TensorIoT

Generative AI: AI Intelligence at Cloud Scale

Unlock the Future with Generative AI

Ben Rodrigue - Generative AI Product Leader - BenRod@TensorIoT.com
Yuxin Yang - Machine Learning Practice Manager - Yuxin.Yang@TensorIoT.com

©2021, TensorIoT • ALL RIGHTS RESERVED

©2021. TensorIoT • ALL RIGHTS RESERVED

aws

PARTNER
Advanced Tier
Services

- ML Services Competency
- IoT Services Competency
- Retail Services Competency
- Industrial Software Service Competency
- Travel & Hospitality Services

2022
AWS Sustainability Partner of the Year
North America

Making Things Intelligent.

World Class IoT, Data, AI & ML on AWS

- Cloud Native Products and Consulting
- Headquartered in Southern California, Global Presence
- Focus on IoT, AI, ML, Data, App Development & Modernization
- Meet the customer wherever they are in their cloud adoption journey
- Named Sustainability Partner of the year at re:Invent 2022

TensorIoT

What is Generative AI?

©2021, TensorIoT • ALL RIGHTS RESERVED

TensorIoT

What if you could feed a Generative AI Model with your company's data to automate business functions, generate content, or drive critical decisions all while retaining control of your data?

©2021, TensorIoT • ALL RIGHTS RESERVED

TensorIoT

Generative AI: Adoption Journey with TensorIoT

Unlock the Future with Generative AI

©2021, TensorIoT • ALL RIGHTS RESERVED

Generative AI adoption journey

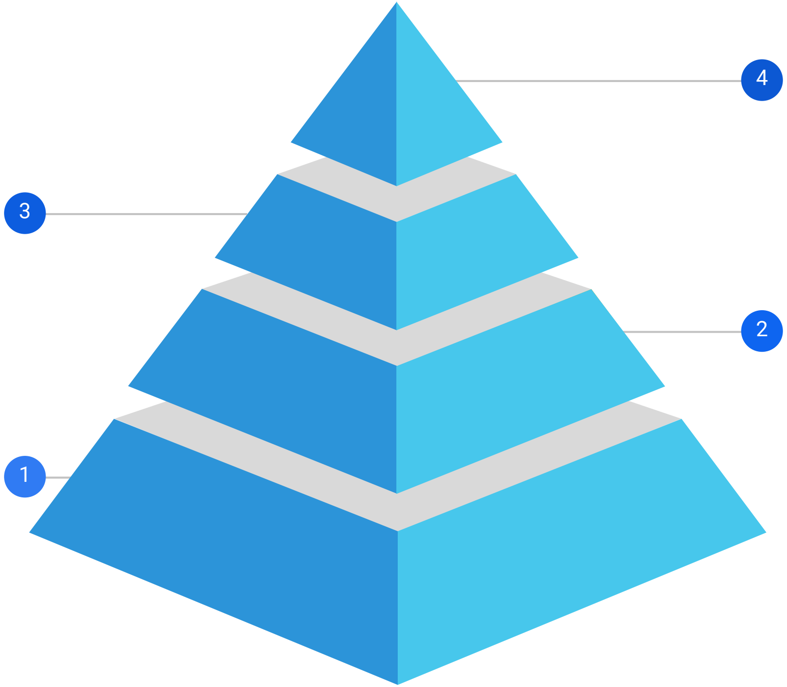
We meet you where you are.

GOVERNANCE

- Review risk, legal, security and design functions
- Obtain design approvals
- Understand data compliance
- Review legislation

VALUE & STRATEGY

- Understanding your need
- Generate stakeholder buy-in
- Curate your data strategy
- Review use cases and impact
- Create a data factory / DevOps PIpeline



VALUE REALIZATION

- Go Live into Production
- Review reusable features
- Deliver Continuous Engineering
- Drive addition use cases

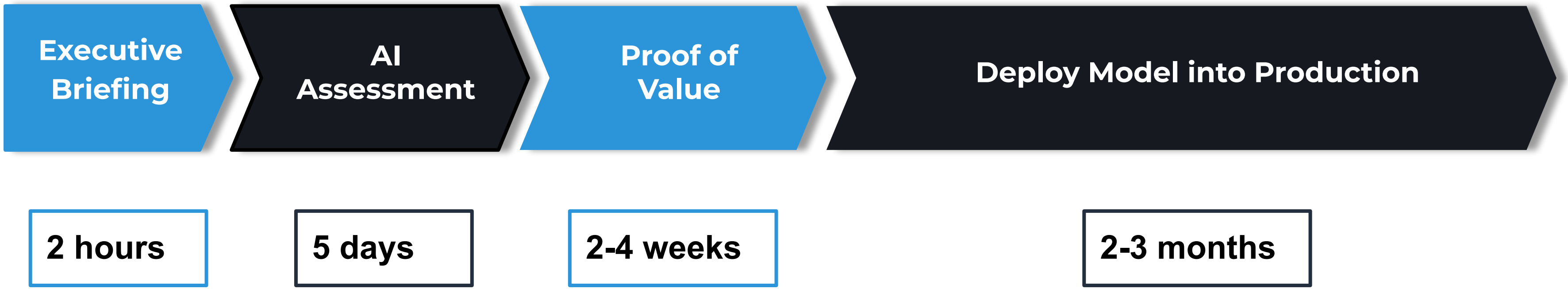
PEOPLE & CAPABILITIES

- Review operating model
- Train and enable your business
- Prepare customers
- Finetune user stories

©2021, TensorIoT • ALL RIGHTS RESERVED

Results at Velocity

We provide tangible results in weeks, not years



Executive Briefing

AI Assessment

Proof of Value

Deploy Model into Production

2 hours

5 days

2-4 weeks

2-3 months

©2021, TensorIoT • ALL RIGHTS RESERVED

Results at Velocity

We provide tangible results in weeks, not years

Executive Briefing: (2 hrs)
Get insights in TensorIoT's AI best practices, capabilities, and high value use cases

Generative AI Assessment: (5 days)
Assess your organization's readiness to embark on a Generative AI journey and build a roadmap to deploy internal AI services.

Proof of Value: (2-4 weeks)
Identify high impact use case and rapidly test a foundational model against your data set.

Deploy Models into Production:(2-3 months)
Train, Refine, Deploy, and Scale your model into production and begin reaping the benefits and economic value of Generative AI.

©2021, TensorIoT • ALL RIGHTS RESERVED

TensorIoT

Generative AI: Strategy Assessment

Unlock the Future with Generative AI

©2021, TensorIoT • ALL RIGHTS RESERVED

Generative AI: Readiness Assessment

Objectives and Benefits

Objective:

Assess your organization's readiness to embark on a Generative AI journey and build a roadmap to deploy internal AI services.

