AUGUST 8TH, 2023

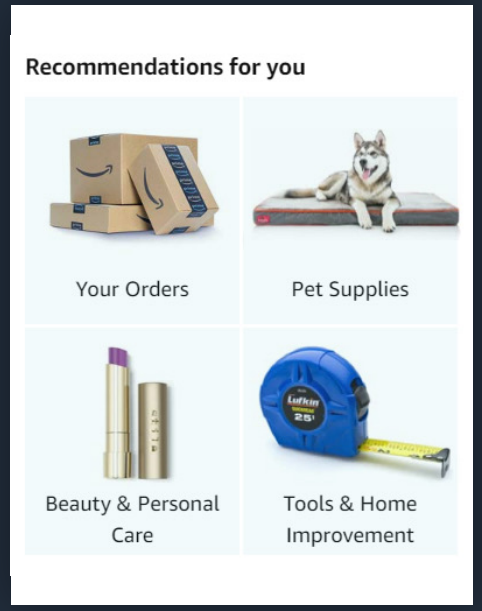
Generative AI on AWS

Keith Johnson

Director Solutions Architecture,
SI Partners, AWS

Innovation can **transform industries**

ML innovation is in Amazon's DNA



4,000 products per minute sold on Amazon.com

1.6M packages every day

Billions of Alexa interactions each week

First Prime Air delivery on **December 7, 2016**

The tipping point for **Generative AI**

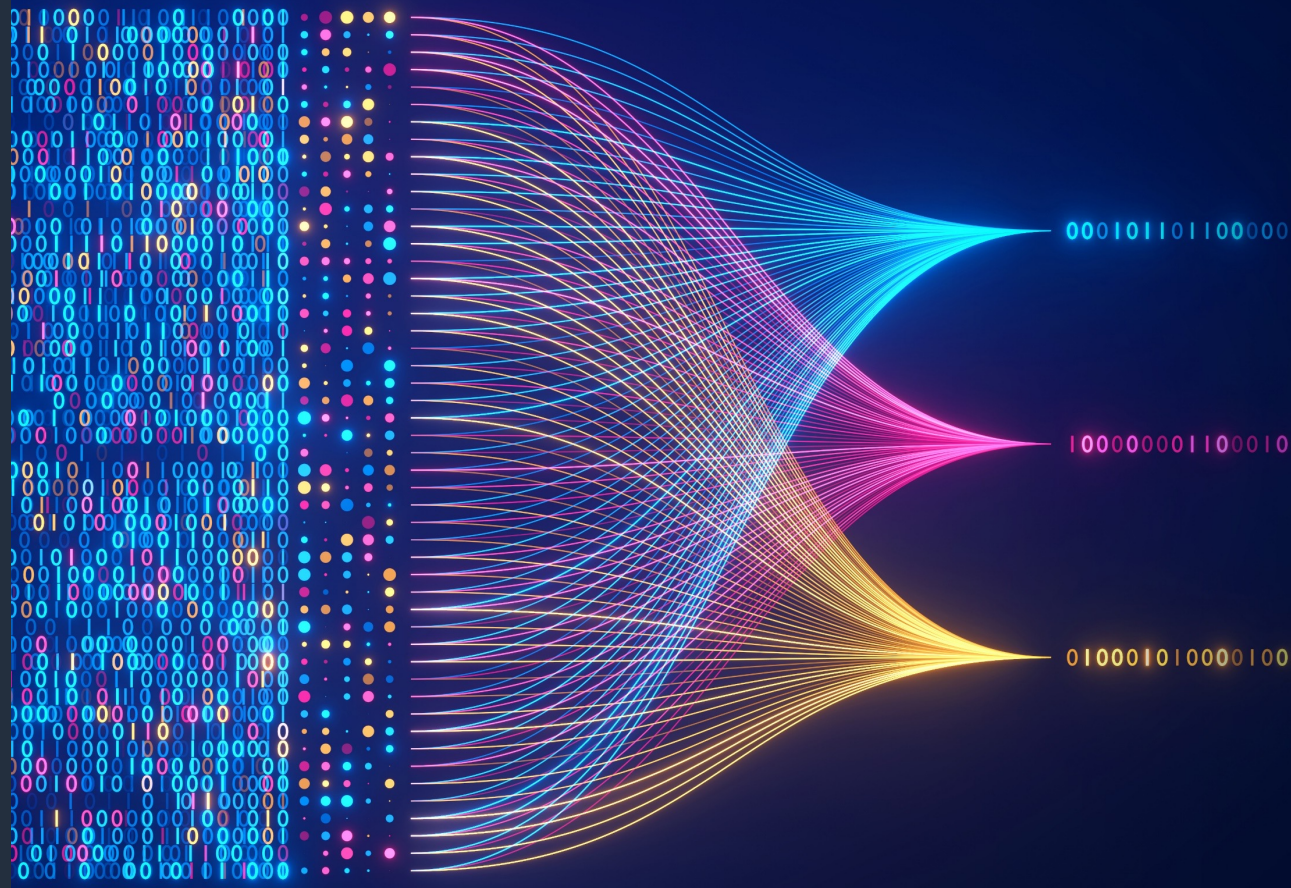
Generative AI is powered by foundation models

