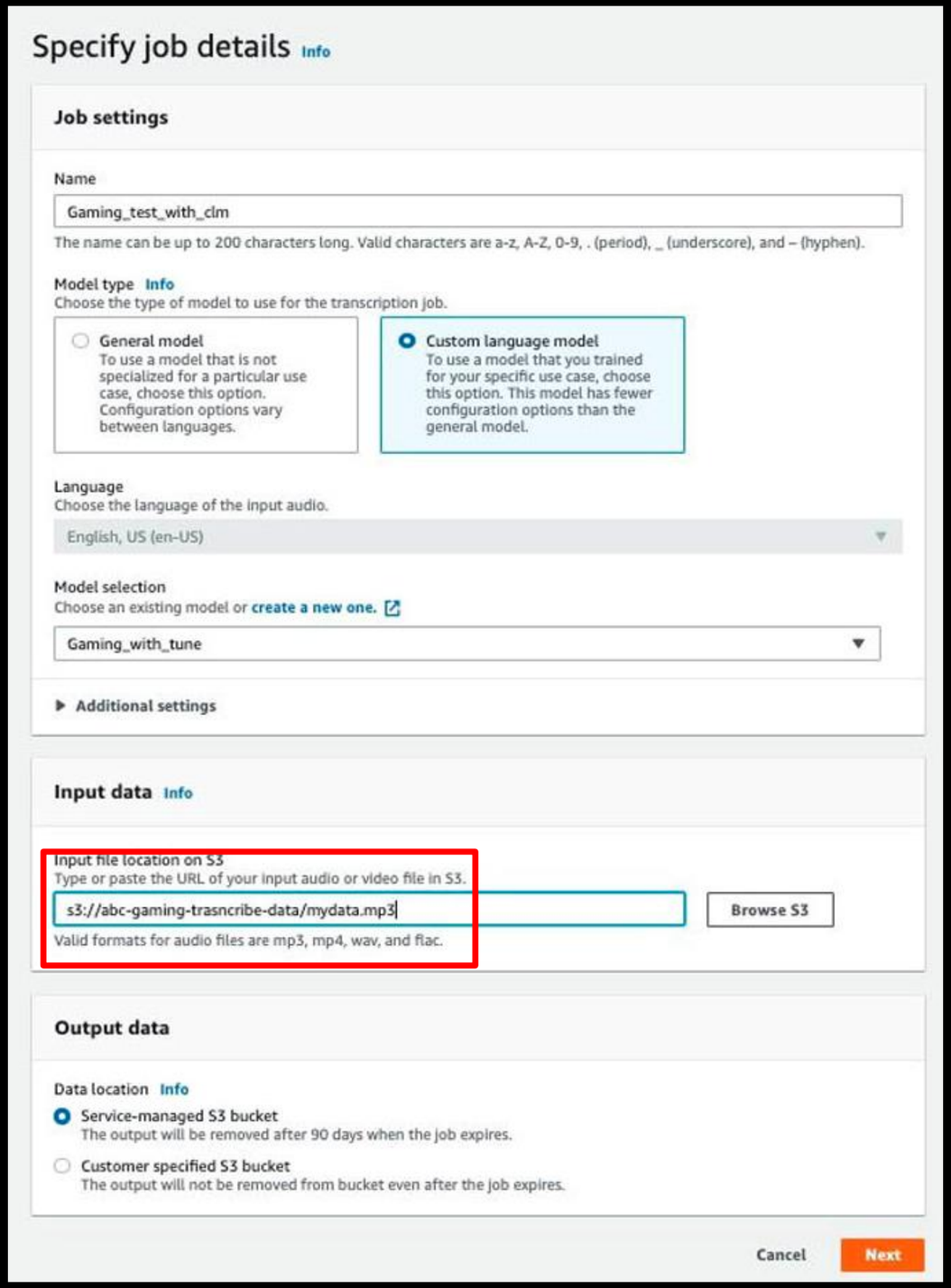
1. Start and name a new transcription job
2. Select the CLM instead of general model
3. If I have tuning data, I can select that
4. Point to the S3 location of audio file
5. Select my transcript output location
6. Then select “Next” to start transcribing



How does CLM work?

