AWS Gen AI Services

NITTTR, Chandigarh

Jan 31st, 2024

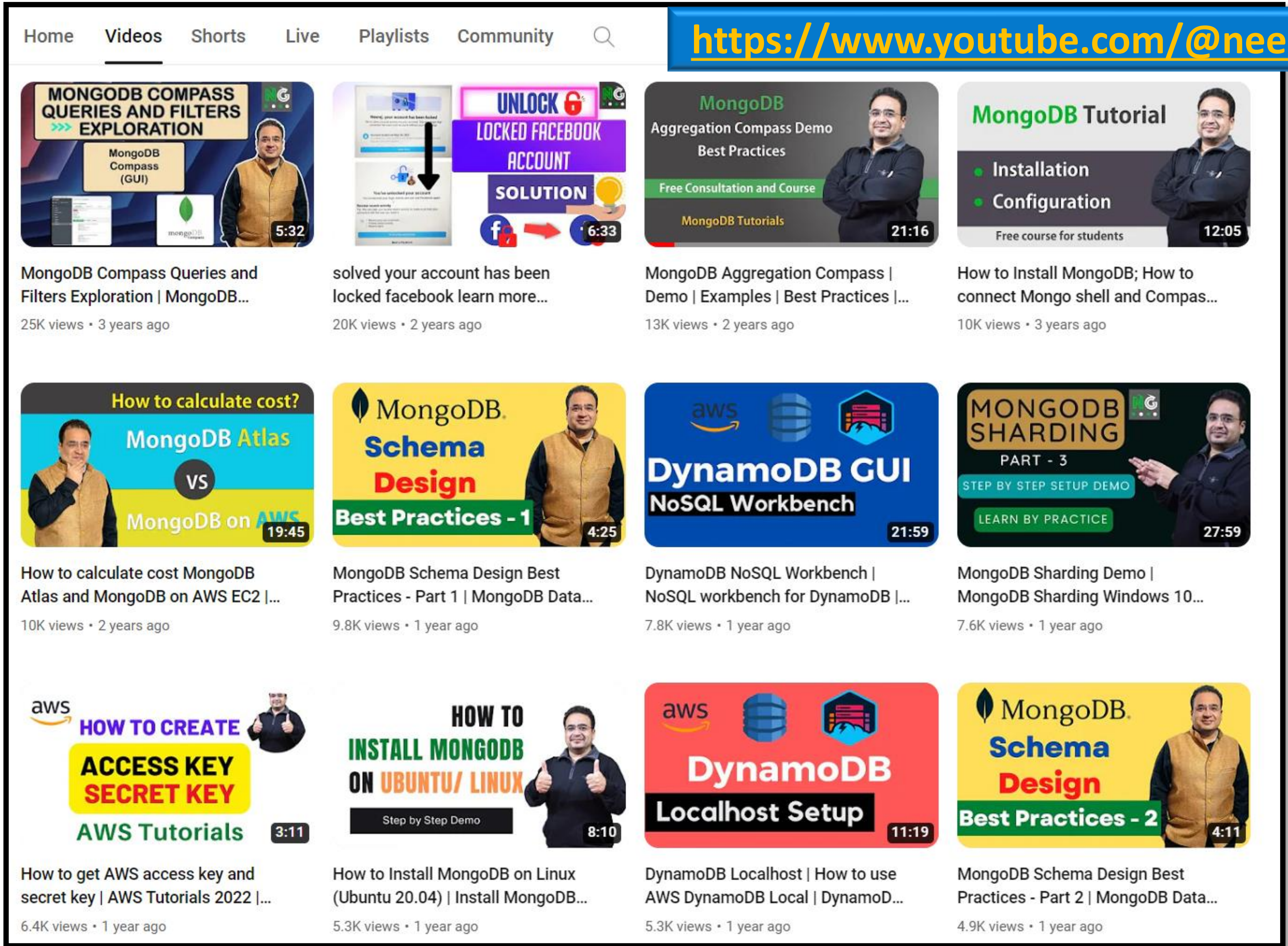
Session Guidelines, please

1. keep yourself mute until you have query/input
2. raise hand to ask a query
3. participate in the session by sharing your inputs, feedback etc.
4. Install prerequisites - https://github.com/neerajg5/gen-ai-aws/blob/main/prerequisites.pdf

About Me

- 19+ years of experience in IT industry
- Design end to end solutions including Web and Big Data
- Multi Cloud Certified (AWS, GCP, Azure)
- Data modelling in NoSQL database : MongoDB and DynamoDB
- Exploring AI/Gen AI
- Interested in research contribution
- https://www.linkedin.com/in/neerajgarg5/
- https://learnwithneeraj.com/
- https://www.youtube.com/@neerajgarg

https://www.youtube.com/@neerajgarg



MongoDB Compass Queries and Filters Exploration | MongoDB...
25K views • 3 years ago

solved your account has been locked facebook learn more...
20K views • 2 years ago

MongoDB Aggregation Compass | Demo | Examples | Best Practices |...
13K views • 2 years ago

How to Install MongoDB; How to connect Mongo shell and Compas...
10K views • 3 years ago

How to calculate cost MongoDB Atlas and MongoDB on AWS EC2 |...
10K views • 2 years ago

MongoDB Schema Design Best Practices - Part 1 | MongoDB Data...
9.8K views • 1 year ago

DynamoDB NoSQL Workbench | NoSQL workbench for DynamoDB |...
7.8K views • 1 year ago

MongoDB Sharding Demo | MongoDB Sharding Windows 10...
7.6K views • 1 year ago

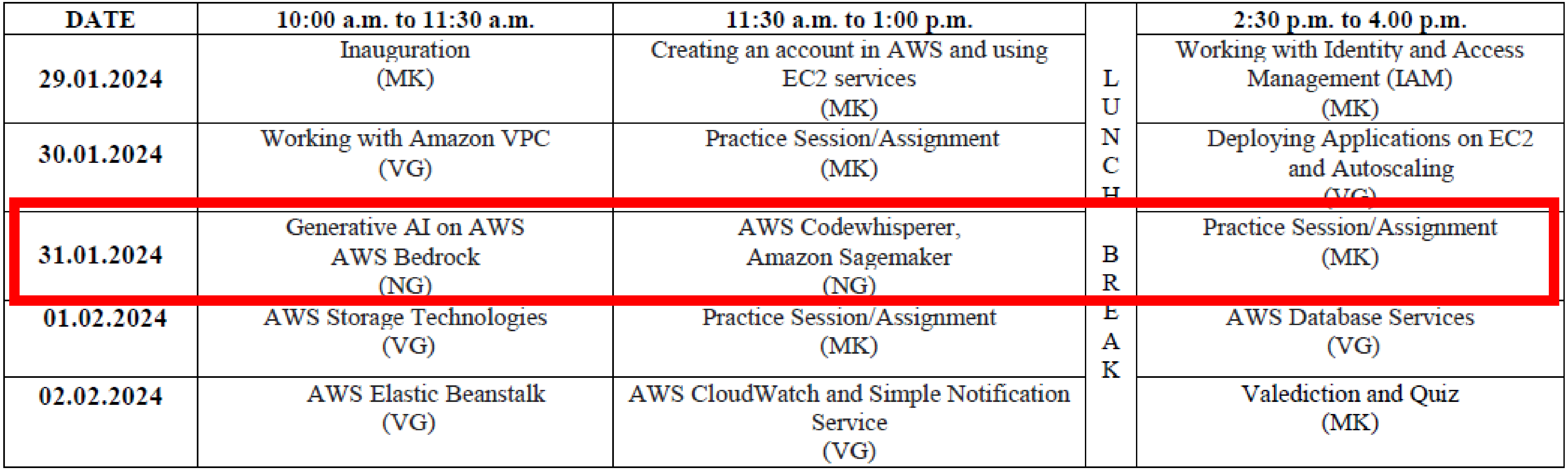
How to get AWS access key and secret key | AWS Tutorials 2022 |...
6.4K views • 1 year ago

How to Install MongoDB on Linux (Ubuntu 20.04) | Install MongoDB...
5.3K views • 1 year ago

DynamoDB Localhost | How to use AWS DynamoDB Local | DynamoD...
5.3K views • 1 year ago

MongoDB Schema Design Best Practices - Part 2 | MongoDB Data...
4.9K views • 1 year ago

Agenda



DATE	10:00 a.m. to 11:30 a.m.	11:30 a.m. to 1:00 p.m.		2:30 p.m. to 4.00 p.m.
29.01.2024	Inauguration (MK)	Creating an account in AWS and using EC2 services (MK)	LUNCH BREAK	Working with Identity and Access Management (IAM) (MK)
30.01.2024	Working with Amazon VPC (VG)	Practice Session/Assignment (MK)		Deploying Applications on EC2 and Autoscaling (VG)
31.01.2024	Generative AI on AWS AWS Bedrock (NG)	AWS Codewhisperer, Amazon Sagemaker (NG)		Practice Session/Assignment (MK)
01.02.2024	AWS Storage Technologies (VG)	Practice Session/Assignment (MK)		AWS Database Services (VG)
02.02.2024	AWS Elastic Beanstalk (VG)	AWS CloudWatch and Simple Notification Service (VG)		Valediction and Quiz (MK)

Prerequisite, reference and other instruction : https://github.com/neerajg5/gen-ai-aws

Prerequisites

- AWS Account
- AWS Builder ID
- Visual Studio Code

https://github.com/neerajg5/gen-ai-aws/blob/main/prerequisites.pdf

Please help me know more about you..

- Your name
- your institute name
- experience (in years)
- area of specialization
- what is your expectation from this session

Quiz - How much knowledge do you have on Gen AI?