Pretrained on vast amounts of unstructured data

Contain large number of parameters that make them capable of learning complex concepts

Can be applied in a wide range of contexts

Customize FMs using your data for domain specific tasks



Generative AI can be used for a wide range of use cases

Enhance citizen experience

CHATBOTS

VIRTUAL ASSISTANTS

AI-POWERED CONTACT CENTER

PERSONALIZATION

Boost productivity

CONVERSATIONAL SEARCH

SUMMARIZATION

CODE GENERATION

DATA TO INSIGHTS

Improve business operations

DOCUMENT PROCESSING

PROCESS OPTIMIZATION

CYBERSECURITY

DATA AUGMENTATION

Creativity & content creation

WRITING

GEOSPATIAL

MOD & SIM

TRAINING

Financial Services

Improve customer loyalty with conversational assistance

Fast track loan approvals to improve profitability

Create personalized financial advice to improve customer service

Education

Save time with automatic generation of questions and answers for exams

Help students find relevant content with concise summaries of lectures and research documents

Automate grading and student performance

Generative AI will transform teaching and learning

Learners and parents

"Create an assignment for students who are struggling with the concept of heat transfer via conduction and convection."

Educators and curriculum designers

Healthcare and Life Sciences

Accelerate drug discovery and research using foundation models to design vaccines, antibodies, and enzymes

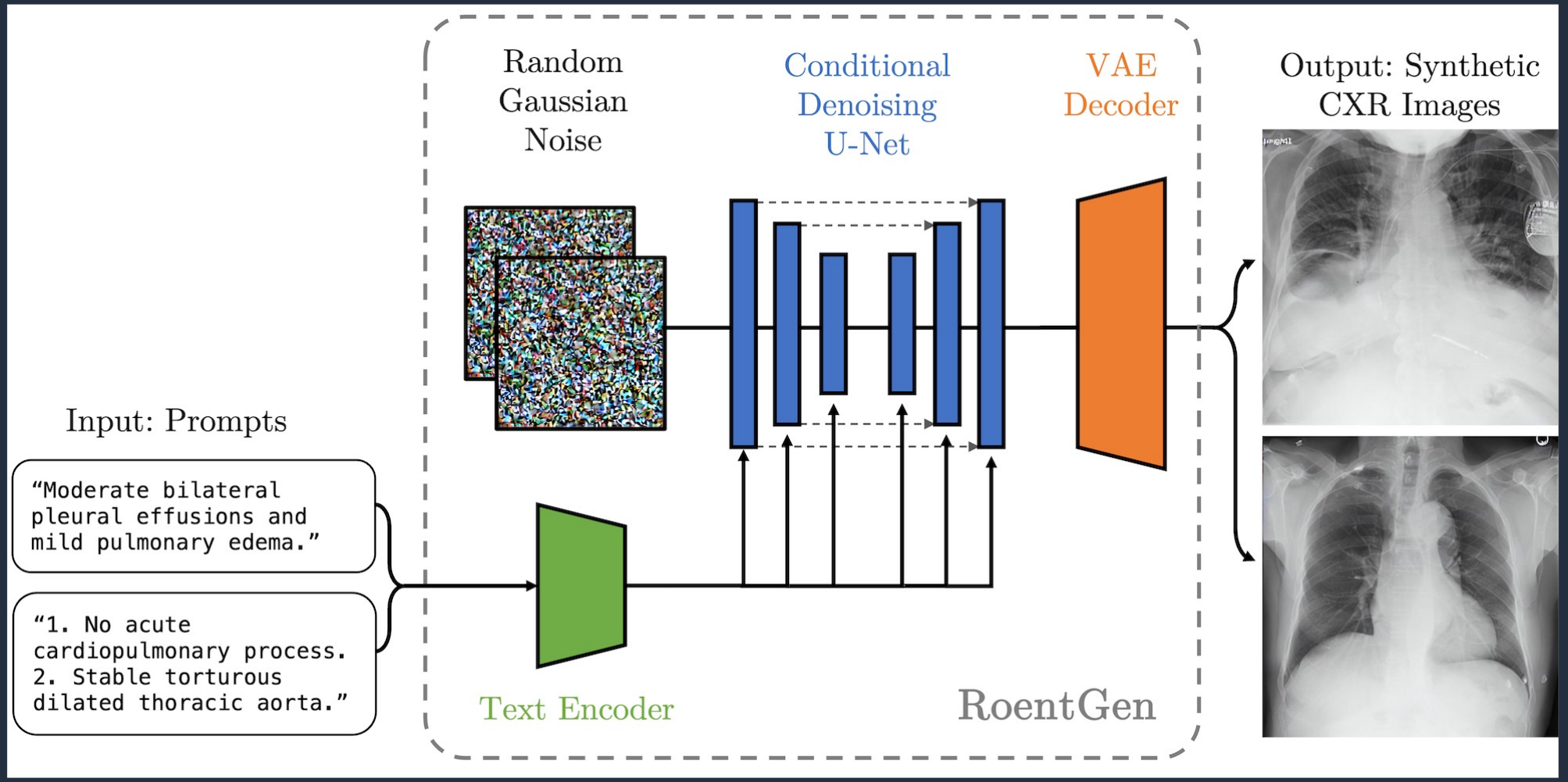
Improve customer support with faster agent assisted claim and document processing