PART 3 – TRANSCRIBING AUDIO

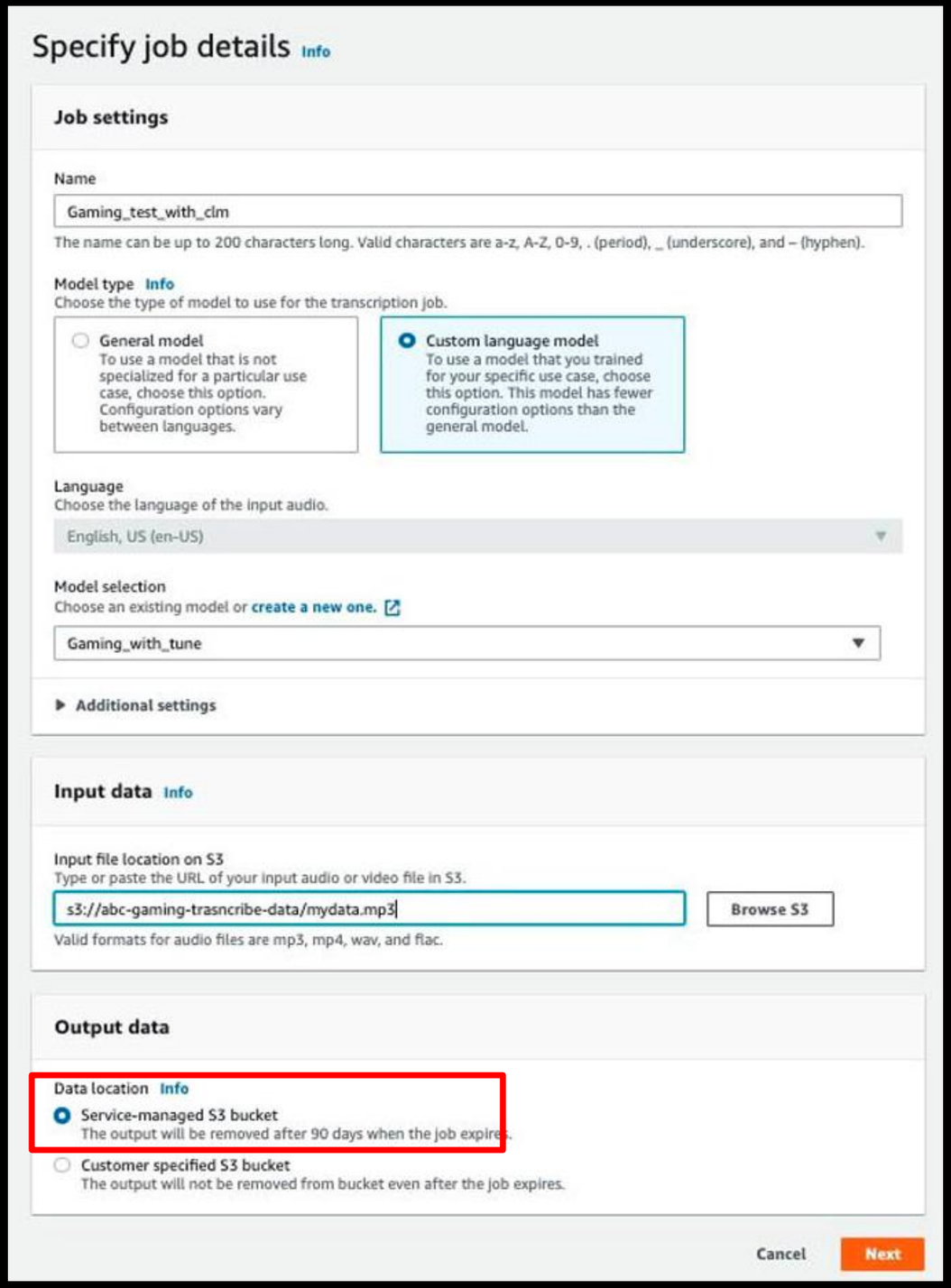
1. Start and name a new transcription job
2. Select the CLM instead of general model
3. If I have tuning data, I can select that
4. Point to the S3 location of audio file
5. Select my transcript output location
6. Then select “Next” to start transcribing



How does CLM work?

PART 3 – TRANSCRIBING AUDIO

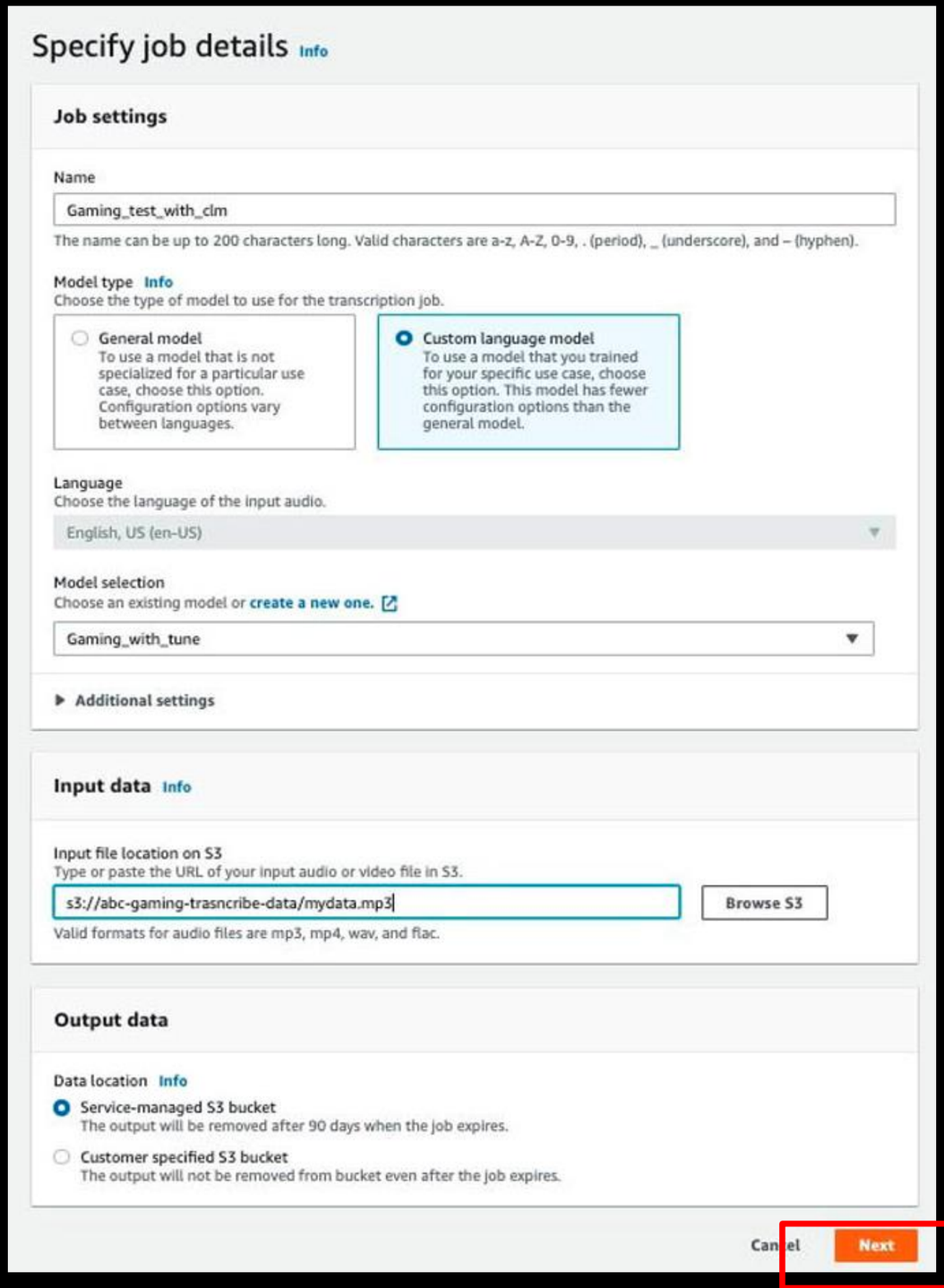
1. Start and name a new transcription job
2. Select the CLM instead of general model
3. If I have tuning data, I can select that
4. Point to the S3 location of audio file
5. Select my transcript output location
6. Then select “Next” to start transcribing