1. Just heard about it
2. Beginner Level – read on internet
3. Intermediate Level – Used in your tasks
4. Advanced level - Research

Quiz – Is every AI use case solvable with Gen AI?

1. Yes

2. No

3. Not sure

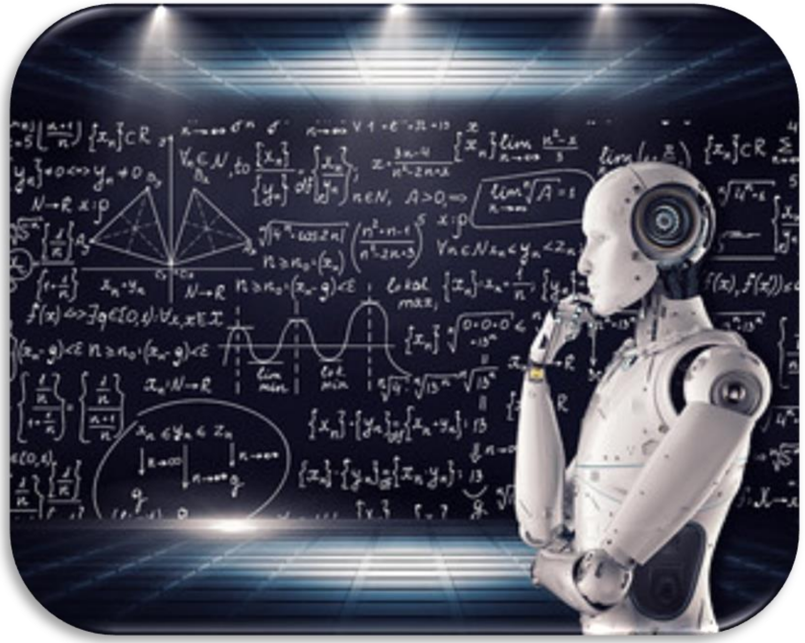
4. Not now, may be in future

Give 2-3 examples of Generative AI

Generative AI – An overview

What is Generative AI?

- The technology which leverages AI to generate different forms of content (Music, Text summary, Translation, Images, Scripts, Analysis etc.)
- Based on Transformers or LLM (**Large Language Models**)
- Some applications in daily use include:
  - Translation
  - Voice commands (Siri, Alexa, Google etc.)



How can **Naive** use it?

Endless possibilities in various fields

- Music Generation
- Text/ Audio/ Video summary
- Translation (between languages)
- Create Images
- Create videos
- Many more..

How can **Experts** use it?

1. Start small with a MVP (**Minimum Viable Product**)
2. Use APIs to build utilities/ products example during demo

Advance use

- Train foundation models on their own data
- Develop affordable applications/ products
- Address real world problems (healthcare, poverty, education, global warming, ESG - Environmental, social, and corporate governance etc.)

A lot of research problems within Gen AI space

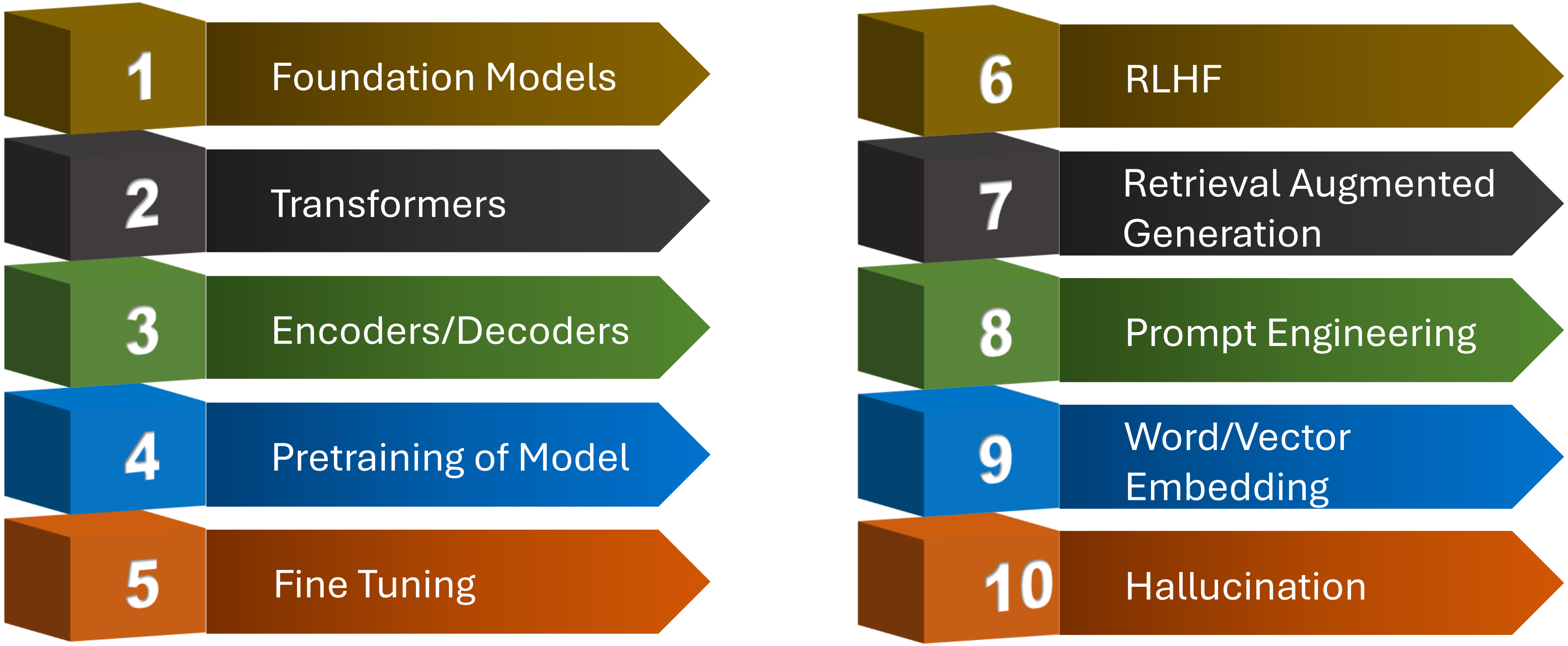
- Removing Bias
- Affordable cost
- Accuracy
- Security
- Many more..

Demos

1. Chat GPT
2. gamma.app

This Photo by Unknown Author is licensed under CC BY-NC-ND

Key Concepts/Terms



1. Foundation Models
2. Transformers
3. Encoders/Decoders
4. Pretraining of Model
5. Fine Tuning
6. RLHF
7. Retrieval Augmented Generation
8. Prompt Engineering
9. Word/Vector Embedding
10. Hallucination

Prompt Engineering

- https://www.promptingguide.ai/

Foundation Models/ Large Language Models (LLMs)

What is a Foundation Model ?