Deliver better care with personalized medicine

Keep patience data safe using synthetic data generation for research

Example Healthcare Research using Foundation Models

Fine-tuning Stable Diffusion for Chest X-ray Generation

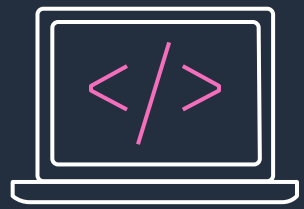


Title: RoentGen: Vision-Language Foundation Model for Chest X-ray Generation
Authors: Pierre Chambon and Christian Bluethgen and Jean-Benoit Delbrouck and Rogier Van der Sluijs and Małgorzata Połacin and Juan Manuel Zambrano Chaves and Tanishq Mathew Abraham and Shivanshu Purohit and Curtis P. Langlotz and Akshay Chaudhari

Aerospace: Design & Engineering

Revolutionize spacecraft design with generative design tools for hardware components

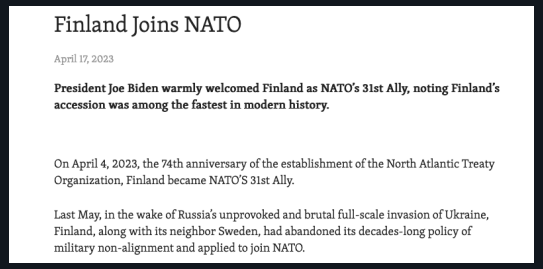
Identify innovative and optimal testing scenarios for digital twins



Reduce risk and technical debt in software development with Amazon CodeWhisperer

Potential Applications for National Security Use Cases

Input ▶ **FM** ▶ **Output**


Finland Joins NATO

April 17, 2023

President Joe Biden warmly welcomed Finland as NATO's 31st Ally, noting Finland's accession was among the fastest in modern history.

On April 4, 2023, the 74th anniversary of the establishment of the North Atlantic Treaty Organization, Finland became NATO'S 31st Ally.

Last May, in the wake of Russia's unprovoked and brutal full-scale invasion of Ukraine, Finland, along with its neighbor Sweden, had abandoned its decades-long policy of military non-alignment and applied to join NATO.

PROMPT:
Why did Finland want to join NATO?

PROMPT:
What is the US stance on NATO?

Text-to-text
Generate text from simple natural-language prompts for various applications

ANSWER:
Russia's invasion of Ukraine

ANSWER:
We will continue to preserve transatlantic security, defend every inch of NATO territory, and meet any and all challenges we face

Text prompt to "paint in" new object into existing overhead image

Multimodal
Diffusion-based model to generate images with natural language prompts

"#function that scrapes a website and stores the data in a s3 buckets"

Text prompt to generate code

Text-to-code
Generate syntactically correct code with natural language prompt

```
#function that scrapes a website and stores the data in a s3 bucket
def scrape_website(url, bucket_name):
    #create a s3 client
    s3_client = boto3.client('s3')
    #use the requests library to get the data from the url
    response = requests.get(url)
    #store the data in a variable
    data = response.text
    #store the data in a s3 bucket
    < 1/4 > Accept Tab Accept Word ⌘ → ⋯ -bucket_name, Key='scraped_data.txt')
    return 'success'
```

Themes we're hearing from our customers

- Protect proprietary content
- Ensure accuracy
- Prevent toxic content
- Address complex bias and fairness issues
- Protect PII and maintain compliance
- Optimize for cost, scalability, and reliability
- Build and maintain competitive differentiation