How does CLM work?

PART 3 – TRANSCRIBING AUDIO

1. Start and name a new transcription job
2. Select the CLM instead of general model
3. If I have tuning data, I can select that
4. Point to the S3 location of audio file
5. Select my transcript output location
6. Then select “Next” to start transcribing



How does CLM work?

PART 4 – EVALUATE CLM PERFORMANCE

Reference transcript

all hope for a little excitement around the next-gen video game consoles coming out soon. So, what's the difference in **hardware specs** between the upcoming Playstation 5 and Xbox Series X? Well, let's take a look under the hoods of each next-gen gaming **console**. The PS5 features an **AMD Zen 2 CPU** with up to 3.5 GHz frequency. It sports **an AMD Radeon** GPU that **touts** 10.3 teraflops, running up to 2.23 GHz. Memory and storage respectively **dial in** at 16 GB and 825 GB. The PS5 supports both PS The PS5 supports both 4K and **8K resolution** screens. Meanwhile, the Xbox Series X also features an **AMD Zen 2 CPU**, but clocks in at **3.8 GHz** instead. **The console boasts a similar AMD custom** GPU with 12 teraflops and 1.825 GHz. Memory is the same as that of the PS5's, **coming in at** 16 GB. But the default storage is where the system has an edge, bringing out a massive 1 TB hard drive. Like the PS5, the **Series X** also supports 4K and 8K resolution screens as well. Those of course, are just the numbers. Therefore, it remains to be exactly how the performance in practice with noting the

Generic model transcript

all hope for a little excitement around the next Gen video game consoles coming out soon. So what's the difference in **heart respects** between the upcoming PlayStation five and Xbox Series X? Well, let's take a look under the hood of each of these **Consul's**. The PS five features in **a M descend to CPU** with up to 3.5 gigahertz frequency is sports **and AM the radio on** GPU that **tells** 10.3 teraflops running up to 2.23 gigahertz memory and storage, respectively. **Dahlin** at 16 gigabytes in 825 gigabytes. The PS five supports both PS. The PS five supports both four K and **A K resolutions**. Meanwhile, the Xbox Series X also features an **AM descend to CPU**, but clocks in at **three point take. It hurts** instead, **the cost. Almost a similar AMG custom** GPU, with 12 teraflops and 1.825 gigahertz memory, is the same as out of the PS. Fives **come in at** 16 gigabytes, but the default storage is where the system has an edge. Bring out a massive one terabyte hard drive. Like the PS five. The **Siri's X** also supports four K and eight K resolution screens as well. Those, of course, are just the numbers. Therefore, it remains to be seen exactly how the

CLM (training only)

for a little excitement around the next Gen video game consoles coming out soon. So what's the difference in **hardware specs** between the upcoming PlayStation five and Xbox Series X? Well, let's take a look under the hood of each of these **consoles**. The PS five features **an A M D Zen two CPU** with up to 3.5 gigahertz frequency it sports **and am the radio** GPU. That **tells** 10.3 teraflops running up to 2.23 gigahertz memory and storage, respectively, **Dahlin** at 16 gigabytes and 825 gigabytes, The PS five supports both PS. The PS five supports both four K and **eight K resolutions**. Meanwhile, the Xbox Series X also features an **A M D Zen two CPU**, but clocks in at 3.8 hertz. Instead. **The console boasts a similar a AMD custom** GPU, with 12 teraflops and 1.825 hertz. Memory is the same as out of the PS five's **come in at** 16 gigabytes, but the default storage is where the system has an edge. Bring out a massive one terabyte hard drive. Like the PS five, the **Series X** also supports four K and eight K resolution screens as well. Those, of course, are just the numbers. Therefore, it remains to be seen exactly how the

CLM (training and tuning)

all hope for a little excitement around the next Gen video game consoles coming out soon. So what's the difference in **hardware specs** between the upcoming PlayStation five and Xbox Series X? Well, let's take a look under the hoods of each of these **consoles**. The PS five features an **AMG Zen two CPU** with up to 3.5 givers frequency it sports an **AMD Radeon** GPU that **touts** 10.3 teraflops running up to 2.23 gigahertz memory and storage, respectively. Dial in at 16 gigabytes and 825 gigabytes. The PS five supports both PS. The PS five supports both four K and eight K resolutions. Meanwhile, the Xbox Series X also features an **AMG Zen two** CPU, but clocks in at 3.8 had hurts. Instead, **the console boasts a similar AMD custom**. GPU, with 12 teraflops and 1.825 gigahertz memory is the same as out of the PS five's **coming in at** 16 gigabytes. But the default storage is where the system has an edge, bringing out a massive one terabyte hard drive. Like the PS five, the **Series X** also supports four K and eight K resolution screens as well. Those, of course, are just the numbers. Therefore, it remains to be seen exactly how the performance

How does CLM work?