Benefits:

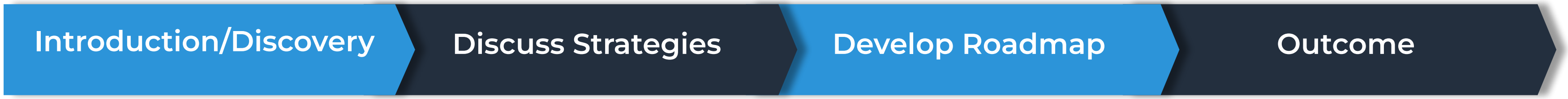
- Prepare for the business disruption from Foundation Models, which are seen in action with technologies like ChatGPT, BARD, Amazon Bedrock & Hugging Face.
- Determine Generative AI's potential value in your business context.
- Assess your organization's readiness for AI implementation.
- Identify AI use cases aligned with your strategic goals.
- Ensure secure, unbiased, and compliant AI deployment.
- Develop a customized roadmap for seamless AI adoption.

©2021, TensorIoT • ALL RIGHTS RESERVED

©2021, TensorIoT • ALL RIGHTS RESERVED

Assessment Overview:

- **Duration:** 5 days with 2-4 hour sessions
- **Participants:** Cross-functional teams from IT, Business, Data Science, and Security
- **Output:** A tailored AI readiness assessment scorecard, use case catalogue, and roadmap to guide your Generative AI journey



Business Overview:

Introduction to Generative AI, assessing your organization's resources and capabilities, identifying high-priority goals, and exploring potential Generative AI use cases.

Strategy and Capabilities:

Reviewing your current cross-organizational strategy to leverage Generative AI, reviewing current data footprint maturity on AWS, and exploring the long-term strategy for Generative AI in your organization.

Recommend Next Steps:

Developing a tailored roadmap for Generative AI maturity and adoption, identifying potential challenges and mitigation strategies, and discussing the next steps to achieve your Generative AI objectives.

Outcome:

Determine an action plan for roadmap implementation, encompassing design choices, AWS architecture, timeline projections, and the identification of high-impact use cases to achieve both short-term and long-term objectives.

TensorIoT

Generative AI: Proof of Value

Unlock the Future with Generative AI

©2021, TensorIoT • ALL RIGHTS RESERVED

Generative AI: Proof of Value

Objectives and Benefits

Overview:

Wizdom AI is an all-inclusive technical framework, enabling you to train and deploy Generative AI, AWS Services, and Foundational Models like Bloom, Titan, and Jurassic within your own AWS account.

Objective:

Discover a high-impact use case and swiftly evaluate a foundational model using your data set, showcasing immediate value to all stakeholders.

Benefits:

- Functional PoV in 2-4 weeks
- Stays in your account giving you full control to the model and your data
- Deployed using out of the box, mature, Foundational Models
- Build on AWS leveraging next-gen services

Wizdom AI

©2021, TensorIoT • ALL RIGHTS RESERVED

Generative AI: Proof of Value

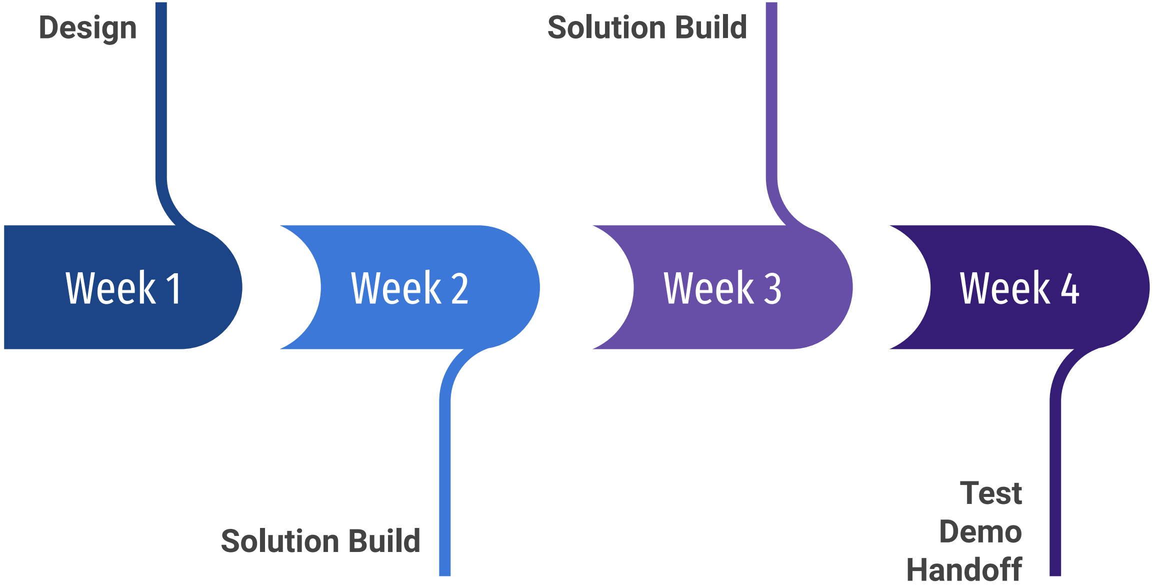
Solution Highlights

	Value	3rd party (OpenAI)	Wizdom AI Framework	Additional Details
1	You control the Model	✗	✓	• All Models are deployed in your AWS account • Foundational Models are Open Source or licensed by you
2	You control your Data	✗	✓	• Full Privacy • Your data never leaves your account
3	End to End Security	✗	✓	• Maintain industry compliance • Maintain data sovereignty • Follows security best practices
4	Fully integrated with AWS services	✗	✓	• Seamlessly integrates with AWS services • On demand, scalable managed services • Cost optimized
5	Dedicated Support and Maintenance	✗	✓	• Support and maintenance available • Ongoing model performance available • Ongoing updates available