Unlocking the potential of generative AI

Easiest way to build

with Foundation Models

The most performant

infrastructure for generative AI

Generative AI-powered

applications on AWS

Flexibility to build

your own Foundation Models

aws

Unlocking the potential of generative AI

Easiest way to build
with Foundation Models

The most performant
infrastructure for generative AI

GenAI-powered
applications on AWS

Flexibility to build
your own Foundation Models

aws

Amazon Bedrock

Choice of foundation models

AI21labs

JURASSIC-2

Multilingual LLMs for text generation in Spanish, French, German, Portuguese, Italian, and Dutch

cohere

NEW

COMMAND + EMBED

Text generation model for business applications and embeddings model for search, clustering, or classification in 100+ languages

stability.ai

NEW

STABLE DIFFUSION XL 1.0

Generation of unique, realistic, high-quality images, art, logos, and designs

ANTHROP\C

NEW

CLAUDE 2

LLM for thoughtful dialogue, content creation, complex reasoning, creativity, and coding, based on Constitutional AI and harmlessness training

amazon

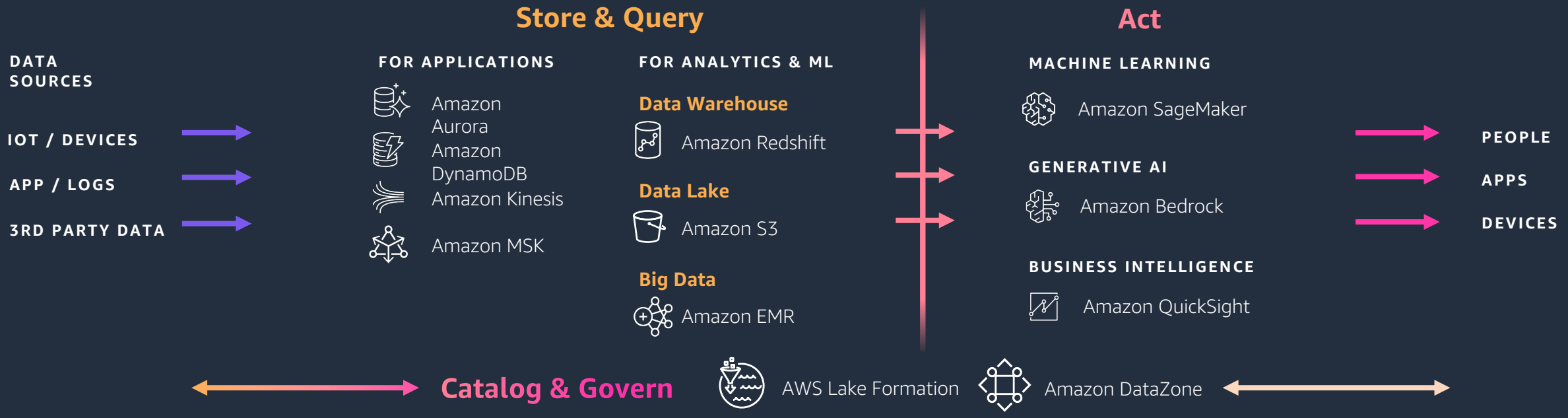
AMAZON TITAN

Text summarization, generation, classification, open-ended Q&A, information extraction, embeddings and search

aws

Your data is
your differentiator

A comprehensive set of services for your data foundation



Unlocking the potential of generative AI

Easiest way to build

with Foundation Models

The most performant

infrastructure for generative AI

GenAI-powered

applications on AWS

Flexibility to build

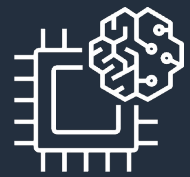
your own Foundation Models

aws

Deep investments
in global infrastructure

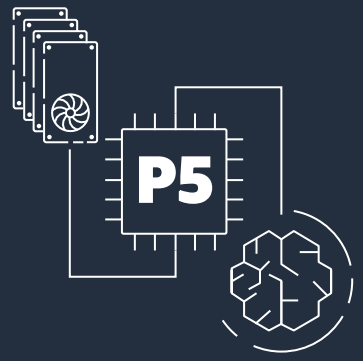
Broad choice of ML accelerators

High performance, low-cost ML infrastructure



10+ years of silicon innovation

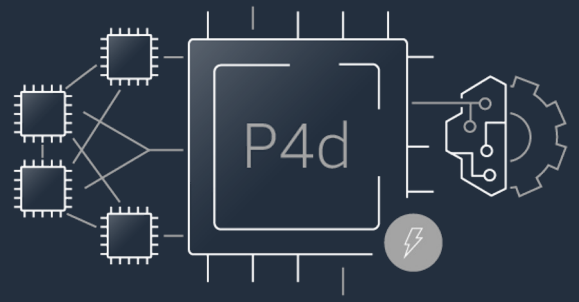
Amazon EC2 P5 instances



Powered by NVIDIA H100 Tensor Core GPUs

Up to 6x faster and up to 40% cost-to-train savings
than previous generation GPU-based instances