1 Train one model on a huge amount of data

2 Adapt it to many applications

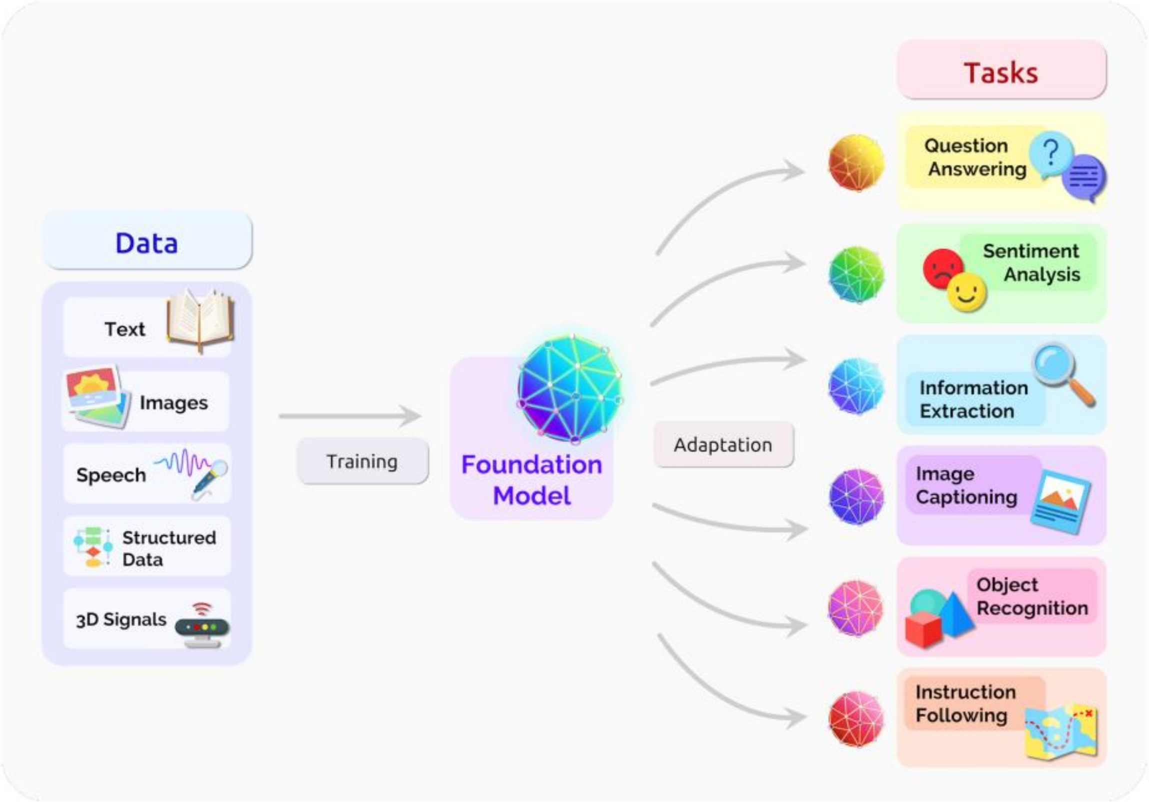
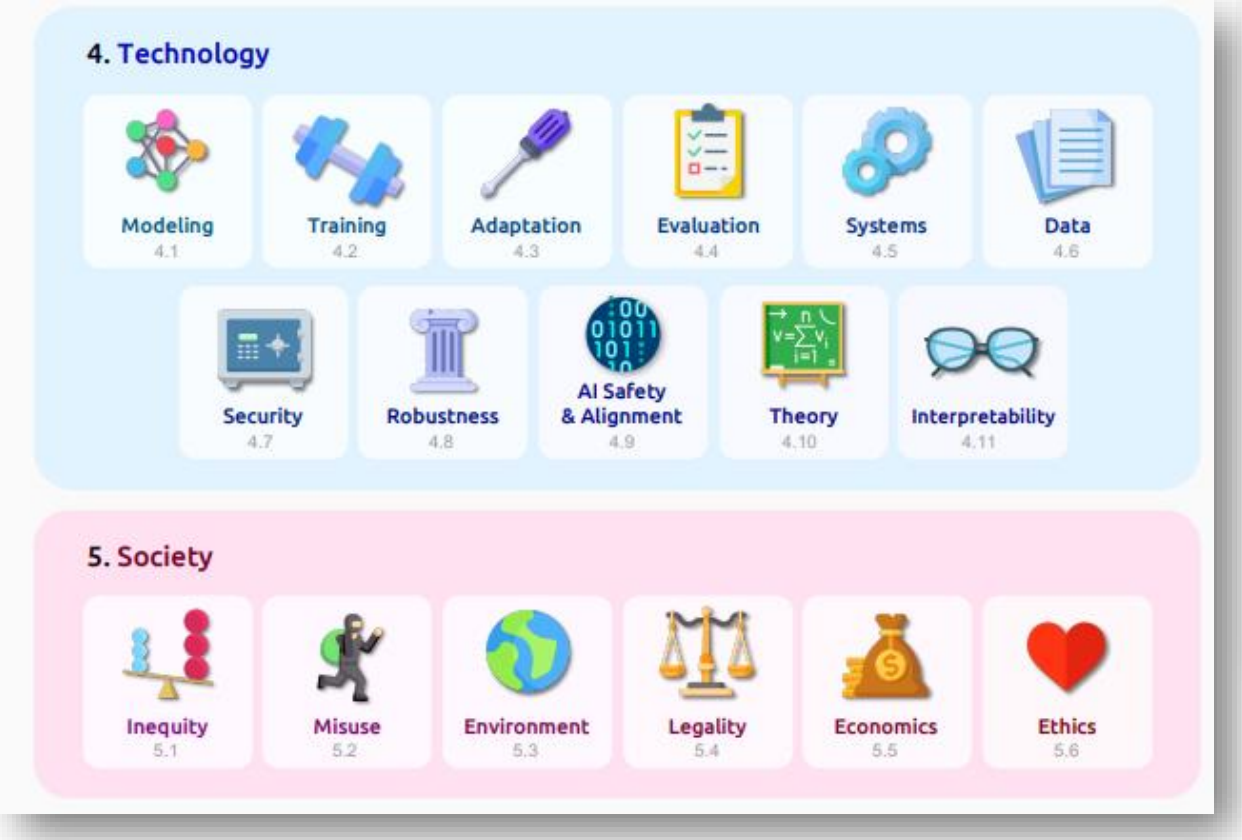


Fig. 2. A foundation model can centralize the information from all the data from various modalities. This one model can then be adapted to a wide range of downstream tasks.

https://crfm.stanford.edu/

Opportunities & Risks of Foundation Models – An overview



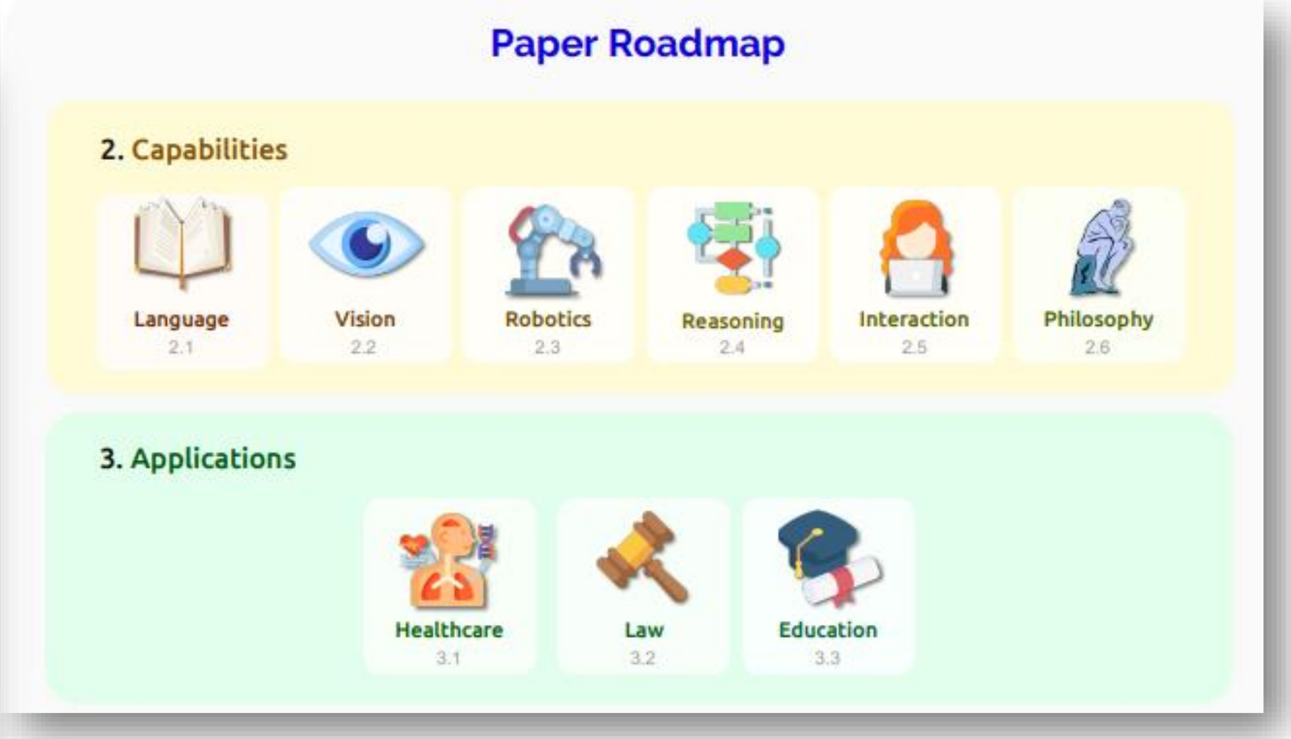
4. Technology

Modeling 4.1 · Training 4.2 · Adaptation 4.3 · Evaluation 4.4 · Systems 4.5 · Data 4.6

Security 4.7 · Robustness 4.8 · AI Safety & Alignment 4.9 · Theory 4.10 · Interpretability 4.11

5. Society

Inequity 5.1 · Misuse 5.2 · Environment 5.3 · Legality 5.4 · Economics 5.5 · Ethics 5.6



Paper Roadmap

2. Capabilities

Language 2.1 · Vision 2.2 · Robotics 2.3 · Reasoning 2.4 · Interaction 2.5 · Philosophy 2.6

3. Applications

Healthcare 3.1 · Law 3.2 · Education 3.3

Extract from Report by Center for Research on Foundation Models (CRFM) Stanford Institute