PART 4 – EVALUATE CLM PERFORMANCE

Word error rate (WER) – Video Game Example

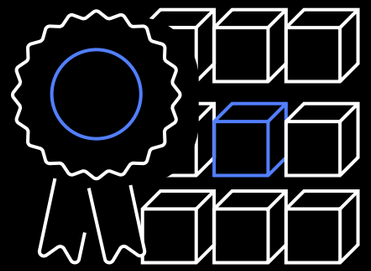
Lower is better

Transcription Type	WER (%)	Accuracy (100-WER)
Amazon Transcribe generic model	31.34%	68.66%
Amazon Transcribe CLM with no tuning split	26.19%	73.81%
Amazon Transcribe CLM with tuning split	20.23%	79.77%

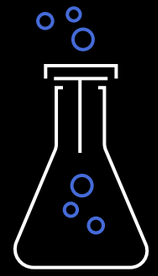
Note: WERs are relative and demonstrate the point of using CLM over generic ASR models. These figures are only specific to this audio experiment and not globally representative of Amazon Transcribe as a service overall. In this example, we used <5MB of training data. In reality, users would use much larger data sets, which would drive further accuracy improvement.

How does CLM work?

CLM BEST PRACTICES



Data quality & quantity



Run experiments

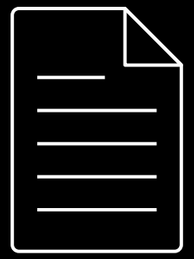


Proper benchmarking

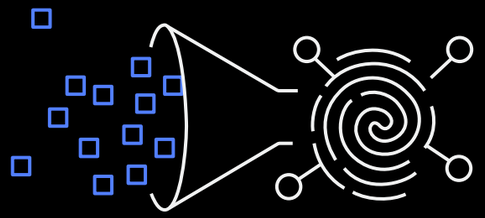
Tech blog:
How to evaluate ASR performance
https://amzn.to/36pfjdO

How does CLM work?

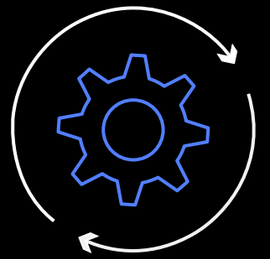
CLM WRAP-UP



Leverage your existing text data



Minimal data wrangling



No machine learning prerequisites

Octopus Energy case study



AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

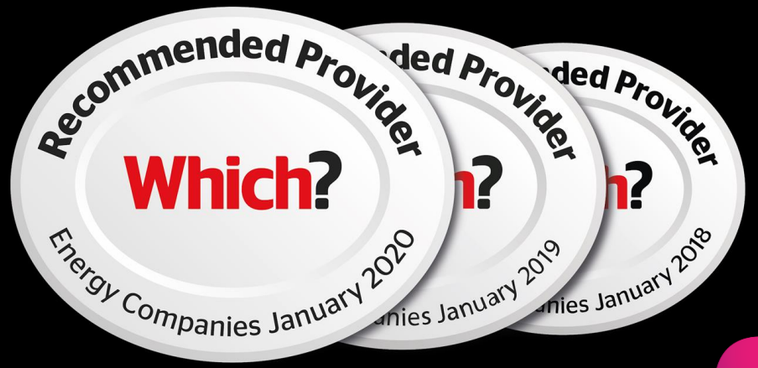
Octopus Energy overview

- Energy supplier and tech company
- Multi-award-winning green energy technology
- 1.7m UK retail customers (5% market share)
- Operating in 5 countries
- £1.8bn annual revenue
- 6m customers via technology licensing