Amazon EC2 P4d/P4de instances



Powered by NVIDIA A100 Tensor Core GPUs

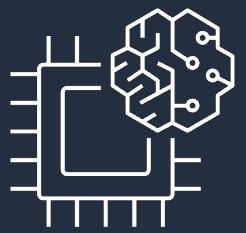
Up to 2.5x faster and up to 60% lower training costs
than previous generation P3/P3dn instances

Amazon EC2 G5 instances

Powered by NVIDIA A10G Tensor Core GPUs

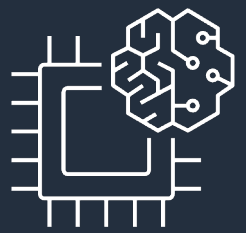
Up to 3.3x higher performance
than previous generation G4dn instances

Purpose-built accelerators
for generative AI



AWS Trainium

Up to 50% savings on training costs
over comparable Amazon EC2 instances



AWS Inferentia2

Up to 40% better price performance
than comparable Amazon EC2 instances

PURPOSED-BUILT ACCELERATORS

Reduced latency while improving efficiency and lowering costs

aws

Unlocking the potential of generative AI

Easiest way to build

with Foundation Models

The most performant

infrastructure for generative AI

GenAI-powered

applications on AWS

Flexibility to build

your own Foundation Models

GENERALLY AVAILABLE

Amazon **CodeWhisperer**

Build apps faster and more securely with an AI coding companion

Generate code suggestions in real-time

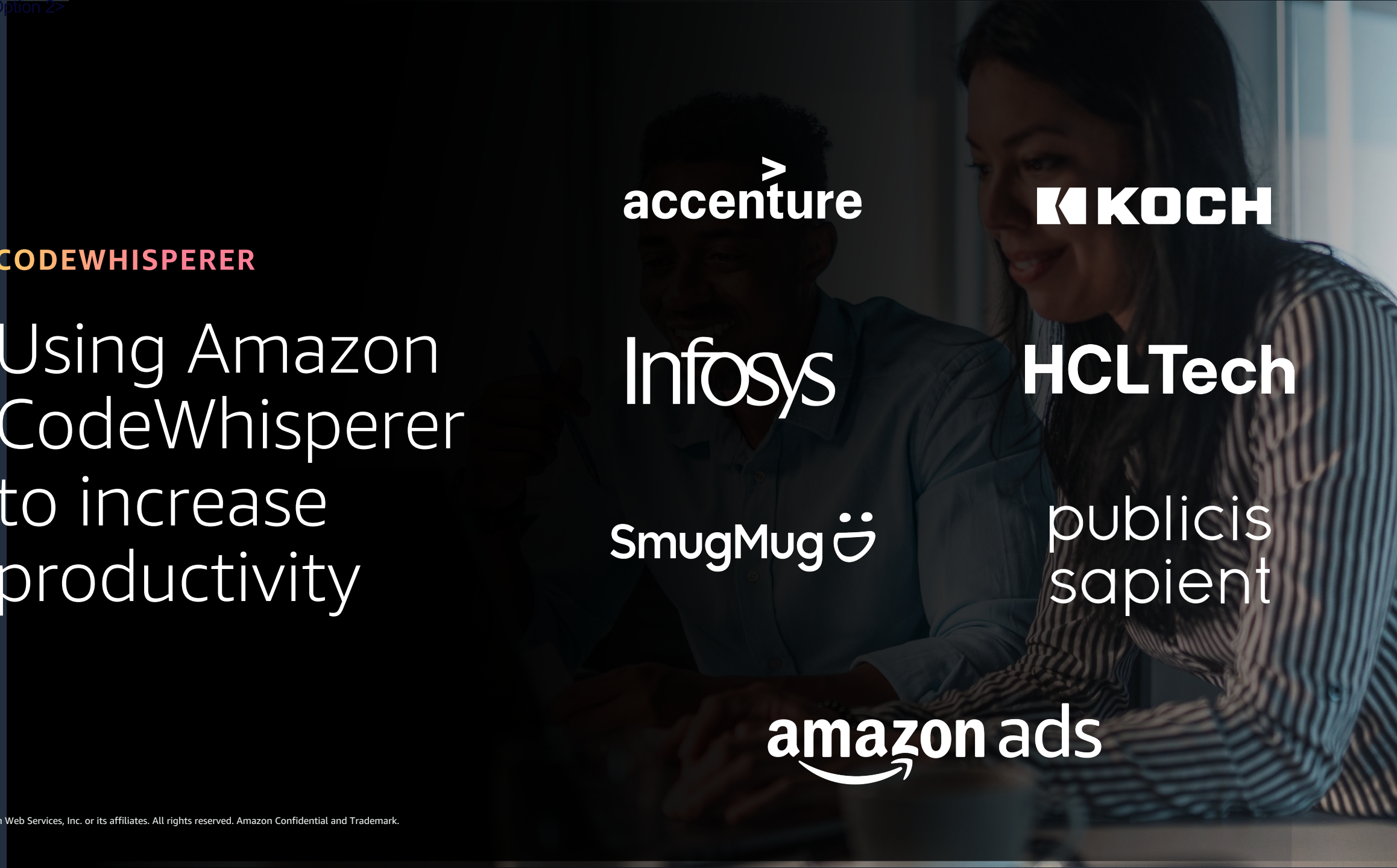
Scan code for hard-to-find vulnerabilities