https://crfm.stanford.edu/assets/report.pdf

Role of Human Beings

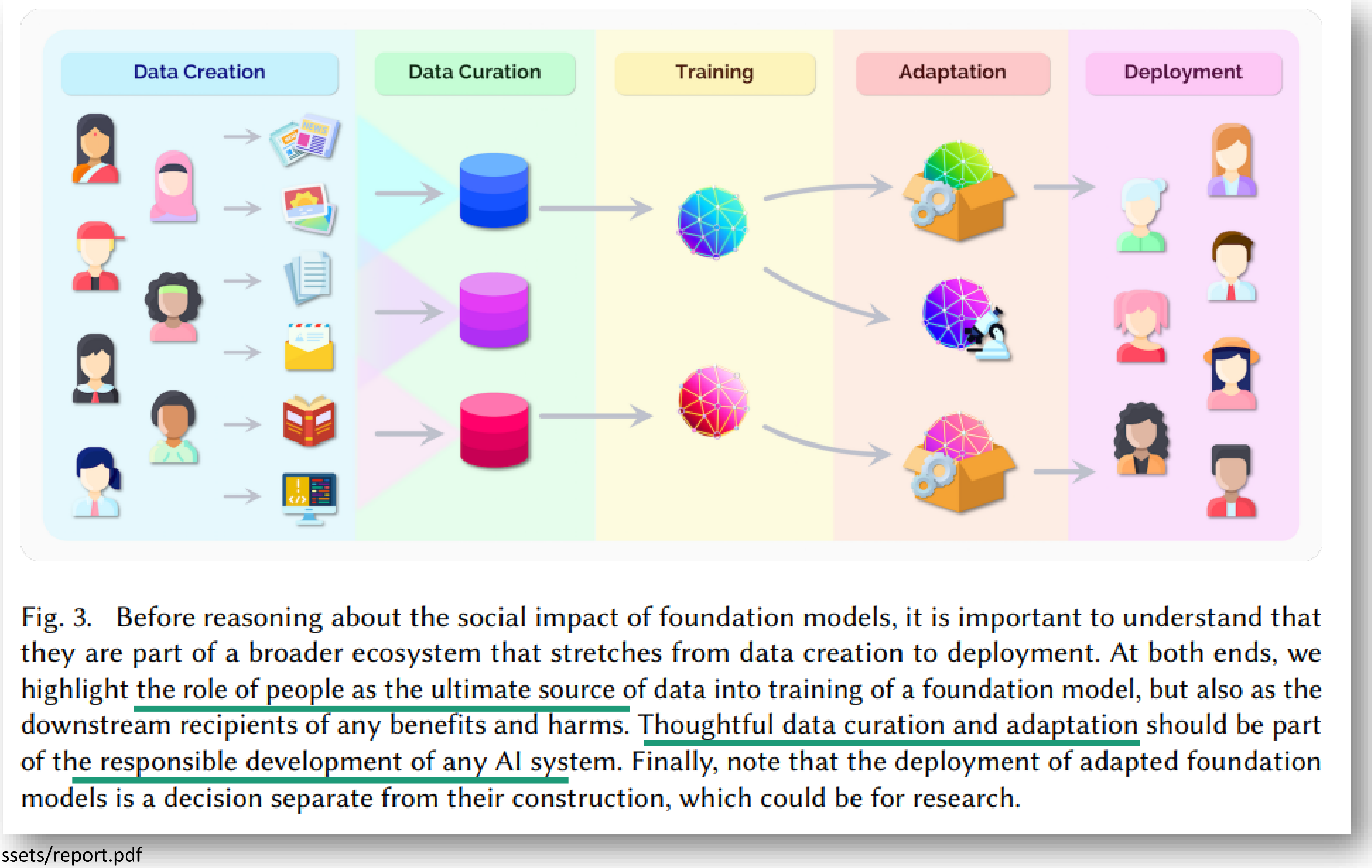


Fig. 3. Before reasoning about the social impact of foundation models, it is important to understand that they are part of a broader ecosystem that stretches from data creation to deployment. At both ends, we highlight the role of people as the ultimate source of data into training of a foundation model, but also as the downstream recipients of any benefits and harms. Thoughtful data curation and adaptation should be part of the responsible development of any AI system. Finally, note that the deployment of adapted foundation models is a decision separate from their construction, which could be for research.

https://crfm.stanford.edu/assets/report.pdf

Coverage of Foundation Models

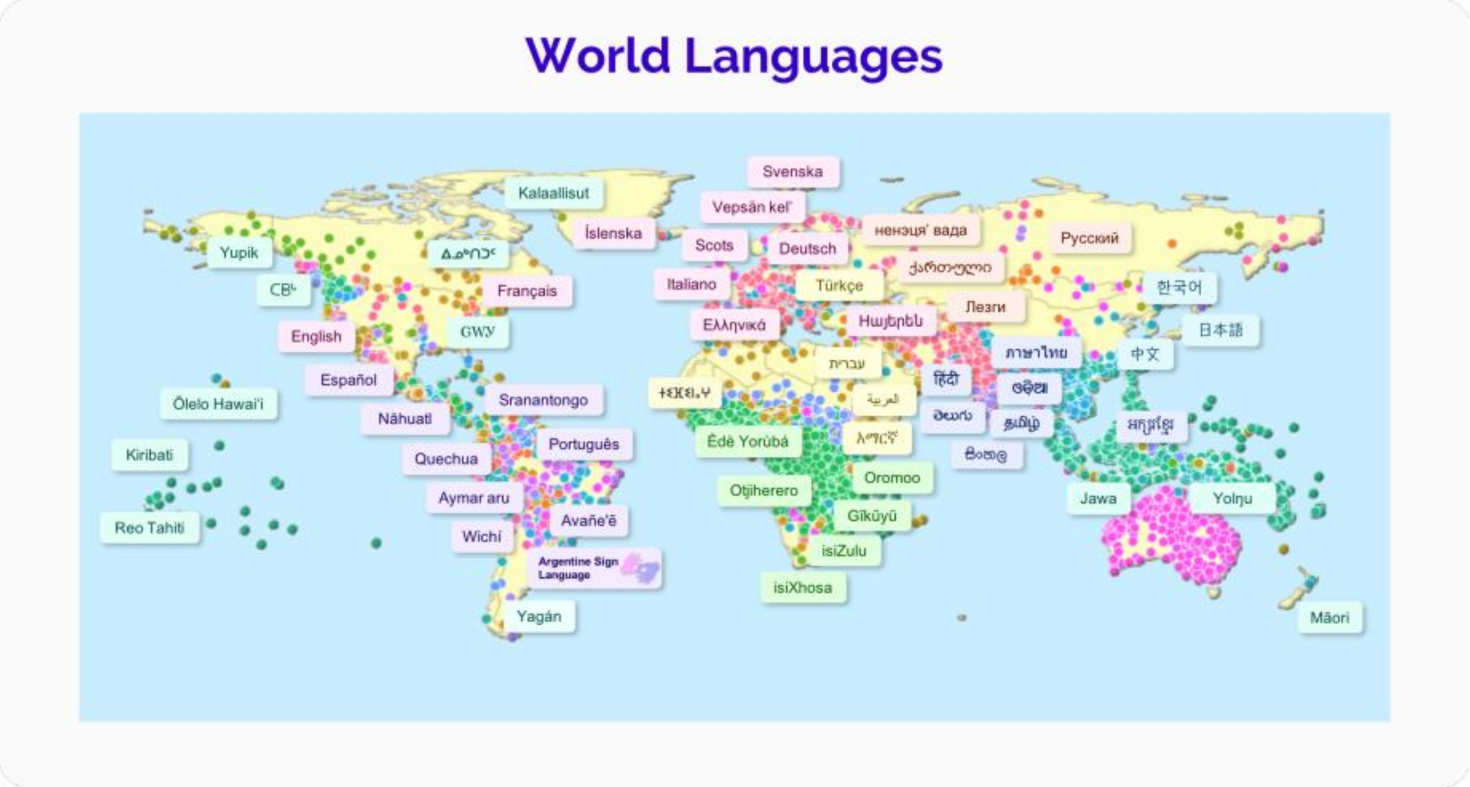


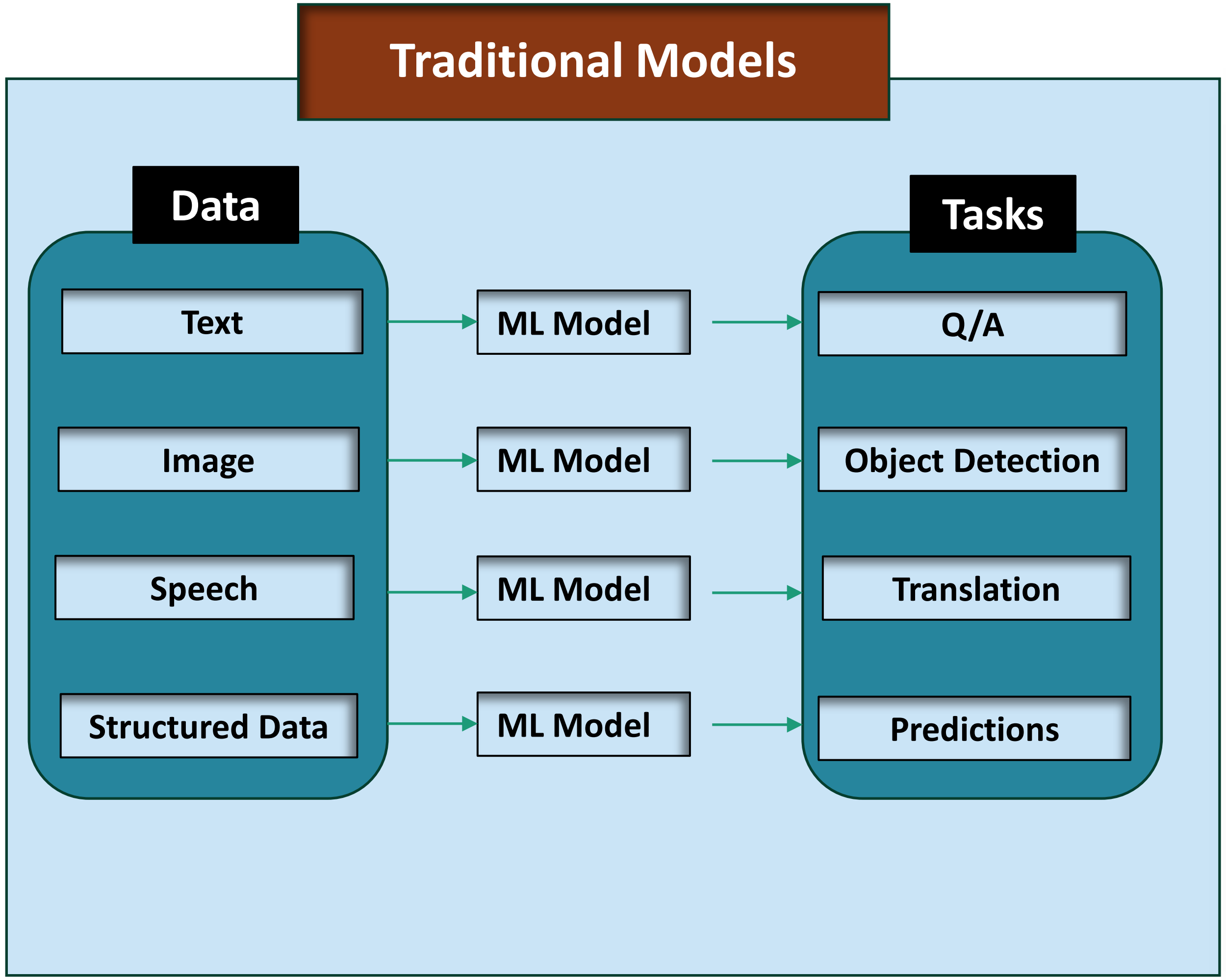
Fig. 5. Only a tiny percentage of the world's languages are currently represented in foundation models. There are over 6,000 languages in the world, with estimates varying due to the inherent uncertainty of what constitutes a separate language [Nordhoff and Hammarström 2011]. This map shows the languages of the world, with each dot representing one language and its color indicating the top-level language family. Data is from Glottolog [Hammarström et al. 2021]. We label a few of the languages on the map as examples.

https://crfm.stanford.edu/assets/report.pdf

Quiz: Are Foundation Models different from Traditional Models?