©2021, TensorIoT • ALL RIGHTS RESERVED

Generative AI: Proof of Value

Execution Journey



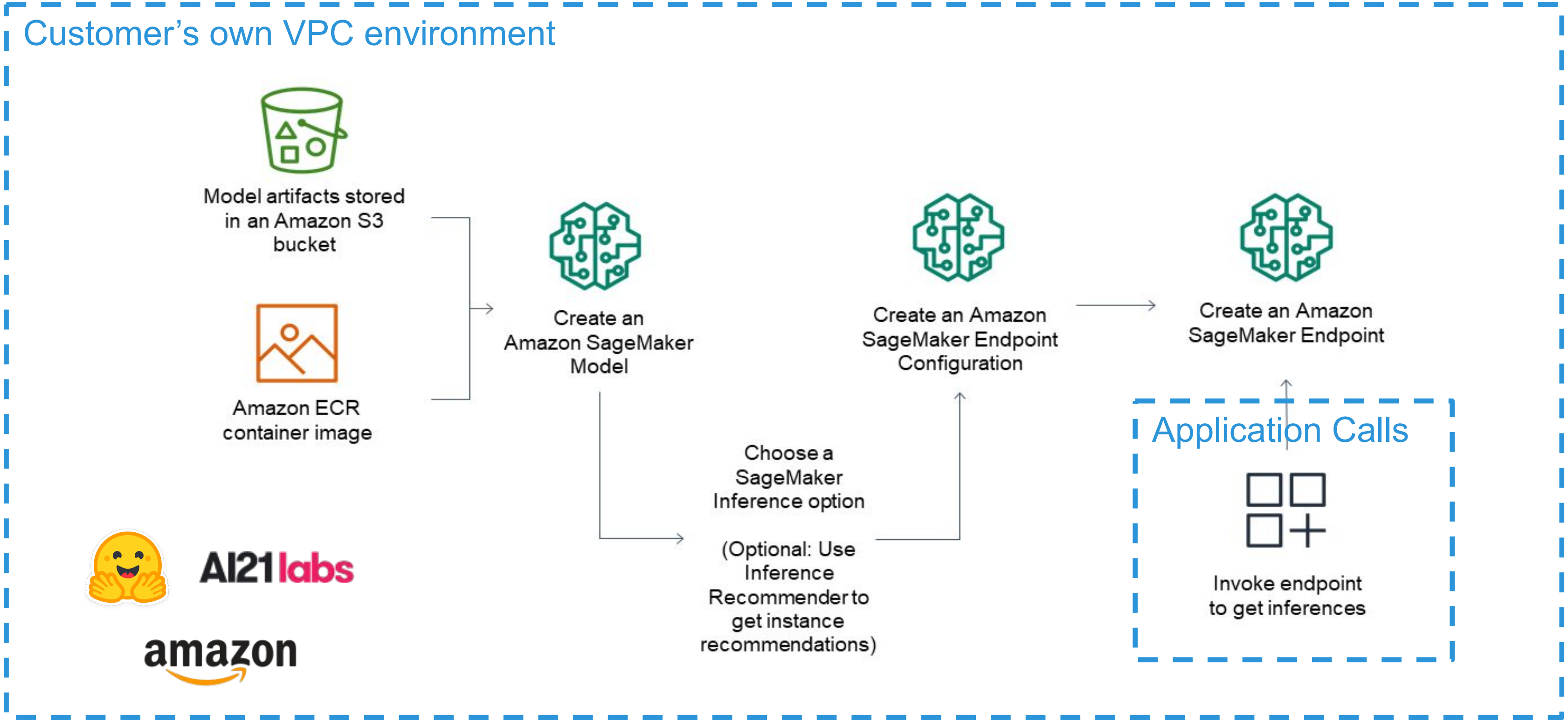
©2021, TensorIoT • ALL RIGHTS RESERVED

TensorIoT

Example Architectures

©2021, TensorIoT • ALL RIGHTS RESERVED

Wizdom AI: Inference Backend



Customer's own VPC environment

Model artifacts stored in an Amazon S3 bucket

Amazon ECR container image

Create an Amazon SageMaker Model

Choose a SageMaker Inference option

(Optional: Use Inference Recommender to get instance recommendations)