Flag code that resembles open-source training data or filter by default

FREE FOR INDIVIDUAL TIER

© 2023, Amazon Web Services, Inc. or its affiliates.

<Slide: Customer Option 2>



CODEWHISPERER

Using Amazon CodeWhisperer to increase productivity

accenture

KOCH

Infosys

HCLTech

SmugMug

publicis sapient

amazon ads

NEW

AWS HealthScribe

A HIPAA-eligible automatic note generation service for clinical applications

IN PREVIEW TODAY

- Enhances clinical productivity
- Enables AI to be used responsibly in clinical settings
- Includes built-in security, privacy, and compliance features

Text-based audio transcript

Skip small talk

Clinician	So, what brings you to see me today?
Patient	Uh, I've noticed this hard lump behind my knee and I'm not sure what caused it, but it's not going away.
Clinician	Okay.
Clinician	Is there any pain in the lump?
Patient	Uh, no. No pain.
Patient	But I can feel it back there when I move.
Clinician	Okay. So, a painless hard subcutaneous mass in the popliteal fossa. How long have you had this lump?
Patient	Oh, I noticed it about three months ago.
Clinician	Okay.
Clinician	Do you play sports?
Patient	Yeah.
Patient	I play soccer a few times a week. Uh yeah. I try to play baseball a couple of times a month if I can.

CLINICAL NOTES

Chief Complaint:

- Hard lump behind the knee

History of Present Illness:

- The patient is seen today for a hard lump behind their knee.
- The patient is not sure what caused the lump, but it has not gone away since they first noticed it 3 months ago.
- They can feel the lump when they move, and deny any pain. Regular activities include playing soccer a few times a week and baseball a couple times a month.

Assessment:

- Baker's cyst

Plan for the condition of Baker's cyst:

- Apply ice to the knee and avoid strenuous activity.
- Use a compression wrap on the knee to help reduce swelling and to elevate the knee during rest.
- If the issue persists, a Cortisone injection will be considered to prevent further fluid accumulation

AI-generated insights

aws

AWS Continues to Invest in Generative AI

- Agents for **Amazon Bedrock**
- AWS **HealthScribe**
- Amazon **CodeWhisperer**
- Vector Engine for Amazon **OpenSearch Serverless**
- Amazon EC2 **P5 Instances**
- Generative BI in Amazon **QuickSight**

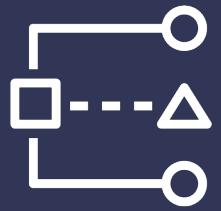
- Plus new models and model providers!

Unlocking the potential of generative AI

Easiest way to build
with Foundation Models

The most performant
infrastructure for generative AI

GenAI-powered
applications on AWS



Flexibility to build
your own Foundation Models

Build with publicly available foundation models

AVAILABLE ON SAGEMAKER JUMPSTART

AI21labs

Models
Jurassic-2 Ultra, Mid
Contextual answers
Summarize
Paraphrase
Grammatical error correction

Tasks
Text generation
Long-form generation
Summarization
Paraphrasing
Chat
Information extraction

Meta AI

Models
Llama 2 7B, 13B, 70B

Tasks
Question answering
Chat
Summarization
Paraphrasing
Sentiment analysis
Text generation

cohere

Models
Cohere
Command XL

Tasks
Text generation
Information extraction
Question answering
Summarization

Hugging Face

Models
Falcon-7B, 40B
Open LlaMA
RedPajama
MPT-7B
BloomZ 176B
Flan T-5 models (8 variants)
DistilGPT2
GPT NeoXT
Bloom models (3 variants)

Tasks
Machine translation
Question answering
Summarization

stability.ai

Models
Stable Diffusion XL 1.0
2.1 base
Upscaling
Inpainting

Tasks
Generate photo-realistic images from text input
Improve quality of generated images