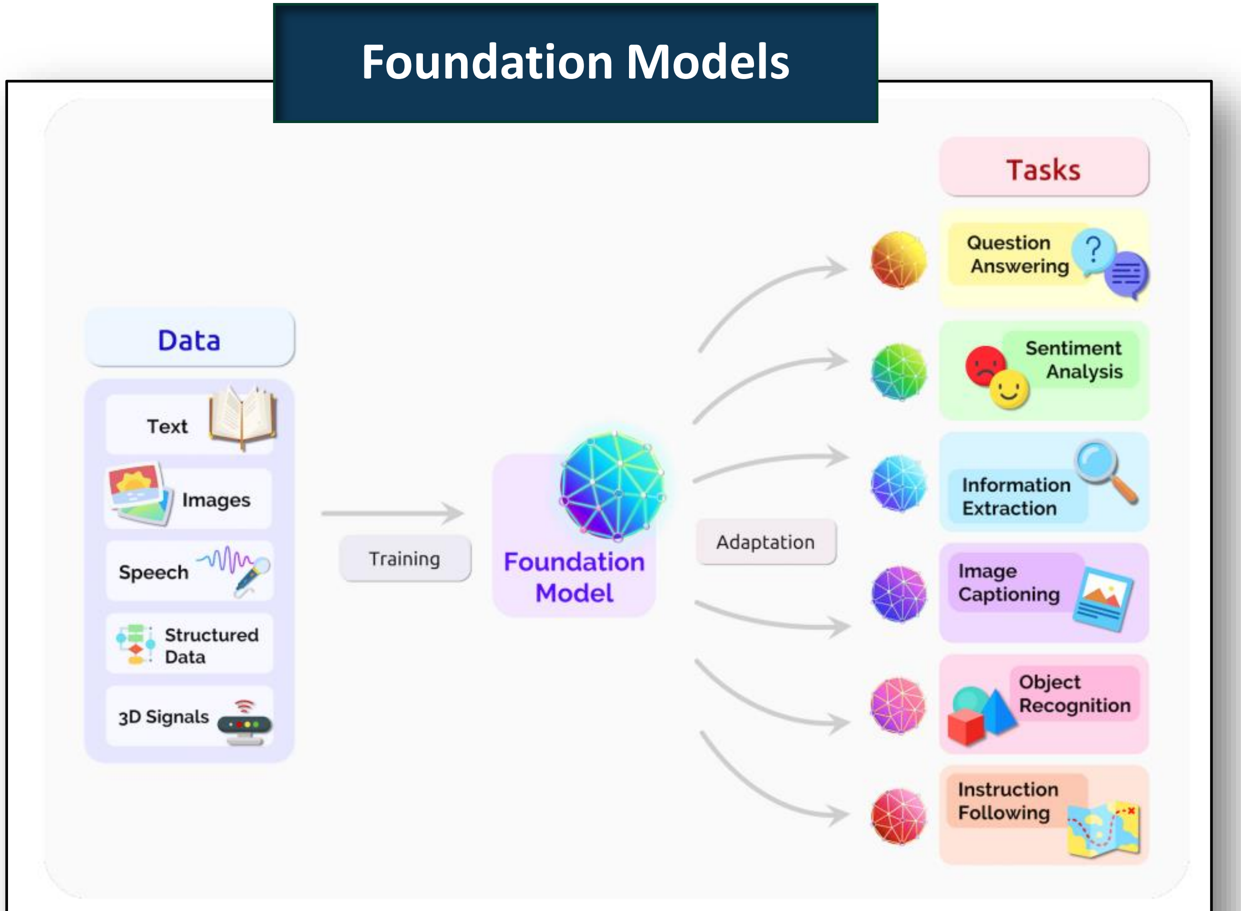
1. Yes
2. No
3. Not sure

Are Foundation Models different from Traditional Models?



Traditional Models

Data		Tasks
Text	ML Model	Q/A
Image	ML Model	Object Detection
Speech	ML Model	Translation
Structured Data	ML Model	Predictions



Foundation Models

Fig. 2. A foundation model can centralize the information from all the data from various modalities. This one model can then be adapted to a wide range of downstream tasks.

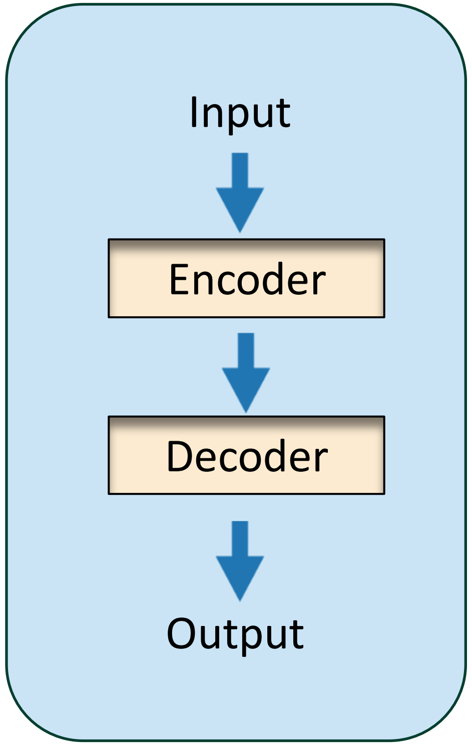
Transformers Architecture (LLM)

Transformers

- Based on the **attention mechanism**
- Foundation of LLMs (Large Language Models)
- Started with use case on language translation

Examples

- Language Translation
- Text summarization
- Named Entity Recognition
- Image Generation
- Etc…



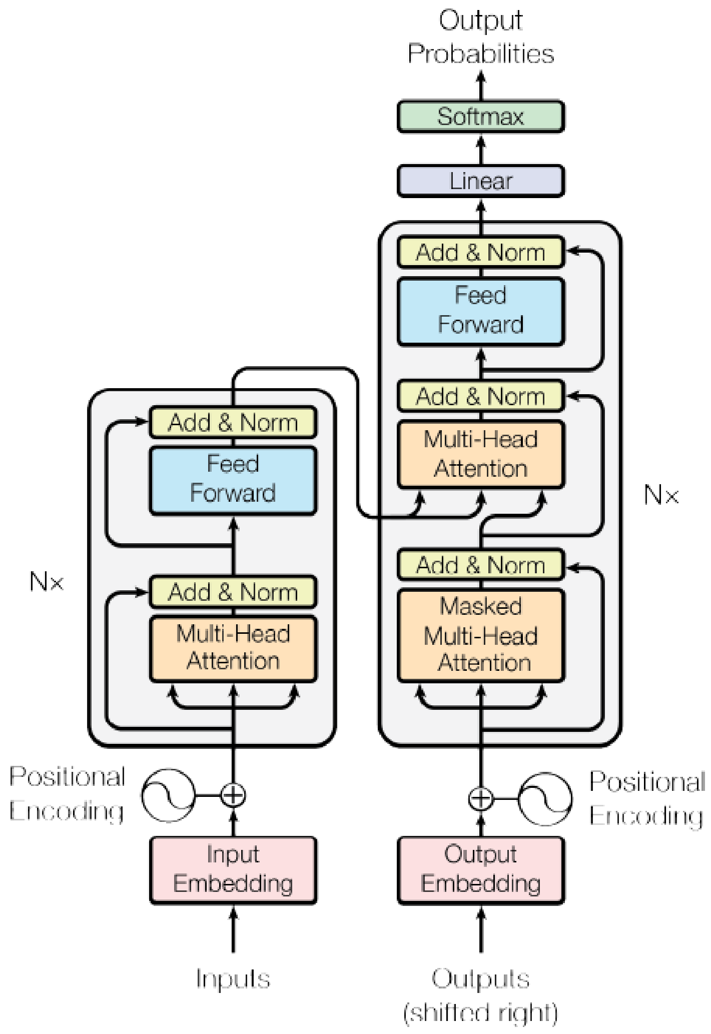
Example translation

"youtube video demo"

"demostración de vídeo de youtube"

Transformers Architecture – Key components

- * Encoders / Decoders
- * Word Embedding
- * Positional Encoding
- * Self Attention, Cross Attention
- * Multi Headed Attention



Source: https://arxiv.org/abs/1706.03762

Quiz

What are the three prompt engineering techniques?

Foundation Models and Large Language Models are same. **True/False**

Generative AI on AWS

Generative AI services on AWS

AWS Bedrock

- Amazon and 3rd party Foundation Models
- Much more

Amazon SageMaker – Fully Managed Service

- Build your models
- Enables end to end automated ML lifecycle
- And beyond that..