Create an Amazon SageMaker Endpoint Configuration

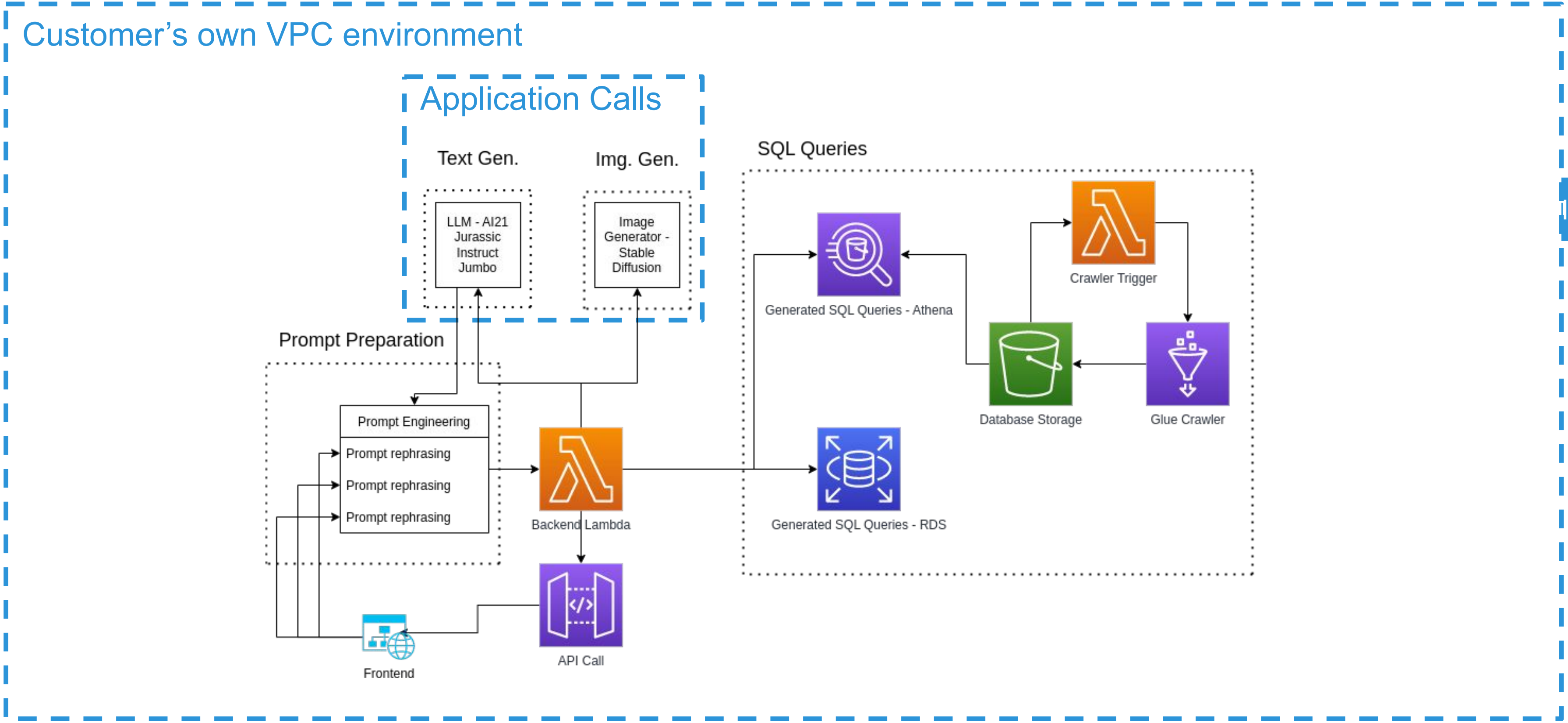
Create an Amazon SageMaker Endpoint

Application Calls

Invoke endpoint to get inferences

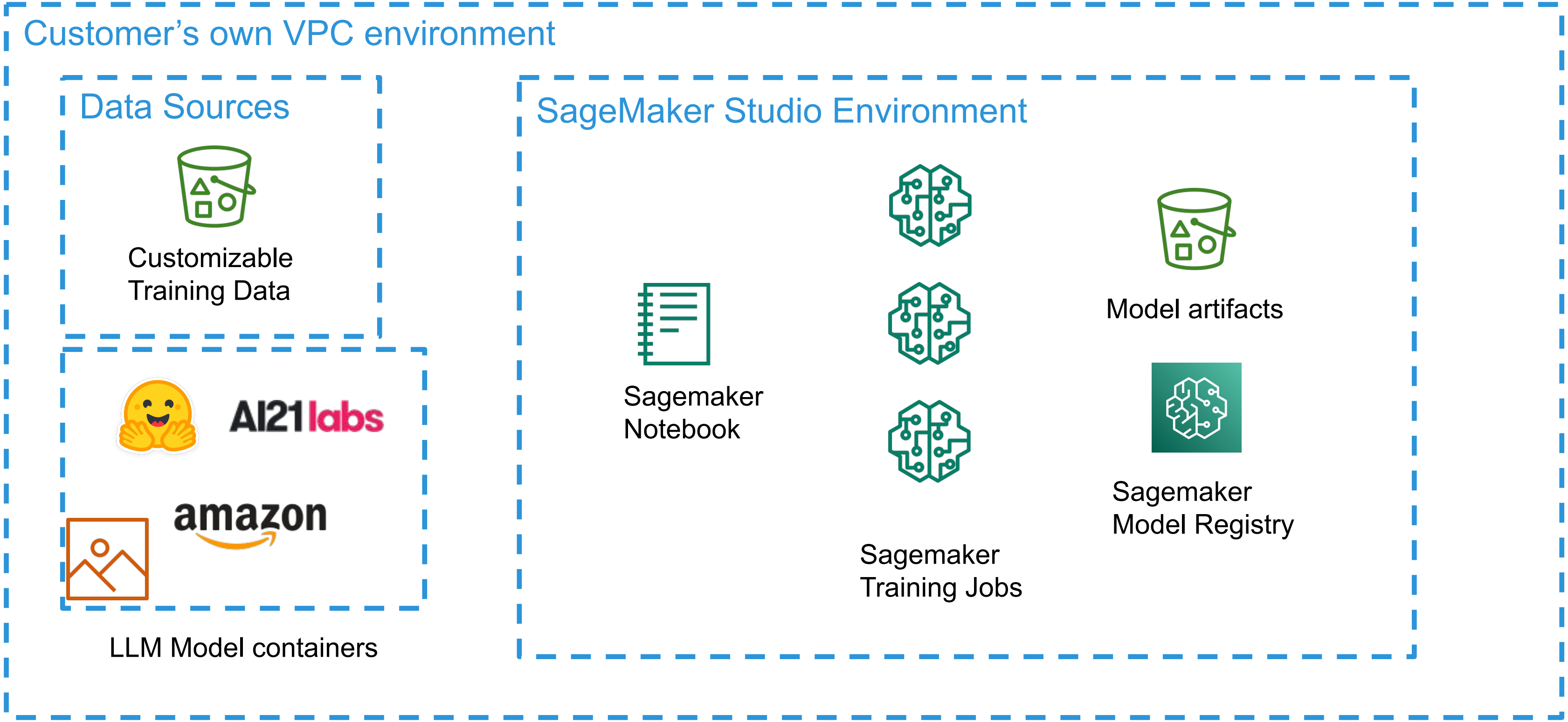
©2021, TensorIoT • ALL RIGHTS RESERVED

Wizdom AI: Core Architecture



©2021, TensorIoT • ALL RIGHTS RESERVED

Wizdom AI: Fine-tuning LLM Architecture



Customer's own VPC environment

Data Sources

Customizable Training Data

LLM Model containers

SageMaker Studio Environment

Sagemaker Notebook

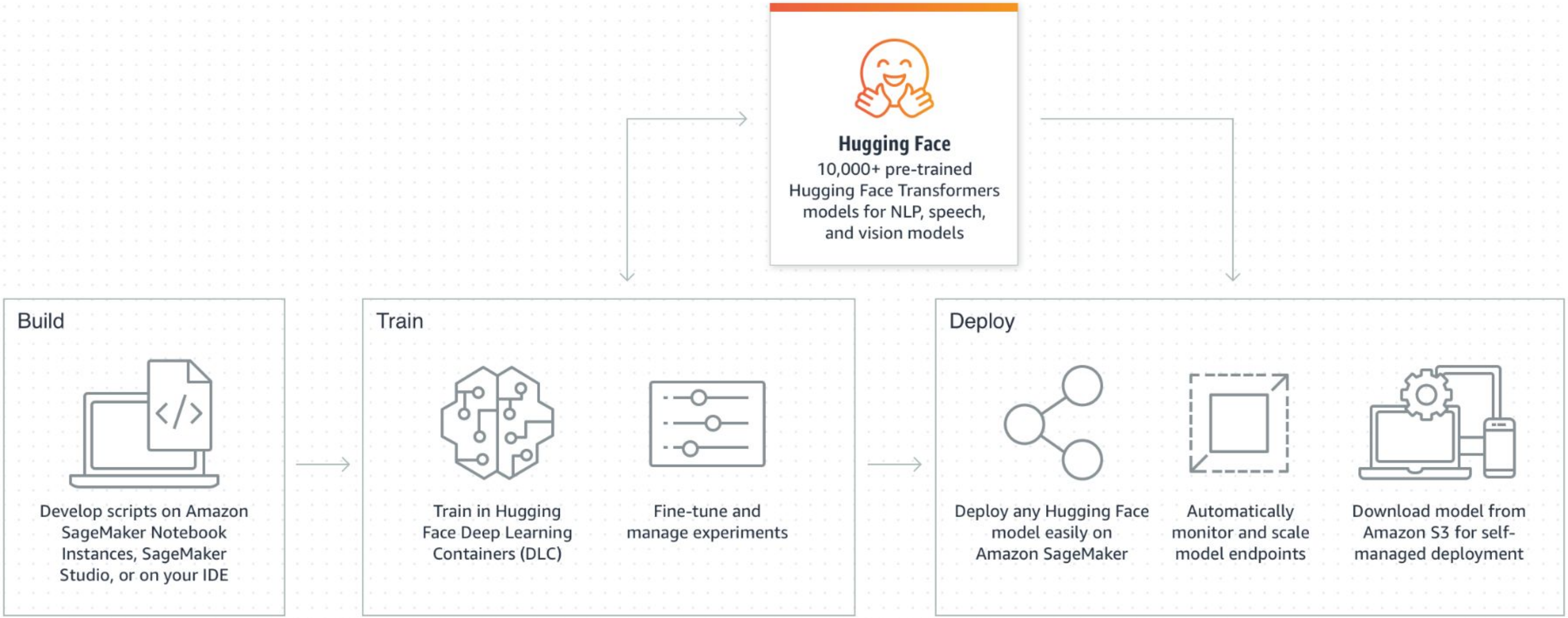
Sagemaker Training Jobs

Model artifacts

Sagemaker Model Registry

©2021, TensorIoT • ALL RIGHTS RESERVED

Transformers Deployed On AWS



Hugging Face

10,000+ pre-trained Hugging Face Transformers models for NLP, speech, and vision models

Build

Develop scripts on Amazon SageMaker Notebook Instances, SageMaker Studio, or on your IDE

Train

Train in Hugging Face Deep Learning Containers (DLC)

Fine-tune and manage experiments

Deploy

Deploy any Hugging Face model easily on Amazon SageMaker

Automatically monitor and scale model endpoints

Download model from Amazon S3 for self-managed deployment

RIGHTS RESERVED

TensorIoT

Thank you

Ben Rodrigue - Generative AI Product Leader - BenRod@TensorIoT.com
Yuxin Yang - Machine Learning Practice Manager - Yuxin.Yang@TensorIoT.com