Features
Fine-tuning on Stable Diffusion 2.1 base model

LightOn

Models
Lyra-Fr
10B, Mini

Tasks
Text generation
Keyword extraction
Information extraction
Question answering
Summarization
Sentiment analysis
Classification

databricks

Models
Dolly

Tasks
Question answering
Chat
Summarization
Paraphrasing
Sentiment analysis
Text generation

alexa

Models
AlexaTM 20B

Tasks
Machine translation
Question answering
Summarization
Annotation
Data generation

Unlocking the potential of generative AI

Easiest way to build

with Foundation Models

The most performant

infrastructure for generative AI

GenAI-powered

applications on AWS

Flexibility to build

your own Foundation Models

aws

Start your generative AI journey today

Deploy Amazon CodeWhisperer for immediate productivity gains and explore FMs through Amazon Bedrock

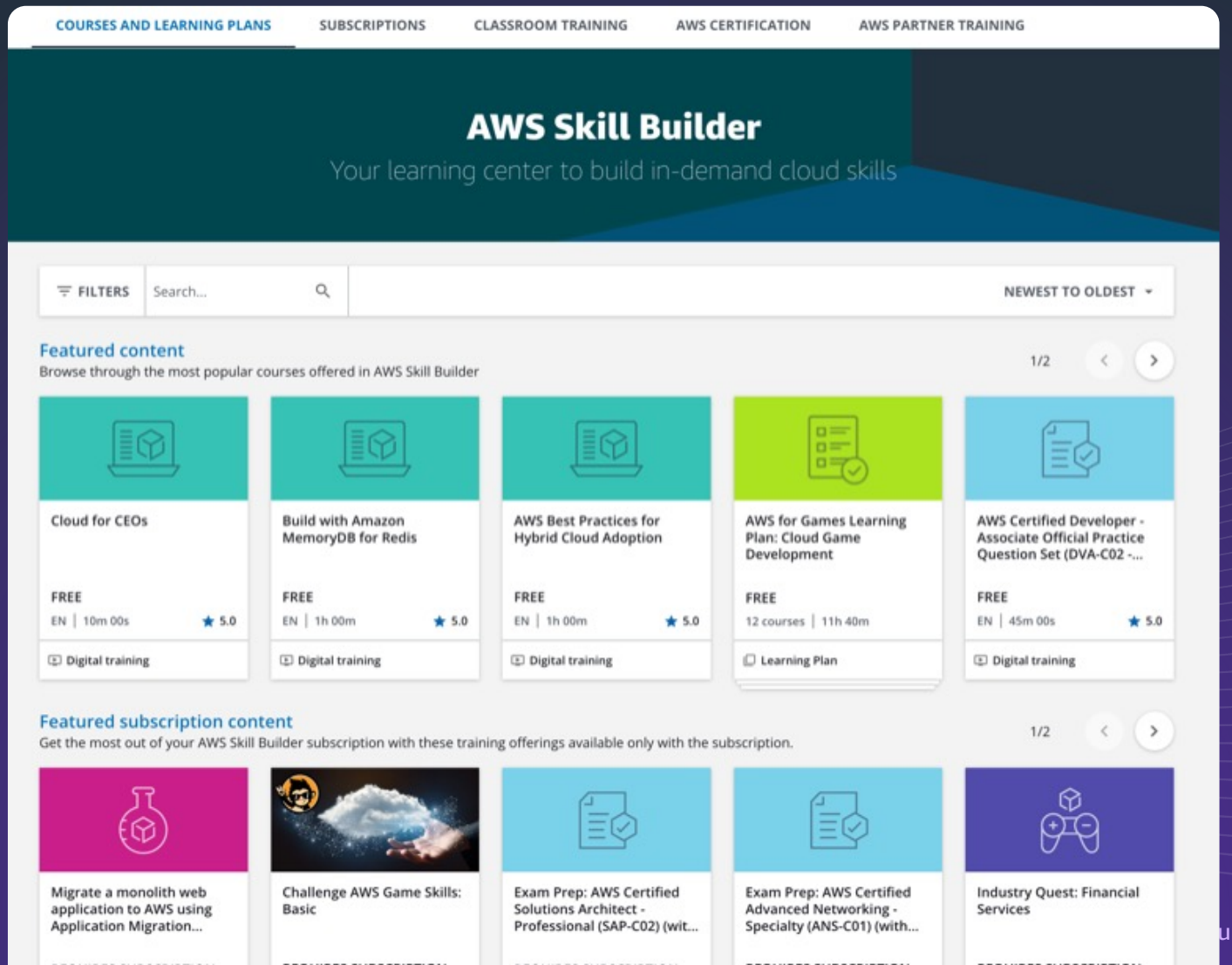
Start your generative AI journey today

Deploy Amazon CodeWhisperer for immediate productivity gains and explore FMs through Amazon Bedrock

Empower developers of all skill levels through a variety of training opportunities