TRUSTPILOT

THE ENERGY AWARDS 19
REWARDING EXCELLENCE IN THE ENERGY BUSINESS
WINNER

WINNER
UtilityWeek AWARDS 2019
Utility of the Year



Octopus Energy's customer care goals

Goals

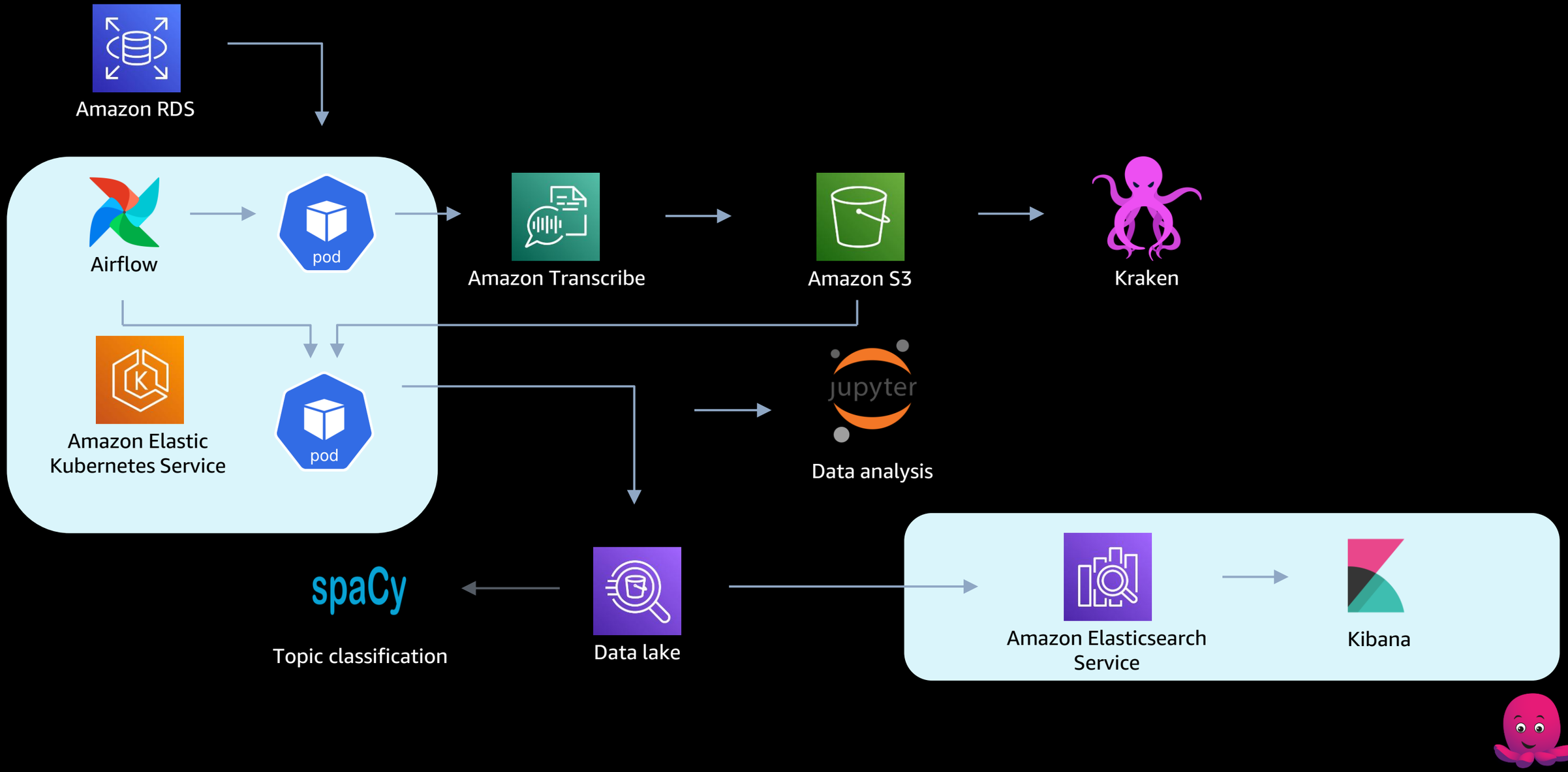
- Improve our customer care with call transcriptions
- Process a million minutes of calls per month
- Open the possibility for call analytics, automatic call topic classification, and sentiment analysis

Solutions

- Selected Amazon Transcribe to convert customer calls into text
- Used custom language models to drive better recognition of business-specific phrases



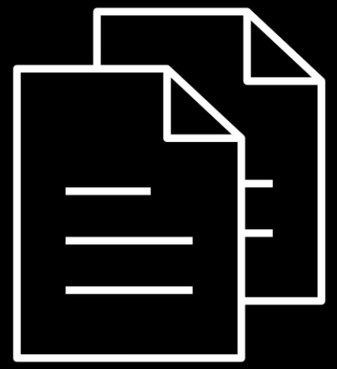
Architecture of customer solution



Model training process



Customer emails



Internal documents



Customer reviews



Case for CLM: Domain-specific language

Smart meter:

```
spk_1: " At least I'm being No , no , no . We got smart me to start today .
```

```
spk_1,"because it's a smart meter , it's a smart meter here , So I've got I've got smart meter ."
```

Tariff names:

```
spk_0: "[...] flexible Taras . But while the flexible tower has changed for other customers [...]"
```

```
spk_0,"[...]. So according to the tariff , the cheapest tariff would be the 12 month fixed tariff on . That's £60.86 per month ."
```

Hello, Octopus Energy:

