Amazon SageMaker を活用した Generative AI への第一歩と第二歩の Tuner へのガイド

呉 和仁
機械学習ソリューションアーキテクト

自己紹介

呉 和仁 (ごう かずひと)

- 推し事と趣味
  - Amazon SageMaker や AWS の AI サービスの推し（押し）売り
  - 機械学習を使ったお遊び
    - Anazon Polly に歌わせて VTuber デビューさせてみた
    - たけのこの里が好きな G くんのために、きのこの山を分別する装置を作ってあげた。
      モデル作成編 分別装置作成編
    - 献立を考えるのが大変なら AI に任せればいいじゃない～Amazon Personalize で献立をレコメンドしてみた～
    - 算数ドリルを秒で解くプログラムに Amazon Textract を使って挑戦してみる
    - テキスト解析 AI サービス Amazon Comprehend で本を読まずに読書感想文に挑戦してみる
    - 生成系 AI を使って仕事をやっているアピールをしてみた ← Today！
    - 生成系 AI と連歌を雅に嗜んでみた ← Today！

本日のおしながき

1. 【デモ】生成系 AI を使ってみる
    1. Text to Image で仕事をやっているアピールしようとする G くん
    2. Text to Text で連歌を雅に嗜もうとする G くん

2. 【一歩目】Amazon SageMaker での生成系 AI の利用を開始する

3. 【デモ】生成系 AI を Fine Tune する
    1. Text to Image で仕事をやっているアピールする G くん
    2. Text to Text で連歌を雅に嗜む G くん

4. 【二歩目 】Amazon SageMaker で生成系 AI を Fine Tune する

※Fine Tune の詳細は後ほど

Demo 1-1.
Text to Image で仕事をやっているアピールしようとする G くん

Amazon SageMaker

Getting started
Studio
Studio Lab
Canvas
RStudio
TensorBoard

Domains

SageMaker dashboard
Images
Lifecycle configurations
Search

JumpStart
Foundation models NEW
Computer vision models
Natural language processing models

Governance
Ground Truth
Notebook
Processing
Training

MACHINE LEARNING

Amazon SageMaker

Build, train, and deploy machine learning models at scale

The quickest and easiest way to get ML models from idea to production.

New to SageMaker?

Get started with Amazon SageMaker by completing the quick start guide.

Get Started

Documentation

Getting started
Tutorials
Documentation
Developer Resources
AWS Developer Forum
Contact us

How it works

What is Amazon SageMaker?

Amazon SageMaker provides machine learning (ML) capabilities for data scientists and developers to prepare, build, train, and deploy high-quality ML models efficiently.

New user onboarding guide NEW

Get started with Amazon SageMaker by completing the quick start onboarding guide.