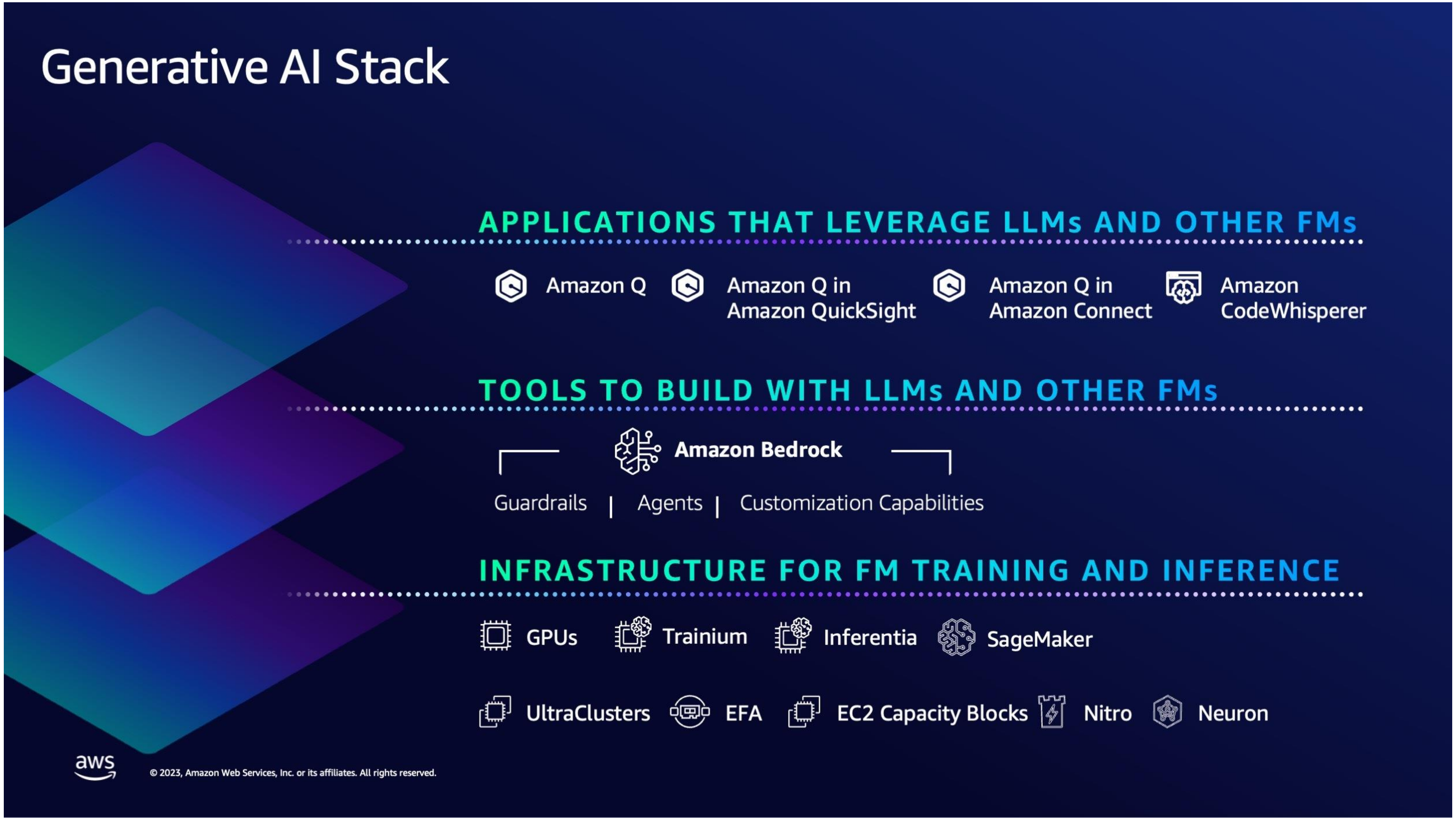
AWS Trainium and Infrentia

- Hardware for accelerated training and inferences

Apps like Amazon Q, CodeWhisperer

- Gen AI powered assistance

Generative AI Stack



Generative AI Stack

APPLICATIONS THAT LEVERAGE LLMs AND OTHER FMs

- Amazon Q
- Amazon Q in Amazon QuickSight
- Amazon Q in Amazon Connect
- Amazon CodeWhisperer

TOOLS TO BUILD WITH LLMs AND OTHER FMs

Amazon Bedrock

Guardrails | Agents | Customization Capabilities

INFRASTRUCTURE FOR FM TRAINING AND INFERENCE

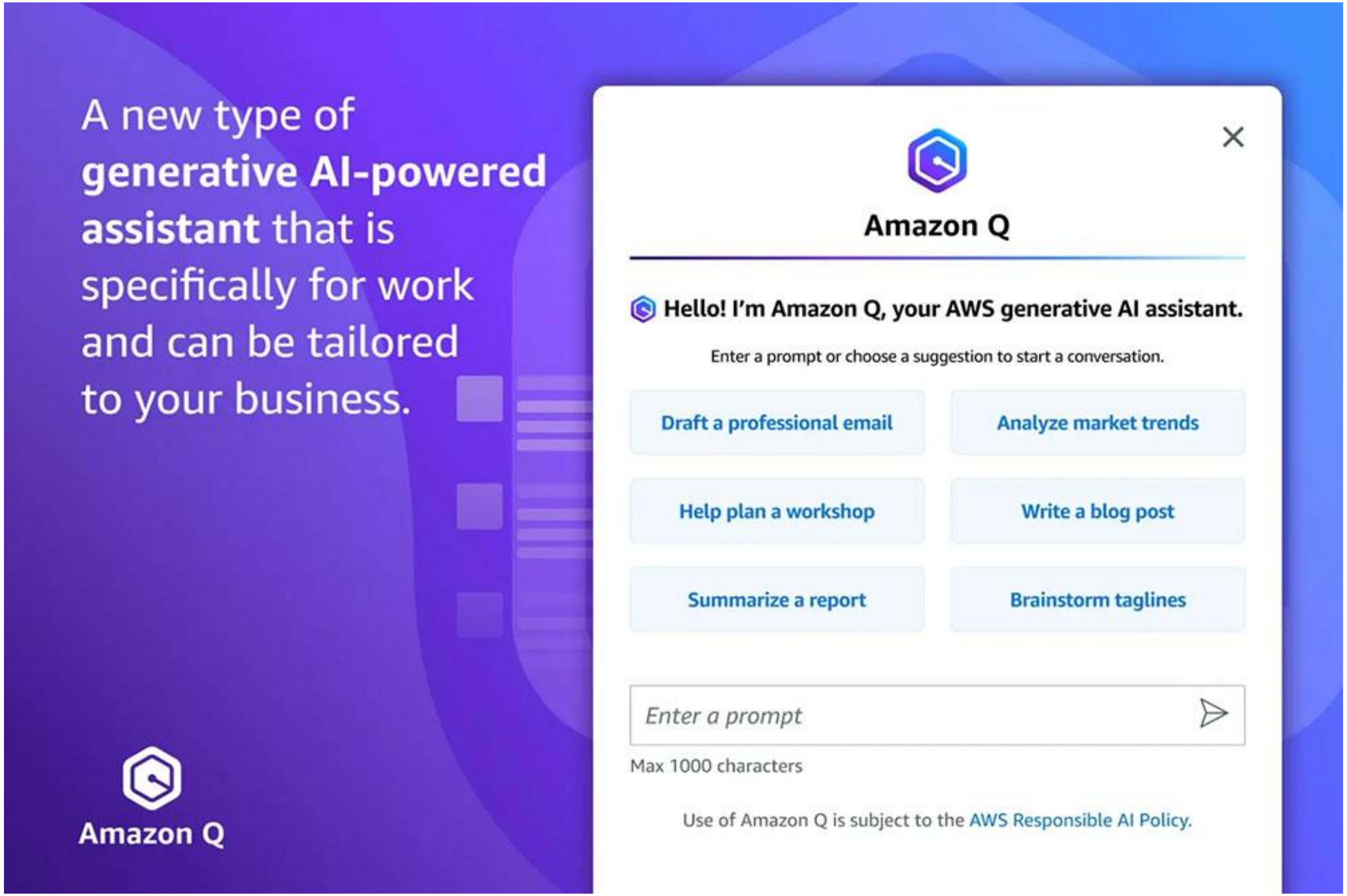
GPUs Trainium Inferentia SageMaker

UltraClusters EFA EC2 Capacity Blocks Nitro Neuron

aws

Source: https://aws.amazon.com/blogs/machine-learning/welcome-to-a-new-era-of-building-in-the-cloud-with-generative-ai-on-aws/

Amazon Q



A new type of **generative AI-powered assistant** that is specifically for work and can be tailored to your business.

Amazon Q

Amazon Q

Hello! I'm Amazon Q, your AWS generative AI assistant.

Enter a prompt or choose a suggestion to start a conversation.

Use of Amazon Q is subject to the AWS Responsible AI Policy.

Source: https://aws.amazon.com/blogs/machine-learning/welcome-to-a-new-era-of-building-in-the-cloud-with-generative-ai-on-aws/

Let’s try this.. An example app

https://aistylist.awsplayer.com/

Another app for Gen AI apps

https://partyrock.aws

- Powered by Amazon Bedrock
- Fun and intuitive hands-on, generative AI app-building playground
- Doesn’t require an AWS account
- Doesn’t need coding

https://partyrock.aws/guide/getStarted

https://aws.amazon.com/about-aws/whats-new/2023/11/partyrock-amazon-bedrock-playground

Quiz

1. Name any two AWS Gen AI services

2. Name any two Foundation Models seen so far

Section Summary

✓ AWS Gen AI Services

✓ Sample apps reflecting Gen AI concepts

Amazon Bedrock

What is Amazon Bedrock?