©2021, TensorIoT • ALL RIGHTS RESERVED

Wizdom AI: Model Training

ML Frameworks

PyTorch

TensorFlow

AWS SageMaker Training Platform

Large-scale Clusters

Fault-tolerant

Hugging Face Integration

SageMaker Training Compiler

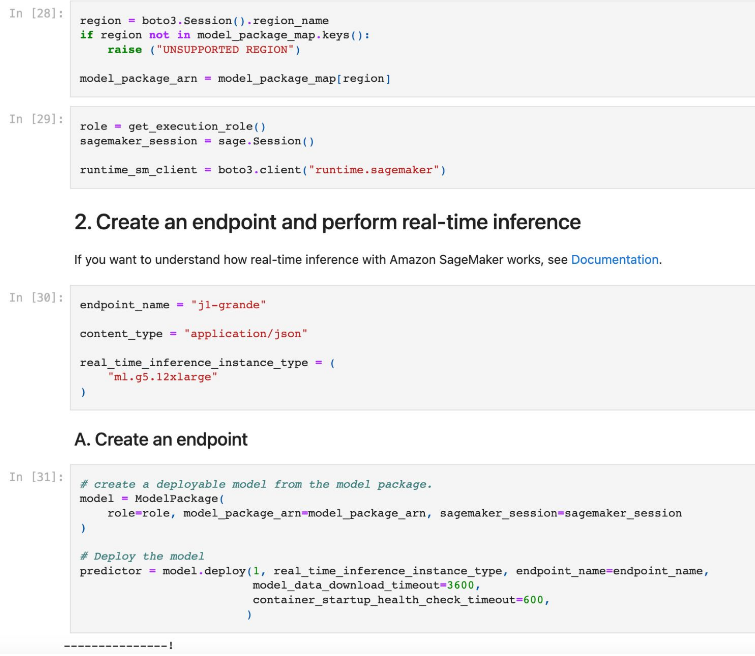
AWS SageMaker Training Parallelism

Model Parallelism

Data Parallelism

©2021, TensorIoT • ALL RIGHTS RESERVED

AWS FM deployment in Production (Deployment + Inference)



In [28]:

```
region = boto3.Session().region_name
if region not in model_package_map.keys():
    raise ("UNSUPPORTED REGION")

model_package_arn = model_package_map[region]
```

In [29]:

```
role = get_execution_role()
sagemaker_session = sage.Session()

runtime_sm_client = boto3.client("runtime.sagemaker")
```

2. Create an endpoint and perform real-time inference

If you want to understand how real-time inference with Amazon SageMaker works, see Documentation.

In [30]:

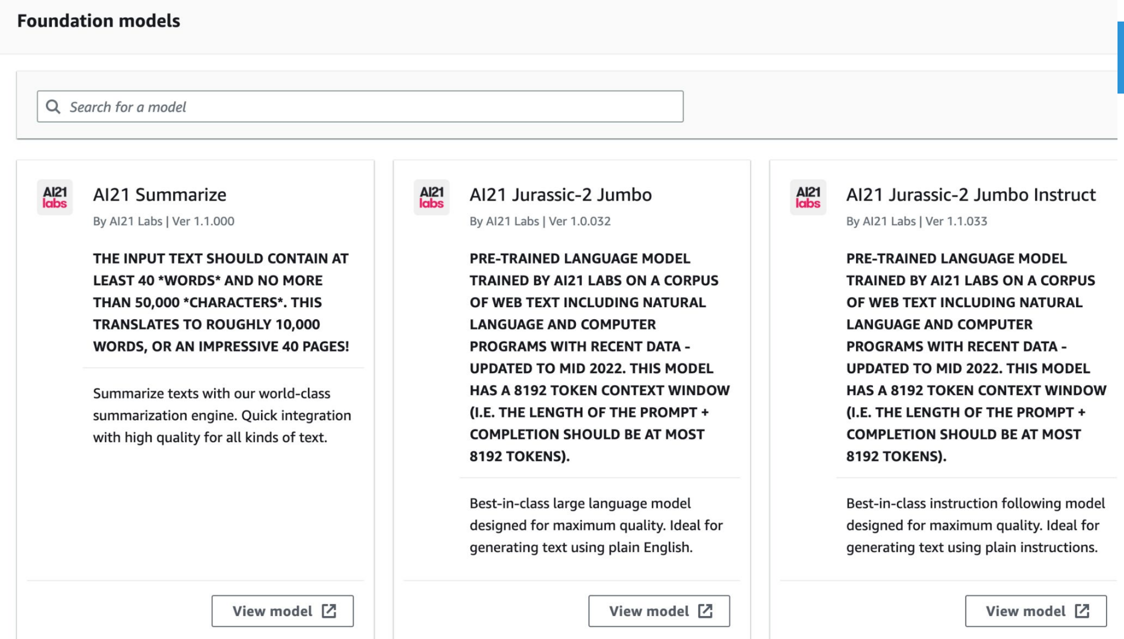
```
endpoint_name = "j1-grande"

content_type = "application/json"

real_time_inference_instance_type = (
    "ml.g5.12xlarge"
)
```

A. Create an endpoint

In [31]:

```
# create a deployable model from the model package.
model = ModelPackage(
    role=role, model_package_arn=model_package_arn, sagemaker_session=sagemaker_session
)

# Deploy the model
predictor = model.deploy(1, real_time_inference_instance_type, endpoint_name=endpoint_name,
                         model_data_download_timeout=3600,
                         container_startup_health_check_timeout=600,
                        )
```

```
---------------!
```



Foundation models

AI21 Summarize

By AI21 Labs | Ver 1.1.000

THE INPUT TEXT SHOULD CONTAIN AT LEAST 40 *WORDS* AND NO MORE THAN 50,000 *CHARACTERS*. THIS TRANSLATES TO ROUGHLY 10,000 WORDS, OR AN IMPRESSIVE 40 PAGES!

Summarize texts with our world-class summarization engine. Quick integration with high quality for all kinds of text.

AI21 Jurassic-2 Jumbo

By AI21 Labs | Ver 1.0.032

PRE-TRAINED LANGUAGE MODEL TRAINED BY AI21 LABS ON A CORPUS OF WEB TEXT INCLUDING NATURAL LANGUAGE AND COMPUTER PROGRAMS WITH RECENT DATA - UPDATED TO MID 2022. THIS MODEL HAS A 8192 TOKEN CONTEXT WINDOW (I.E. THE LENGTH OF THE PROMPT + COMPLETION SHOULD BE AT MOST 8192 TOKENS).

Best-in-class large language model designed for maximum quality. Ideal for generating text using plain English.