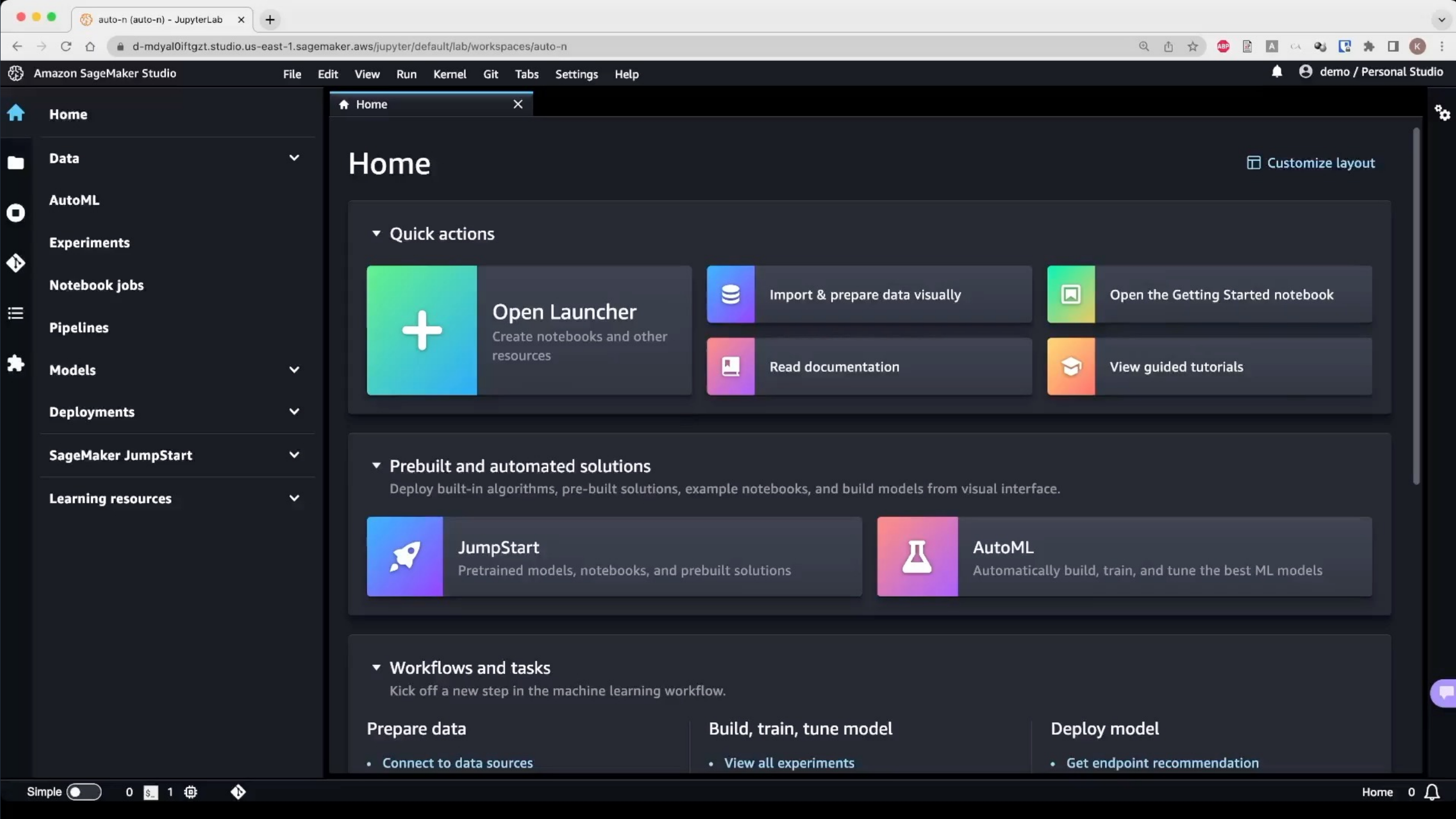
Get started with SageMaker

Typical SageMaker workflow

1. Label data

Set up and manage labeling jobs for highly accurate training datasets within Amazon SageMaker, using active learning and human labeling.

Demo 1-2. Text to Text で連歌を雅に嗜もうとする G くん


auto-n (auto-n) - JupyterLab
d-mdyal0iftgzt.studio.us-east-1.sagemaker.aws/jupyter/default/lab/workspaces/auto-n
Amazon SageMaker Studio
File
Edit
View
Run
Kernel
Git
Tabs
Settings
Help
demo / Personal Studio
Home
Data
AutoML
Experiments
Notebook jobs
Pipelines
Models
Deployments
SageMaker JumpStart
Learning resources
Home
Customize layout
Quick actions
Open Launcher
Create notebooks and other resources
Import & prepare data visually
Open the Getting Started notebook
Read documentation
View guided tutorials
Prebuilt and automated solutions
Deploy built-in algorithms, pre-built solutions, example notebooks, and build models from visual interface.
JumpStart
Pretrained models, notebooks, and prebuilt solutions
AutoML
Automatically build, train, and tune the best ML models
Workflows and tasks
Kick off a new step in the machine learning workflow.
Prepare data
Connect to data sources
Build, train, tune model
View all experiments
Deploy model
Get endpoint recommendation
Simple
0
1
Home
0

AWS のミッション：すべてのお客様に機械学習を

AI SERVICES

Generative AI

SCALING GENERATIVE AI
- Amazon Bedrock (Limited Preview)

FOUNDATION MODELS (Limited Preview)
- Amazon Titan
- Claude
- Stable Diffusion
- Jurassic

Specialized AI Services

BUSINESS PROCESSES
- Amazon Personalize
- Amazon Forecast
- Amazon Fraud Detector
- Amazon Lookout for Metrics

SEARCH
- Amazon Kendra

CONVERSATION
- Amazon Lex
- Contact Lens
- Voice ID
- Amazon Transcribe Call Analytics

CODE + DEVOPS
- Amazon CodeGuru
- Amazon CodeWhisperer
- Amazon DevOps Guru

INDUSTRIAL
- Amazon Monitron
- Amazon Lookout for Equipment
- Amazon Lookout for Vision

HEALTH
- Amazon HealthLake
- Amazon Comprehend Medical
- Amazon Transcribe Medical
- Amazon Omics