- fully managed service that provides access to FMs from third-party providers and Amazon

- offers a choice of high-performing foundation models (FMs) from leading AI companies like
  - Amazon Titan
  - AI21 Labs
  - Anthropic
  - Cohere
  - Meta
  - Stability AI
  - Jurrasic

Benefits

- Model Choice
- Customization
- RAG (Retrieval Augmented Generation)
- Agents

Use Cases

- Text Generation
- Virtual Assistants
- Text and Image Search
- Text Summarization
- Image Generation
- Many more…

Amazon Bedrock Workshop

- https://catalog.us-east-1.prod.workshops.aws/workshops/a4bdb007-5600-4368-81c5-ff5b4154f518/en-US

- https://github.com/aws-samples/amazon-bedrock-workshop

Quiz

Name two model variants from Anthropic

Which models are offered by Amazon Titan? Name any three models.

Can we train custom models using Amazon Bedrock? **True/False**

Section Summary

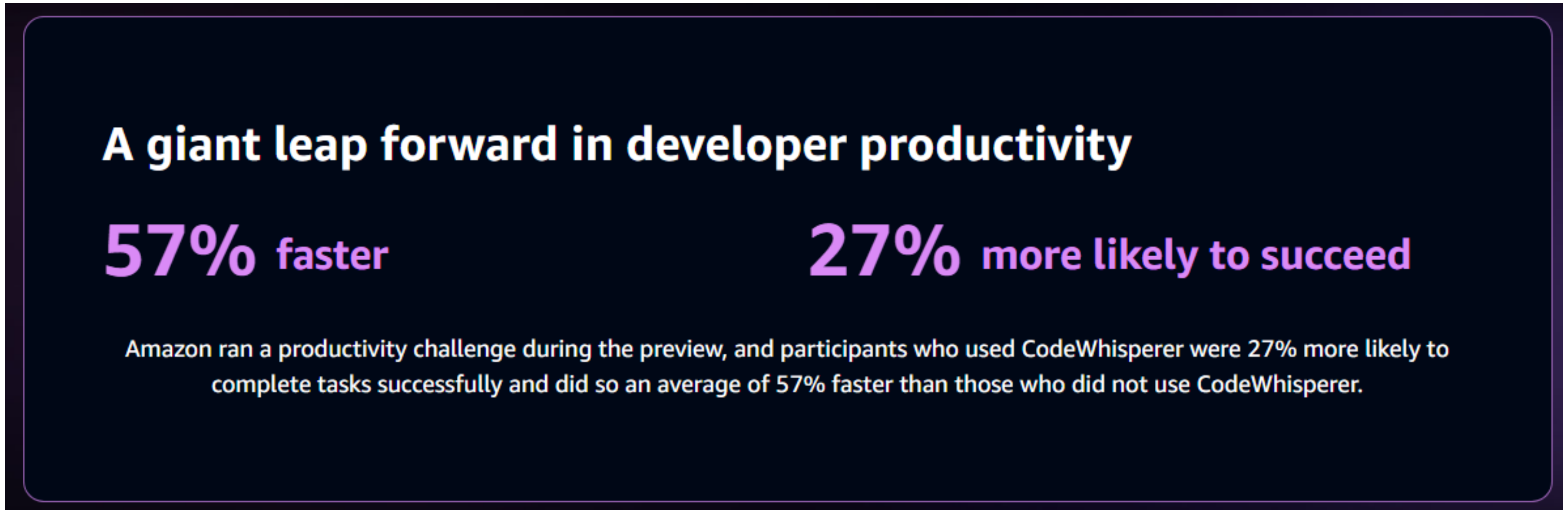
- ✓ Amazon Bedrock
- ✓ Benefits
- ✓ Lab Guides for self practice

Amazon CodeWhisperer

AI Code Generator

CodeWhisperer

1. AI powered productivity tool for IDE and command line
2. Generates real time code suggestions
3. Flag/filter code suggestions
4. Can use Amazon Q (an assistant)
5. Scan your code
6. Can work with multiple programming languages Python, Java, Javascript etc.
7. Can work with multiple IDEs (Integrated Development Envs)
8. Optimized for use with AWS services



A giant leap forward in developer productivity

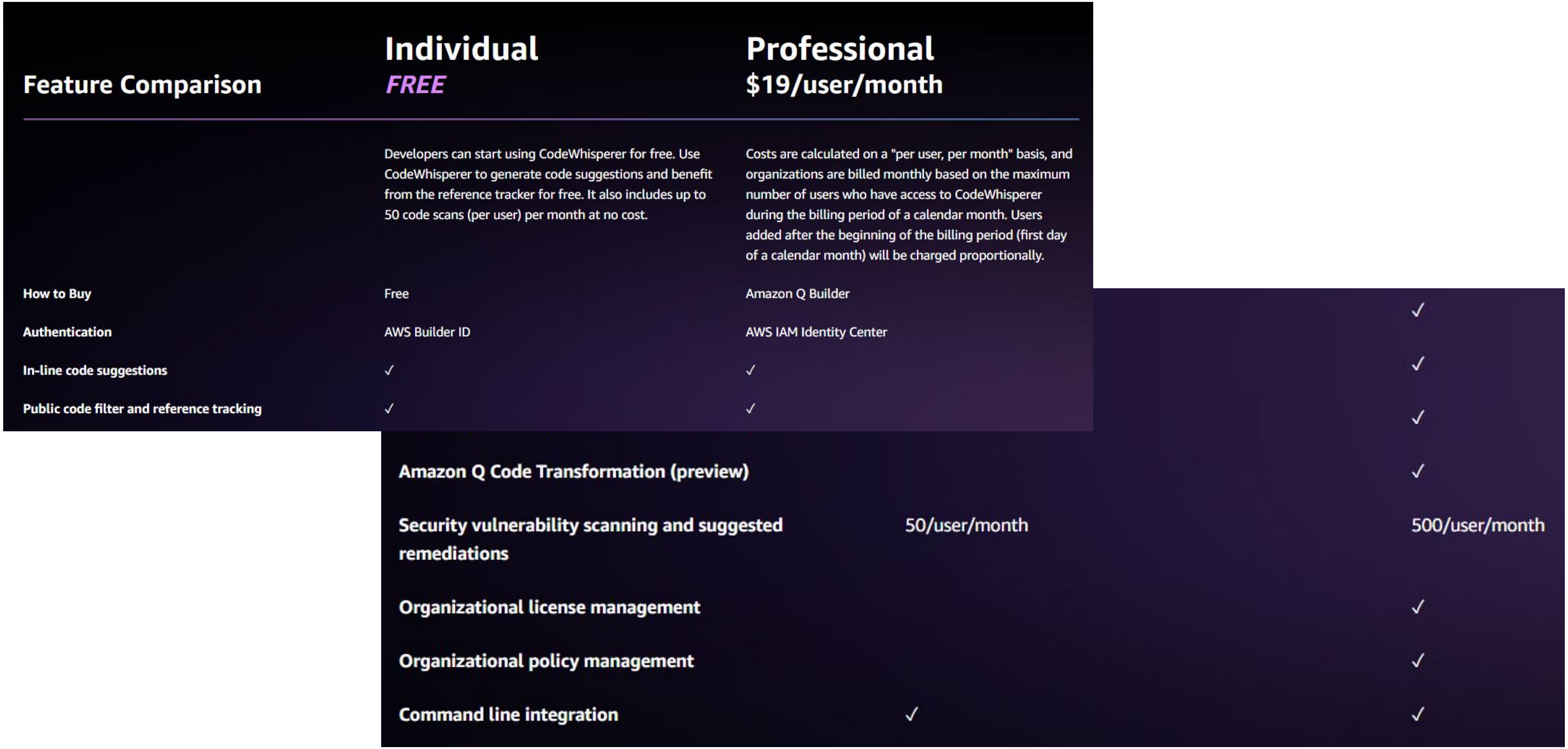
57% faster

27% more likely to succeed

Amazon ran a productivity challenge during the preview, and participants who used CodeWhisperer were 27% more likely to complete tasks successfully and did so an average of 57% faster than those who did not use CodeWhisperer.

Source: https://aws.amazon.com/codewhisperer/

CodeWhisperer Pricing



Feature Comparison	Individual *FREE*	Professional $19/user/month
	Developers can start using CodeWhisperer for free. Use CodeWhisperer to generate code suggestions and benefit from the reference tracker for free. It also includes up to 50 code scans (per user) per month at no cost.	Costs are calculated on a "per user, per month" basis, and organizations are billed monthly based on the maximum number of users who have access to CodeWhisperer during the billing period of a calendar month. Users added after the beginning of the billing period (first day of a calendar month) will be charged proportionally.
How to Buy	Free	Amazon Q Builder
Authentication	AWS Builder ID	AWS IAM Identity Center
In-line code suggestions	✓	✓
Public code filter and reference tracking	✓	✓

		✓
		✓
		✓
Amazon Q Code Transformation (preview)		✓
Security vulnerability scanning and suggested remediations	50/user/month	500/user/month
Organizational license management		✓
Organizational policy management		✓
Command line integration	✓	✓

CodeWhisperer Lab

- Create Builder ID https://profile.aws.amazon.com/
- Workshop URL https://catalog.us-east-1.prod.workshops.aws/workshops/6838a1a5-4516-4153-90ce-ac49ca8e1357/en-US

Quiz

CodeWhisperer is capable of scanning the code **automatically**? **True/ False**

Which of the following are true for Gen AI based code assistants

1. Helps increase in developer productivity by upto 57%
2. Supports C++ language along with other languages like Javascript, Java and Python
3. It is free for individual use with some limitation