AI21 Jurassic-2 Jumbo Instruct

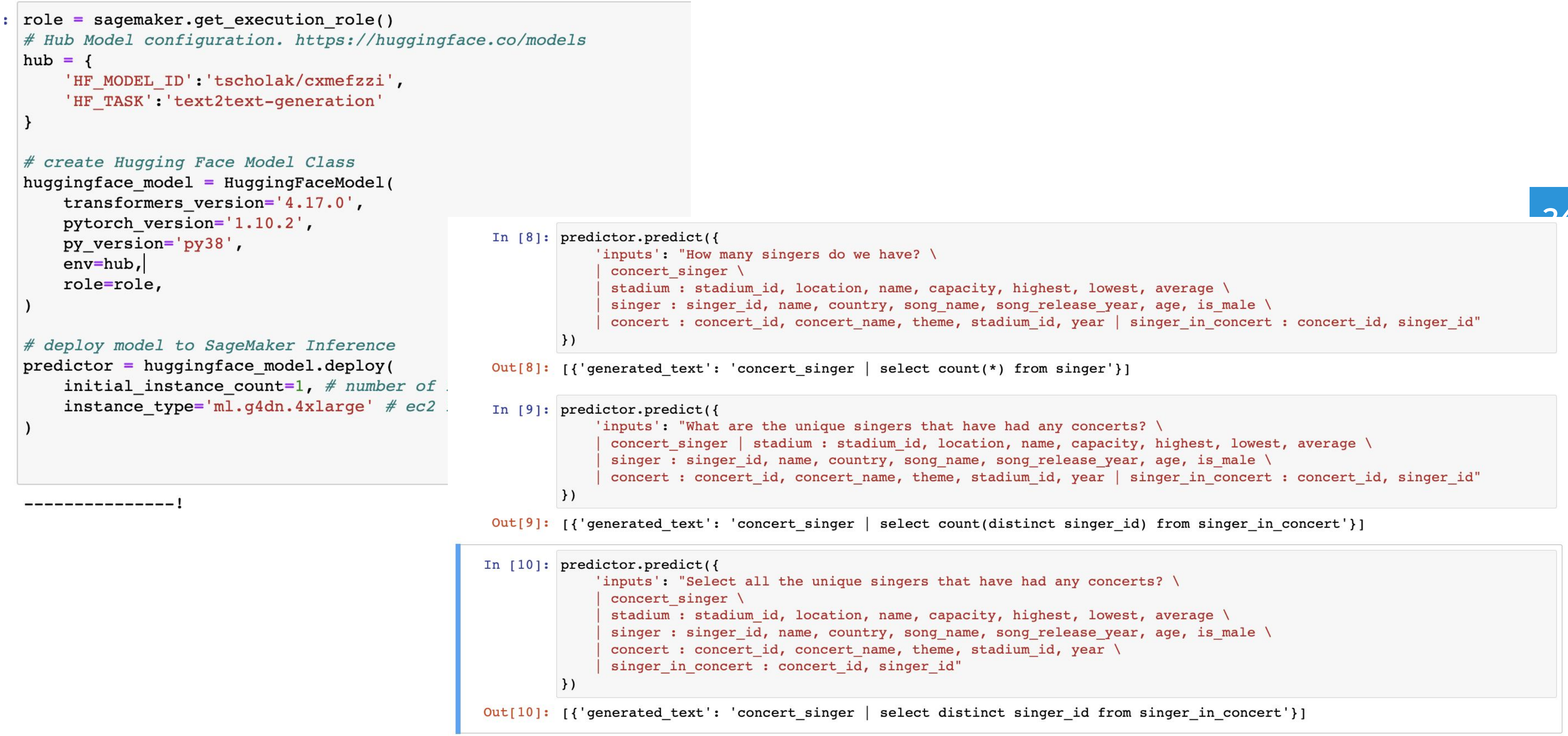
By AI21 Labs | Ver 1.1.033

PRE-TRAINED LANGUAGE MODEL TRAINED BY AI21 LABS ON A CORPUS OF WEB TEXT INCLUDING NATURAL LANGUAGE AND COMPUTER PROGRAMS WITH RECENT DATA - UPDATED TO MID 2022. THIS MODEL HAS A 8192 TOKEN CONTEXT WINDOW (I.E. THE LENGTH OF THE PROMPT + COMPLETION SHOULD BE AT MOST 8192 TOKENS).

Best-in-class instruction following model designed for maximum quality. Ideal for generating text using plain instructions.

©2021, TensorIoT • ALL RIGHTS RESERVED

Hugging Face in Production (Deployment + Inference)

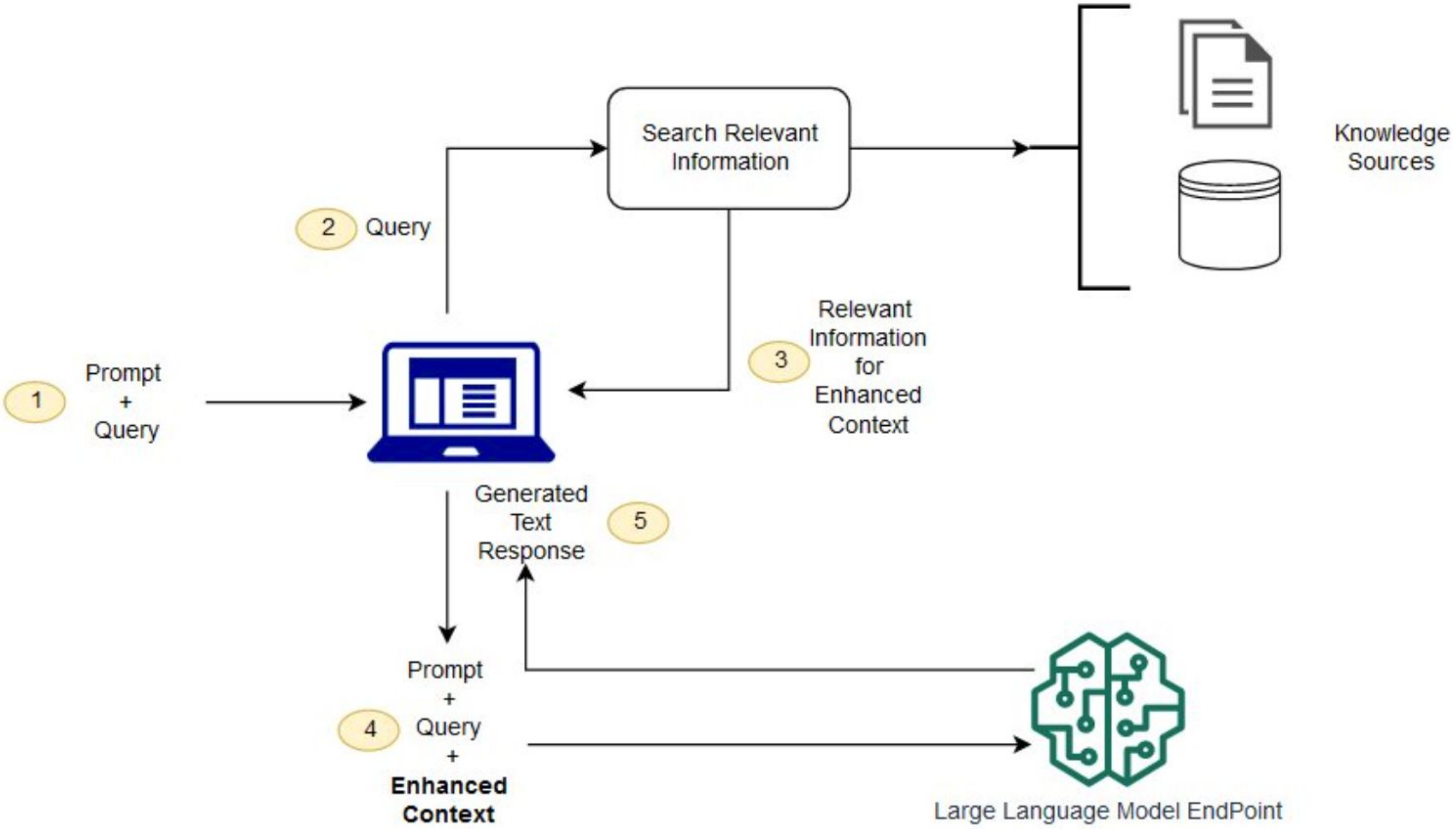


```
role = sagemaker.get_execution_role()
# Hub Model configuration. https://huggingface.co/models
hub = {
    'HF_MODEL_ID':'tscholak/cxmefzzi',
    'HF_TASK':'text2text-generation'
}

# create Hugging Face Model Class
huggingface_model = HuggingFaceModel(
    transformers_version='4.17.0',
    pytorch_version='1.10.2',
    py_version='py38',
    env=hub,
    role=role,
)

# deploy model to SageMaker Inference
predictor = huggingface_model.deploy(
    initial_instance_count=1, # number of
    instance_type='ml.g4dn.4xlarge' # ec2
)
```

```
---------------!
```

```
In [8]:  predictor.predict({
             'inputs': "How many singers do we have? \
             | concert_singer \
             | stadium : stadium_id, location, name, capacity, highest, lowest, average \
             | singer : singer_id, name, country, song_name, song_release_year, age, is_male \
             | concert : concert_id, concert_name, theme, stadium_id, year | singer_in_concert : concert_id, singer_id"
         })

Out[8]:  [{'generated_text': 'concert_singer | select count(*) from singer'}]

In [9]:  predictor.predict({
             'inputs': "What are the unique singers that have had any concerts? \
             | concert_singer | stadium : stadium_id, location, name, capacity, highest, lowest, average \
             | singer : singer_id, name, country, song_name, song_release_year, age, is_male \
             | concert : concert_id, concert_name, theme, stadium_id, year | singer_in_concert : concert_id, singer_id"
         })

Out[9]:  [{'generated_text': 'concert_singer | select count(distinct singer_id) from singer_in_concert'}]

In [10]: predictor.predict({
             'inputs': "Select all the unique singers that have had any concerts? \
             | concert_singer \
             | stadium : stadium_id, location, name, capacity, highest, lowest, average \
             | singer : singer_id, name, country, song_name, song_release_year, age, is_male \
             | concert : concert_id, concert_name, theme, stadium_id, year \
             | singer_in_concert : concert_id, singer_id"
         })

Out[10]: [{'generated_text': 'concert_singer | select distinct singer_id from singer_in_concert'}]
```