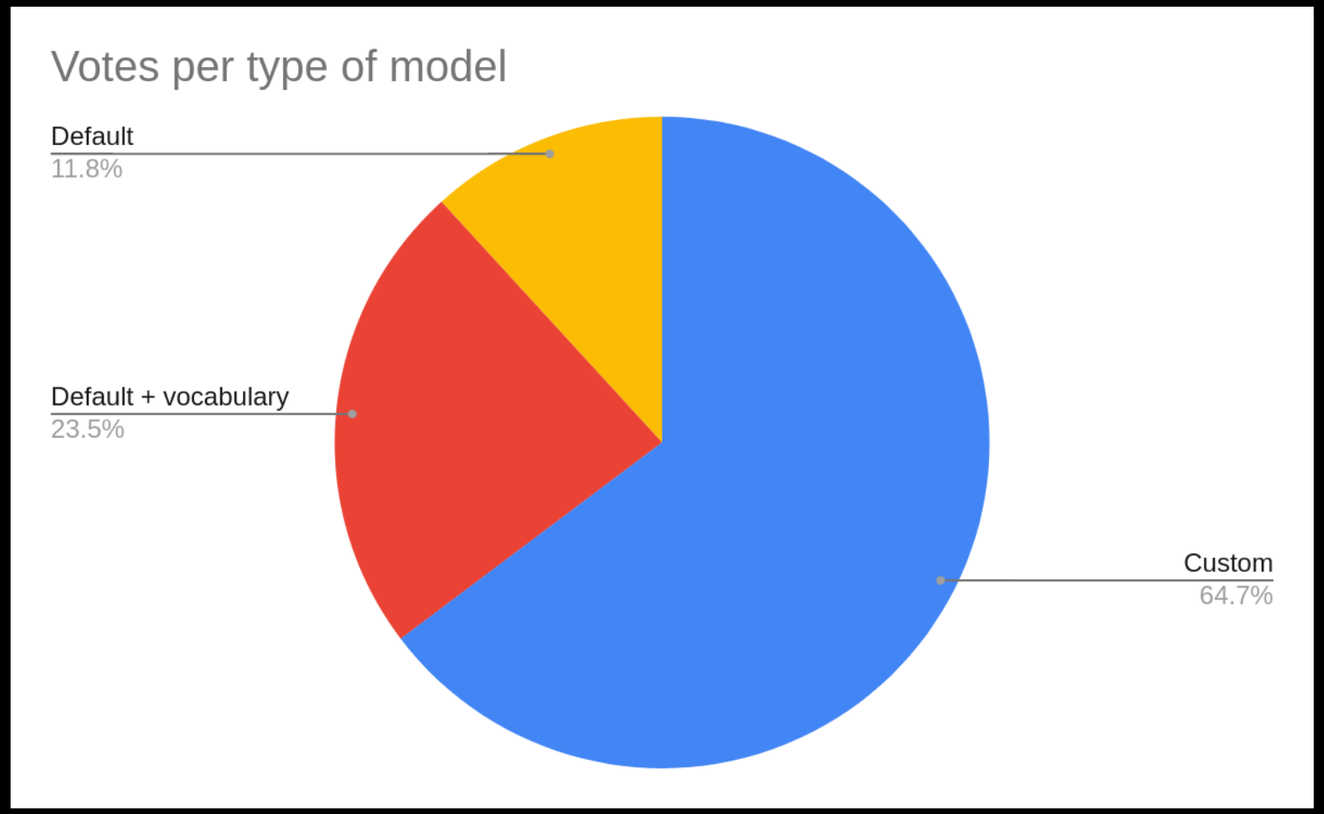
```
spk_0,"earlier . Three adults . Energy . How could help ?"
```

```
spk_0,"Good afternoon , Emily Octopus Energy . How can I help ?"
```



Model evaluation

- Qualitative evaluation (voting)