Section Summary

- ✓ What is Amazon CodeWhisperer
- ✓ CodeWhisperer Benefits
- ✓ Lab Resources

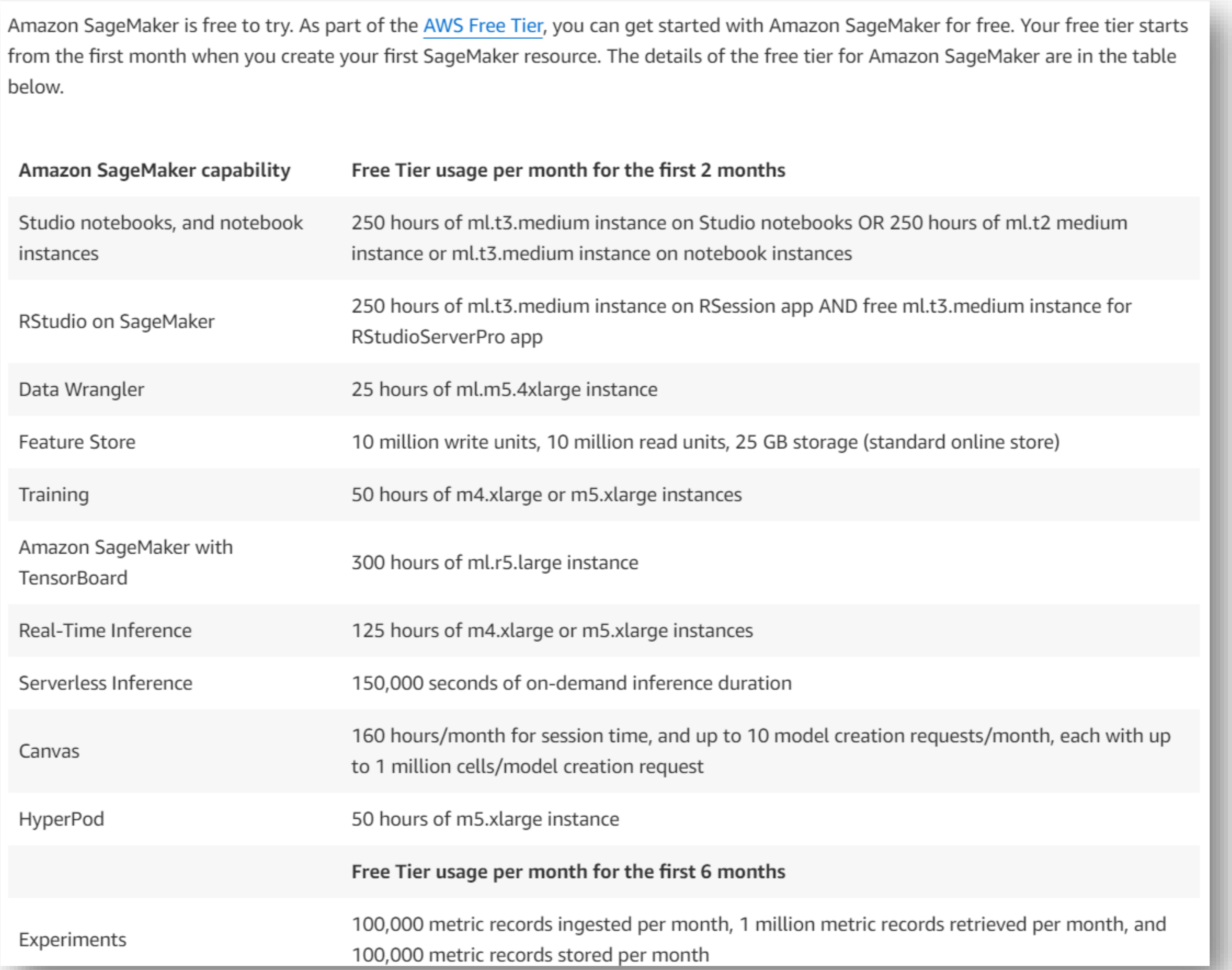
Amazon SageMaker

What is Amazon SageMaker?

- Fully managed service enables high performance, low cost ML
- Data preparation
- Model building
- Model deployment
- Monitoring – Cloudwatch and SageMaker
- MLOps
- supports

jupyter TensorFlow PyTorch mxnet Hugging Face scikit learn python R

SageMaker Pricing



Amazon SageMaker is free to try. As part of the AWS Free Tier, you can get started with Amazon SageMaker for free. Your free tier starts from the first month when you create your first SageMaker resource. The details of the free tier for Amazon SageMaker are in the table below.

Amazon SageMaker capability	Free Tier usage per month for the first 2 months
Studio notebooks, and notebook instances	250 hours of ml.t3.medium instance on Studio notebooks OR 250 hours of ml.t2 medium instance or ml.t3.medium instance on notebook instances
RStudio on SageMaker	250 hours of ml.t3.medium instance on RSession app AND free ml.t3.medium instance for RStudioServerPro app
Data Wrangler	25 hours of ml.m5.4xlarge instance
Feature Store	10 million write units, 10 million read units, 25 GB storage (standard online store)
Training	50 hours of m4.xlarge or m5.xlarge instances
Amazon SageMaker with TensorBoard	300 hours of ml.r5.large instance
Real-Time Inference	125 hours of m4.xlarge or m5.xlarge instances
Serverless Inference	150,000 seconds of on-demand inference duration
Canvas	160 hours/month for session time, and up to 10 model creation requests/month, each with up to 1 million cells/model creation request
HyperPod	50 hours of m5.xlarge instance
	Free Tier usage per month for the first 6 months
Experiments	100,000 metric records ingested per month, 1 million metric records retrieved per month, and 100,000 metric records stored per month

Source: https://aws.amazon.com/sagemaker/pricing/

SageMaker Labs/ workshop

- https://catalog.us-east-1.prod.workshops.aws/workshops/63069e26-921c-4ce1-9cc7-dd882ff62575/en-US

- https://github.com/aws-samples/sagemaker-distributed-training-workshop.git

- https://github.com/aws/amazon-sagemaker-examples.git

Quiz

Which one of the following is the most recent AWS offering in SageMaker?

1. HyperPod
2. Jumpstart
3. Ground Truth

Amazon SageMaker doesn’t offer free tier? **True/ False**

Does Amazon SageMaker provide same functionality as Amazon Bedrock?

Section Summary

- ✓ An overview of Amazon SageMaker
- ✓ Lab Resources

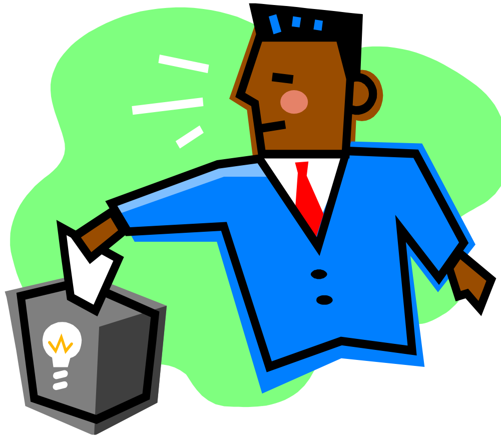
References

- https://aws.amazon.com/generative-ai/
- https://aws.amazon.com/bedrock/
- https://github.com/aws-samples/amazon-bedrock-workshop
- https://workshops.aws/categories/Amazon%20Bedrock
- https://aws.amazon.com/blogs/aws/amazon-codewhisperer-free-for-individual-use-is-now-generally-available/
- https://aws.amazon.com/blogs/machine-learning/how-accenture-is-using-amazon-codewhisperer-to-improve-developer-productivity/
- https://www.amazon.science/blog/the-science-behind-alexas-new-interactive-story-creation-experience
- https://aws.amazon.com/compare/the-difference-between-machine-learning-and-deep-learning/
- https://aws.amazon.com/what-is/foundation-models/
- https://aws.amazon.com/what-is/large-language-model/
- https://aws.amazon.com/generative-ai/use-cases/
- https://aws.amazon.com/solutions/case-studies/dataminr-case-study/
- https://aws.amazon.com/solutions/case-studies/finchcomputing-case-study/
- https://aws.amazon.com/blogs/machine-learning/falcon-180b-foundation-model-from-tii-is-now-available-via-amazon-sagemaker-jumpstart/
- https://aws.amazon.com/what-is/retrieval-augmented-generation/
- https://aws.amazon.com/solutions/case-studies/booking-case-study/
- https://aws.amazon.com/solutions/case-studies/fox-summit-ny-2023-keynote/

LLM- Foundation Models

- High-Resolution Image Synthesis with Latent Diffusion Models https://arxiv.org/abs/2112.10752
- Self-Consistency Improves Chain of Thought Reasoning in Language Models https://arxiv.org/abs/2203.11171
- Tree of Thoughts: Deliberate Problem Solving with Large Language Models https://arxiv.org/abs/2305.10601
- Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks https://arxiv.org/abs/2005.11401
- ART: Automatic multi-step reasoning and tool-use for large language models https://arxiv.org/abs/2303.09014
- ReAct: Synergizing Reasoning and Acting in Language Models https://arxiv.org/abs/2210.03629
- https://docs.anthropic.com/claude/docs/introduction-to-prompt-design
- https://docs.ai21.com/docs/prompt-engineering
- https://aws.amazon.com/what-is/vector-databases/

Thank You



For any query/ feedback, please contact at

https://www.youtube.com/@neerajgarg

https://www.learnwithneeraj.com

https://github.com/neerajg5/gen-ai-aws

https://www.linkedin.com/in/neerajgarg5/