©2021, TensorIoT • ALL RIGHTS RESERVED


Search Relevant Information
Knowledge Sources
2 Query
1 Prompt + Query
3 Relevant Information for Enhanced Context
Generated Text Response 5
4 Prompt + Query + Enhanced Context
Large Language Model EndPoint

Trends in Artificial Intelligence

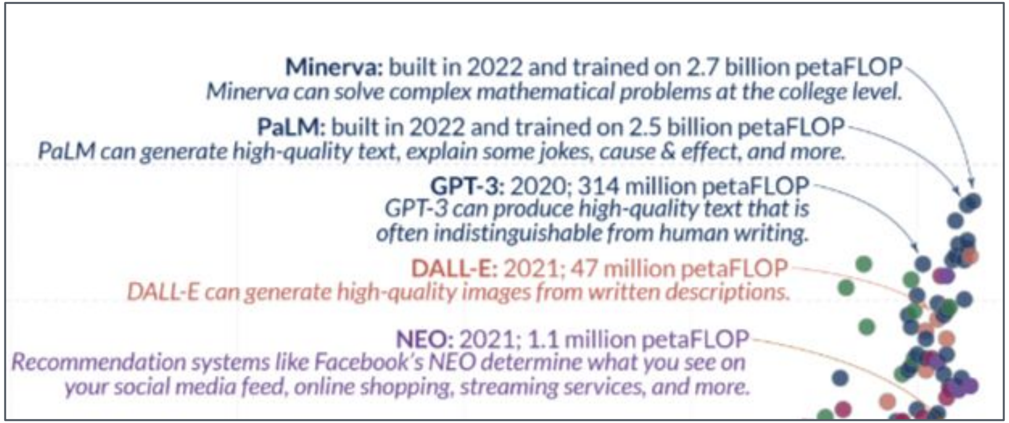
AI Index Annual Report 2023

1. Industry races ahead of academia
2. Performance saturation on traditional benchmarks.
3. AI is both helping and harming the environment.
4. The world's best new scientist … AI?
5. The number of incidents concerning the misuse of AI is rapidly rising.
6. The demand for AI-related professional skills is increasing across virtually every American industrial sector.
7. While the proportion of companies adopting AI has plateaued, the companies that have adopted AI continue to pull ahead.
8. Policymaker interest in AI is on the rise.
9. Chinese citizens are among those who feel the most positively about AI products and services. Americans … not so much.

SOURCE: https://aiindex.stanford.edu/report/

©2021, TensorIoT • ALL RIGHTS RESERVED

History of AI



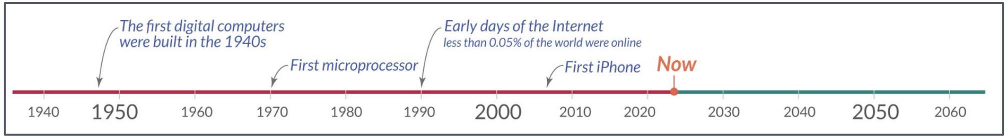
Minerva: built in 2022 and trained on 2.7 billion petaFLOP
Minerva can solve complex mathematical problems at the college level.

PaLM: built in 2022 and trained on 2.5 billion petaFLOP
PaLM can generate high-quality text, explain some jokes, cause & effect, and more.

GPT-3: 2020; 314 million petaFLOP
GPT-3 can produce high-quality text that is often indistinguishable from human writing.

DALL-E: 2021; 47 million petaFLOP
DALL-E can generate high-quality images from written descriptions.

NEO: 2021; 1.1 million petaFLOP
Recommendation systems like Facebook's NEO determine what you see on your social media feed, online shopping, streaming services, and more.