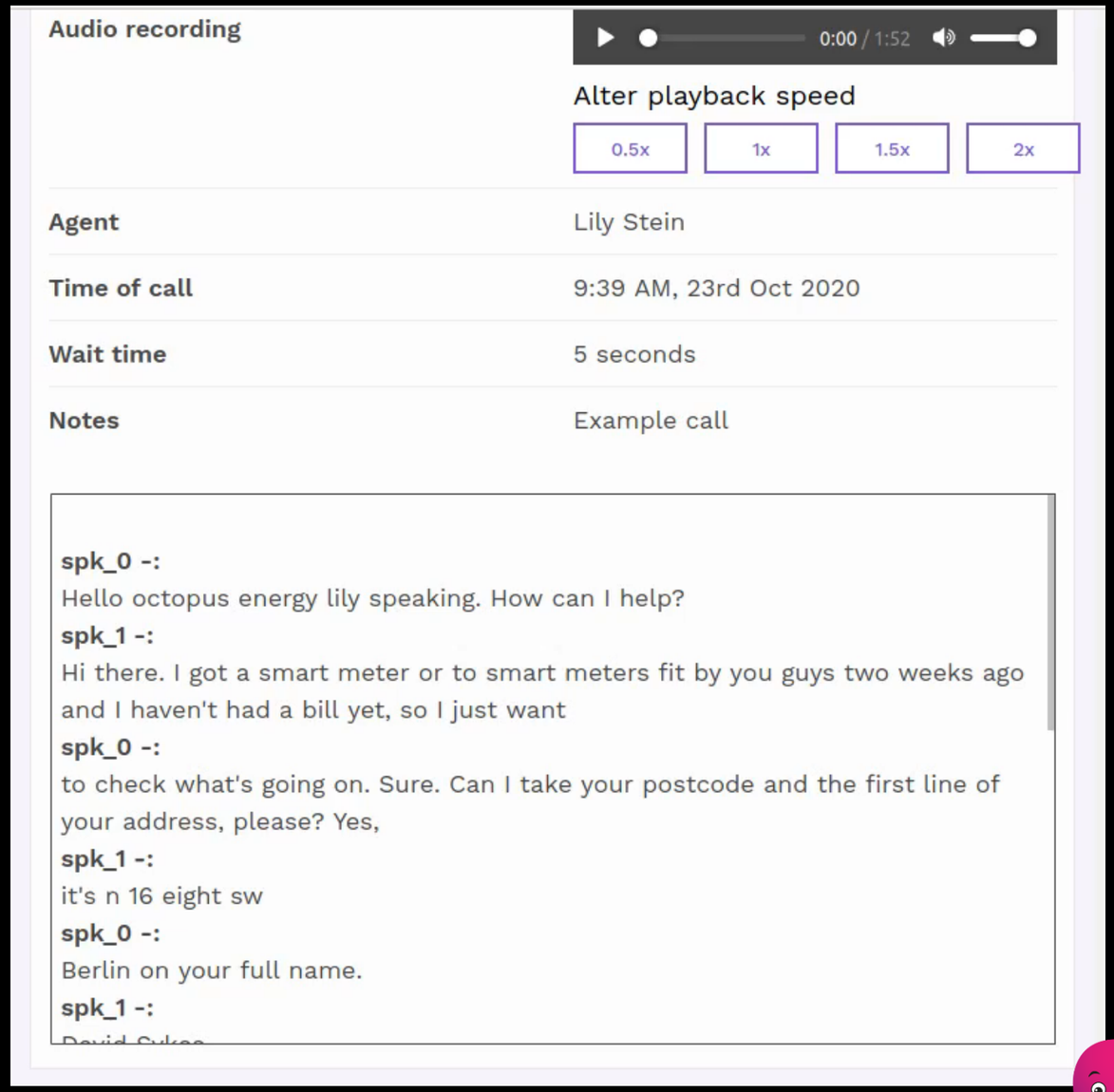


- Quantitative evaluation (WER)

Depending on the file, between 12% and 21% relative accuracy improvement

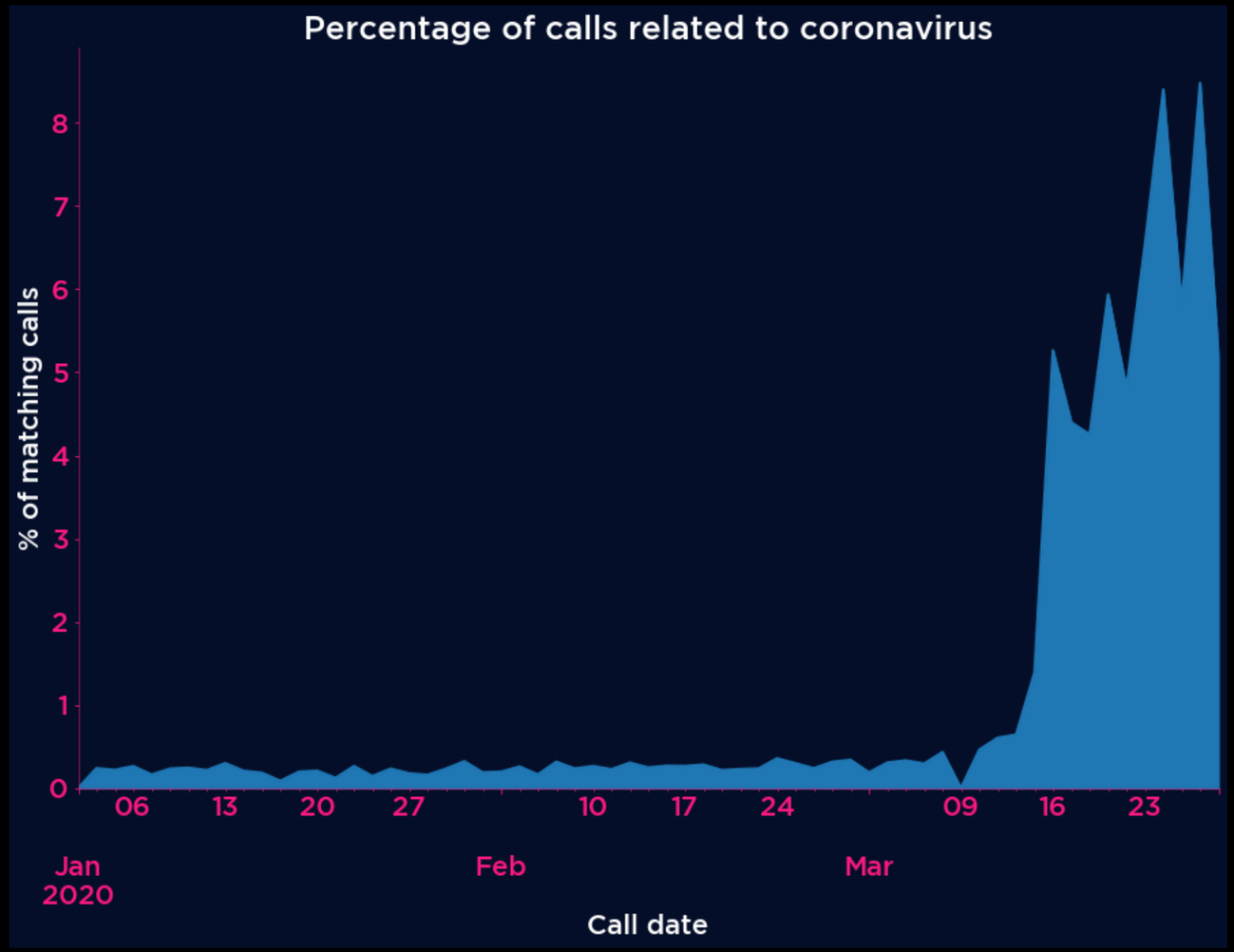
Business benefit and impact

- Processed 1.8 years’ worth of customer calls every month
- Immediate access to past customer interactions and issue tracking, which leads to better customer service



Business benefit and impact

- Uncover new business insights
- Use text analytics to extract trends (i.e., COVID-19) for fast government reporting



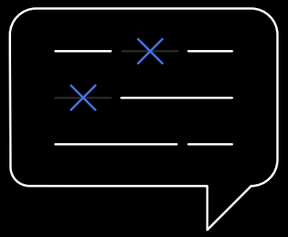
Octopus Energy roadmap



Real-time transcription and next best actions



Sentiment and customer satisfaction metrics



Compliance monitoring and dashboards

“Octopus Energy is a UK-based retail electricity and gas supplier specialising in sustainable energy. In our mission to provide customers with fair and transparent pricing along with excellent service, we analyse millions of minutes of contact center calls each month. We’ve leveraged our own domain-specific data to train custom speech recognition models using Amazon Transcribe. These custom models provide enhanced transcription accuracy and enable us to more efficiently and intelligently identify, understand, and serve our customers’ energy needs.”