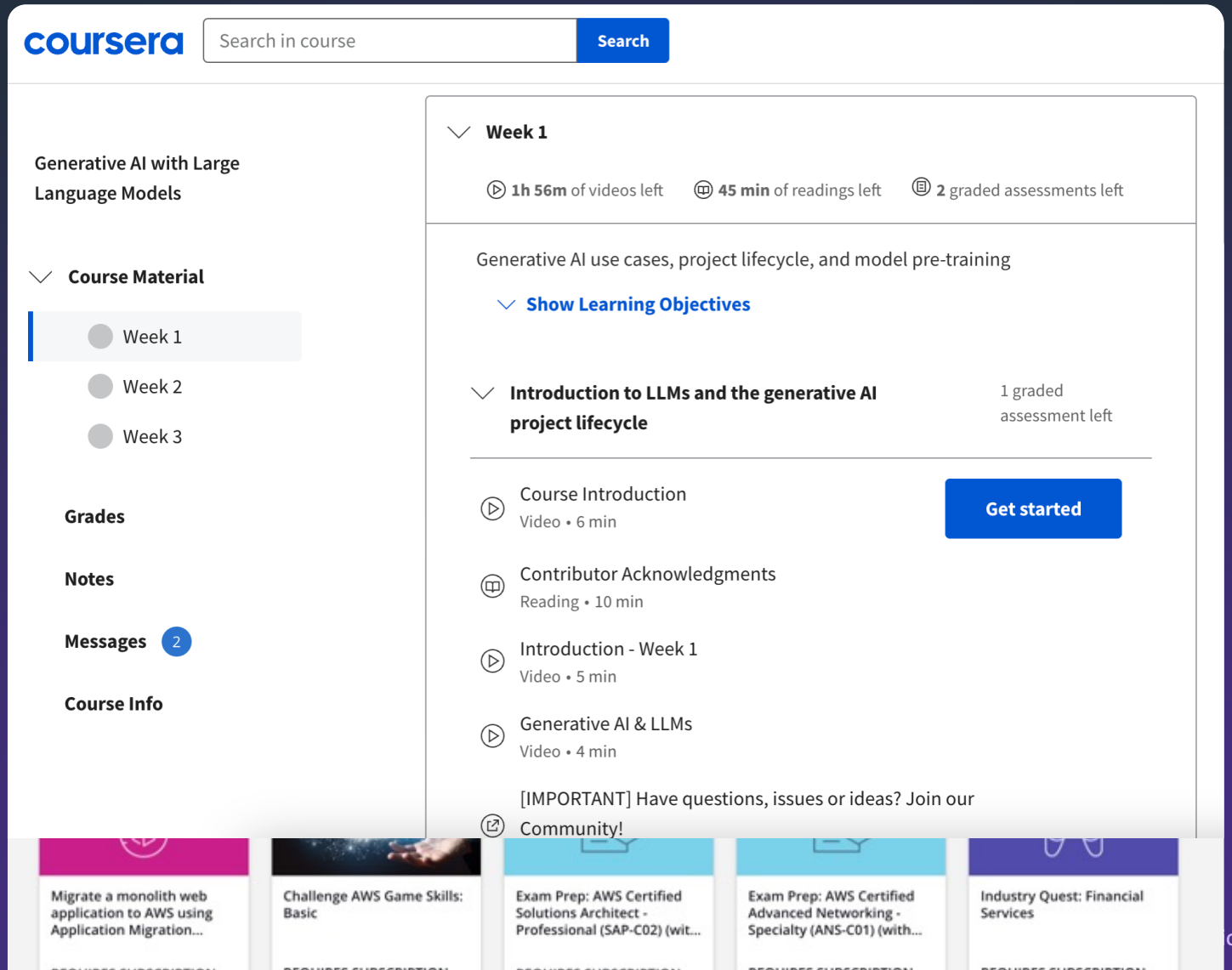
SKILLBUILDER.AWS

600+ digital courses



Hands-on generative AI training opportunities

Learn the fundamentals of generative AI for real-world applications



aws

c Sector

Start your generative AI journey today

Deploy Amazon CodeWhisperer for immediate productivity gains and explore FMs through Amazon Bedrock

Empower developers of all skill levels through a variety of training opportunities

Get started on a PoC for your top use cases

AWS Generative AI Innovation Center

Everything you need to accelerate **your generative AI journey**

Foundation models to build your applications

Generative AI services to enhance productivity

Purpose-built ML infrastructure for low latency and reduced costs

Education and training to accelerate employee productivity

Reimagine and transform your applications **with generative AI on AWS**

View product page

Get started with foundation models

Watch deep dive demo videos

Read blog posts

Hands on workshop

Thank you!

Keith Johnson
kjohnz@amazon.com