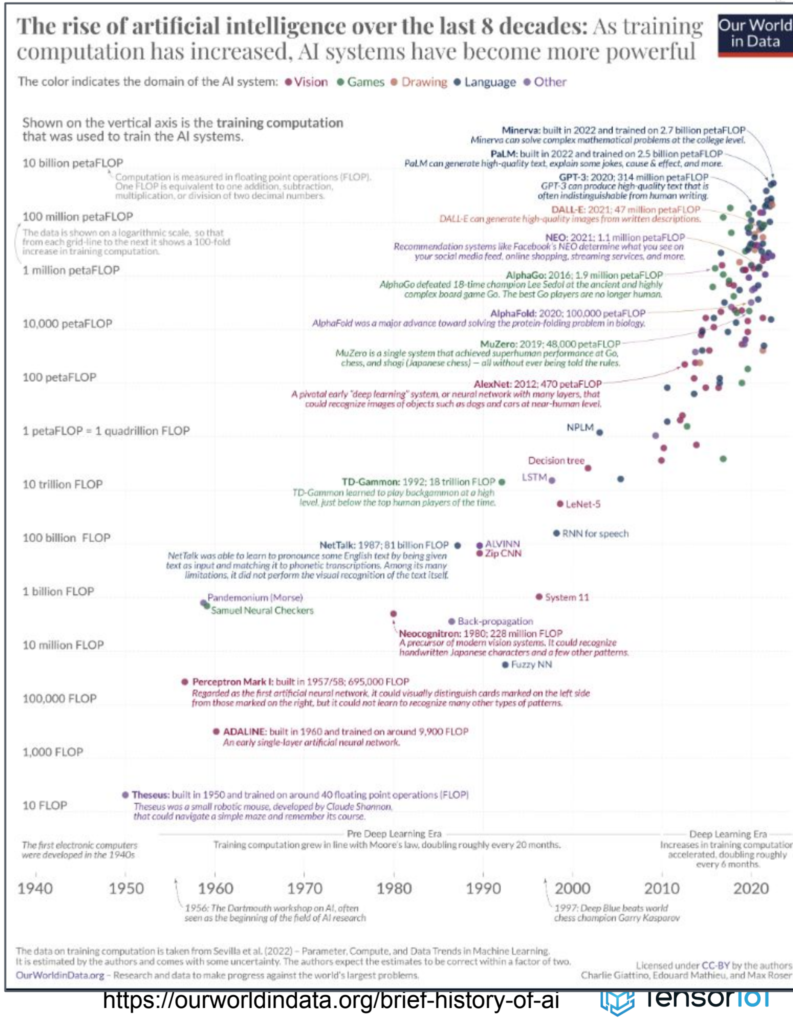
The first digital computers were built in the 1940s

First microprocessor

Early days of the Internet
less than 0.05% of the world were online

First iPhone

Now



The rise of artificial intelligence over the last 8 decades: As training computation has increased, AI systems have become more powerful

Our World in Data

The color indicates the domain of the AI system: ● Vision ● Games ● Drawing ● Language ● Other

Shown on the vertical axis is the **training computation** that was used to train the AI systems.

Computation is measured in floating point operations (FLOP). One FLOP is equivalent to one addition, subtraction, multiplication, or division of two decimal numbers.

The data is shown on a logarithmic scale, so that from each grid-line to the next it shows a 100-fold increase in training computation.

Minerva: built in 2022 and trained on 2.7 billion petaFLOP
Minerva can solve complex mathematical problems at the college level.

PaLM: built in 2022 and trained on 2.5 billion petaFLOP
PaLM can generate high-quality text, explain some jokes, cause & effect, and more.

GPT-3: 2020; 314 million petaFLOP
GPT-3 can produce high-quality text that is often indistinguishable from human writing.

DALL-E: 2021; 47 million petaFLOP
DALL-E can generate high-quality images from written descriptions.

NEO: 2021; 1.1 million petaFLOP
Recommendation systems like Facebook's NEO determine what you see on your social media feed, online shopping, streaming services, and more.

AlphaGo: 2016; 1.9 million petaFLOP
AlphaGo defeated 18-time champion Lee Sedol at the ancient and highly complex board game Go. The best Go players are no longer human.

AlphaFold: 2020; 100,000 petaFLOP
AlphaFold was a major advance toward solving the protein-folding problem in biology.

MuZero: 2019; 48,000 petaFLOP
MuZero is a single system that achieved superhuman performance at Go, chess, and shogi (Japanese chess) — all without ever being told the rules.

AlexNet: 2012; 470 petaFLOP
A pivotal early "deep learning" system, or neural network with many layers, that could recognize images of objects such as dogs and cars at near-human level.

TD-Gammon: 1992; 18 trillion FLOP
TD-Gammon learned to play backgammon at a high level, just below the top human players of the time.

NetTalk: 1987; 81 billion FLOP
NetTalk was able to learn to pronounce some English text by being given text as input and matching it to phonetic transcriptions. Among its many limitations, it did not perform the visual recognition of the text itself.

Neocognitron: 1980; 228 million FLOP
A precursor of modern vision systems. It could recognize handwritten Japanese characters and a few other patterns.

Perceptron Mark I: built in 1957/58; 695,000 FLOP
Regarded as the first artificial neural network, it could visually distinguish cards marked on the left side from those marked on the right, but it could not learn to recognize many other types of patterns.

ADALINE: built in 1960 and trained on around 9,900 FLOP
An early single-layer artificial neural network.

Theseus: built in 1950 and trained on around 40 floating point operations (FLOP)
Theseus was a small robotic mouse, developed by Claude Shannon, that could navigate a simple maze and remember its course.

The first electronic computers were developed in the 1940s

Pre Deep Learning Era
Training computation grew in line with Moore's law, doubling roughly every 20 months.

Deep Learning Era
Increases in training computation accelerated, doubling roughly every 6 months.

1956: The Dartmouth workshop on AI, often seen as the beginning of the field of AI research

1997: Deep Blue beats world chess champion Garry Kasparov

The data on training computation is taken from Sevilla et al. (2022) – Parameter, Compute, and Data Trends in Machine Learning. It is estimated by the authors and comes with some uncertainty. The authors expect the estimates to be correct within a factor of two.

OurWorldinData.org – Research and data to make progress against the world's largest problems.

Licensed under CC-BY by the authors Charlie Giattino, Edouard Mathieu, and Max Roser

https://ourworldindata.org/brief-history-of-ai

Scale - Good / Bad

Here are some of the potential benefits of generative AI:

- It can be used to create new and innovative content.
- It can be used to personalize content for individual users.
- It can be used to automate tasks that are currently done by humans.
- It can be used to generate new insights and ideas.

Here are some of the potential risks of generative AI:

- It could be used to create fake news and propaganda.
- It could be used to generate harmful or offensive content.
- It could be used to automate jobs that are currently done by humans.
- It could lead to job losses and economic disruption.

©2021, TensorIoT • ALL RIGHTS RESERVED

Generative AI: Advisory Services

Unlock the Future with Generative AI

©2021, TensorIoT • ALL RIGHTS RESERVED

Generative AI: Advisory Services

Objectives and Benefits

Overview:

Wizdom AI is an all-inclusive technical framework, enabling you to train and deploy Generative AI, AWS Services, and Foundational Models like Bloom, Titan, and Jurassic within your own AWS account.

Objective:

Discover a high-impact use case and swiftly evaluate a foundational model using your data set, showcasing immediate value to all stakeholders.

Benefits:

- Functional PoV in 2-4 weeks
- Stays in your account giving you full control to the model and your data
- Deployed using out of the box, mature, Foundational Models
- Build on AWS leveraging next-gen services

Wizdom AI

©2021, TensorIoT • ALL RIGHTS RESERVED