Dr. Javier Asensio-Cubero

Lead of Data Platform

Octopus Energy

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Next steps



Build custom language models

https://amzn.to/32BDkNJ



Engage with a solutions architect

Brainstorm use cases and get support for solution implementation



Connect with an AWS Partner

Let AWS Partners help you with the POC and build integrations



Get started with Amazon Transcribe

Sign up for an AWS account and start transcribing your content today

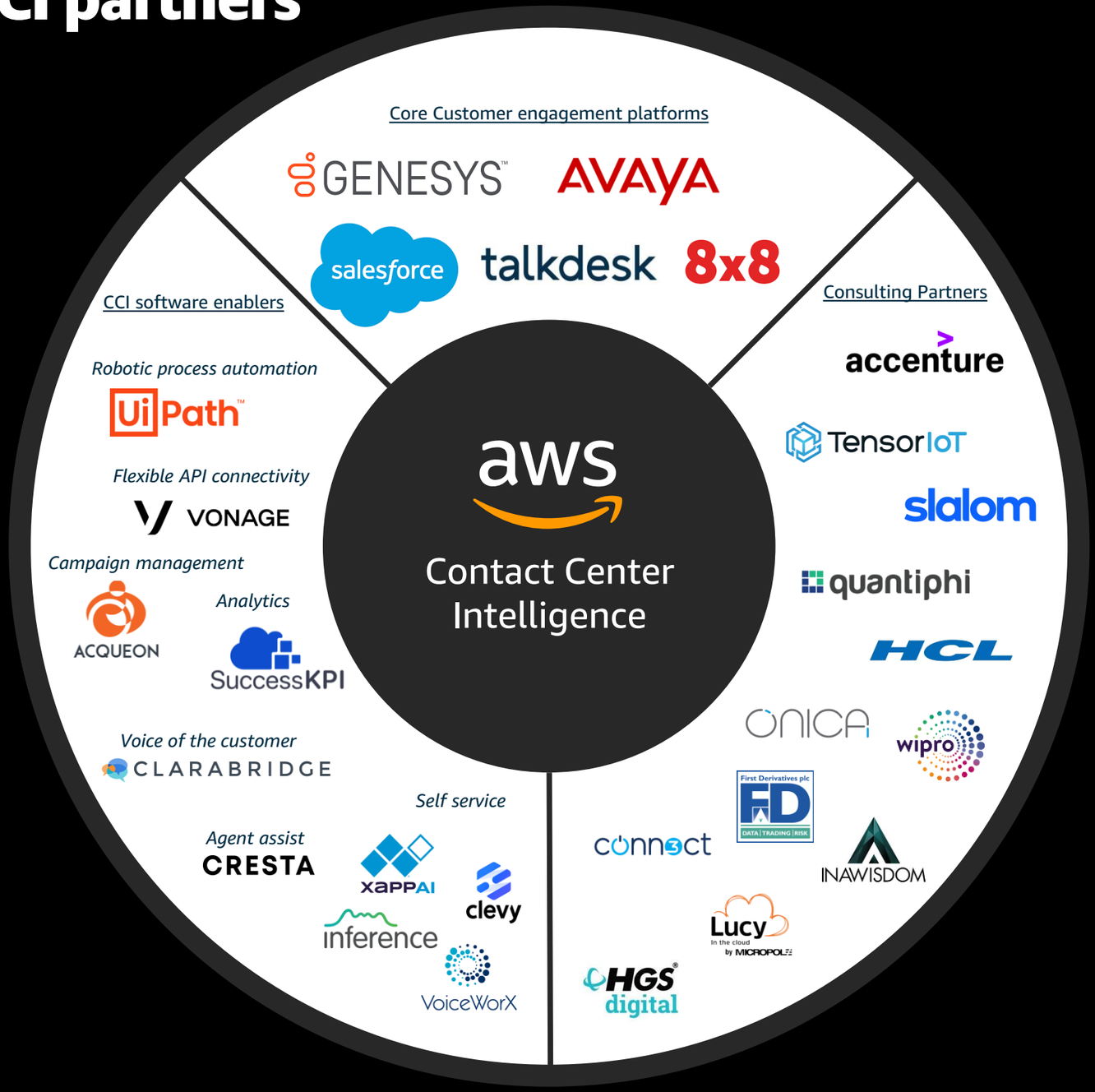
Announcing a global network of CCI partners

Technology partners

GENESYS · VONAGE · UiPath · salesforce · SuccessKPI · inference · ACQUEON · VoiceWorx · talkdesk · 8x8 · AVAYA · XAPPAI · CRESTA · CLARABRIDGE · clevy

Consulting partners

ONICA · quantiphi · slalom · HCL · TensorIoT · HGS digital · accenture · wipro · INAWISDOM · connect · MICROPOLE · First Derivatives



Thank you!



AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Please complete the session survey

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.