Core

TEXT
- Amazon Translate
- Amazon Comprehend

SPEECH
- Amazon Polly
- Amazon Transcribe

VISION
- Amazon Textract
- Panorama
- Amazon Rekognition
- AWS

AMAZON SAGEMAKER

JUMPSTART
Use pre-built models in SageMaker

CANVAS
No-code ML for business analysts

STUDIO LAB
Learn ML

GROUND TRUTH
Label data

STUDIO IDE
- Prepare data
- Store features
- Geospatial ML
- Build with notebooks
- Train models
- Tune parameters
- Deploy in production
- Manage and monitor

CI/CD | GOVERNANCE | RESPONSIBLE ML

ML FRAMEWORKS & INFRASTRUCTURE

PyTorch, TensorFlow | Amazon EC2 | CPUs | GPUs | AWS Inferentia | AWS Trainium | Habana Gaudi | FPGA

AWS のミッション：すべてのお客様に機械学習を

AI SERVICES

Generative AI

SCALING GENERATIVE AI
Amazon Bedrock (Limited Preview)

FOUNDATION MODELS (Limited Preview)
Amazon Titan | Claude | Stable Diffusion | Jurassic

Specialized AI Services

BUSINESS PROCESSES
Amazon Personalize
Amazon Forecast
Amazon Fraud Detector
Amazon Lookout for Metrics

SEARCH
Amazon Kendra

CONVERSATION
Amazon Lex
Contact Lens
Voice ID
Amazon Transcribe Call Analytics

CODE + DEVOPS
Amazon CodeGuru
Amazon CodeWhisperer
Amazon DevOps Guru

INDUSTRIAL
Amazon Monitron
Amazon Lookout for Equipment
Amazon Lookout for Vision

HEALTH
Amazon HealthLake
Amazon Comprehend Medical
Amazon Transcribe Medical
Amazon Omics

Core

TEXT
Amazon Translate | Amazon Comprehend

SPEECH
Amazon Polly | Amazon Transcribe

VISION
Amazon Textract | Amazon Rekognition | AWS Panorama

AMAZON SAGEMAKER

JUMPSTART
Use pre-built models in SageMaker

CANVAS
No-code ML for business analysts

STUDIO LAB
Learn ML

GROUND TRUTH
Label data

STUDIO IDE

Prepare data | Geospatial ML | Build with notebooks | Train models | Deploy in production
Store features | Tune parameters | Manage and monitor

CI/CD | GOVERNANCE | RESPONSIBLE ML

ML FRAMEWORKS & INFRASTRUCTURE

PyTorch, TensorFlow | Amazon EC2 | CPUs | GPUs | AWS Inferentia | AWS Trainium | Habana Gaudi | FPGA

機械学習の開始には数々の苦労が伴う

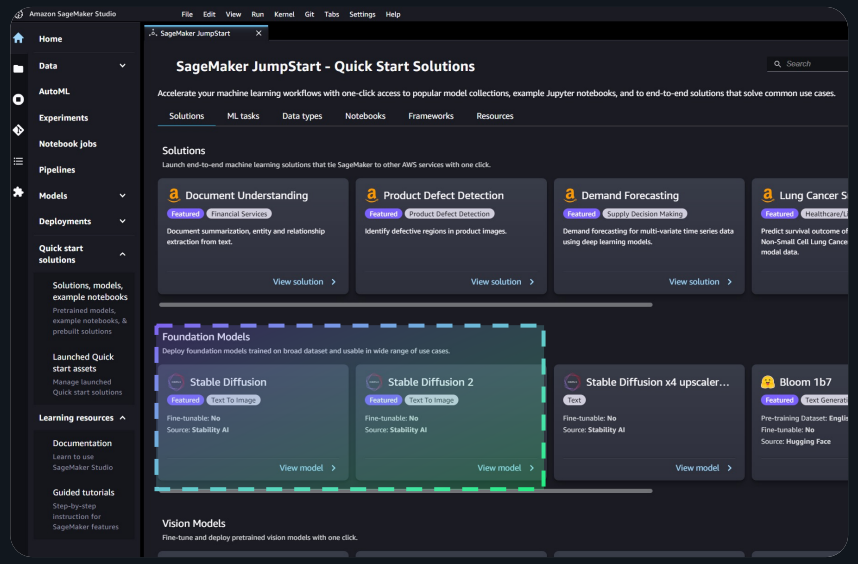
機械学習を開始するまでの道のりは

長すぎる…

- 公開されたアルゴリズムやモデルを動かすまでの苦労
- スクリプトを保守し、更新し続ける苦労
- コンピューティングリソースを準備する苦労
- NLP や Vision モデルのスクラッチ開発
- そもそもの機械学習の知識獲得

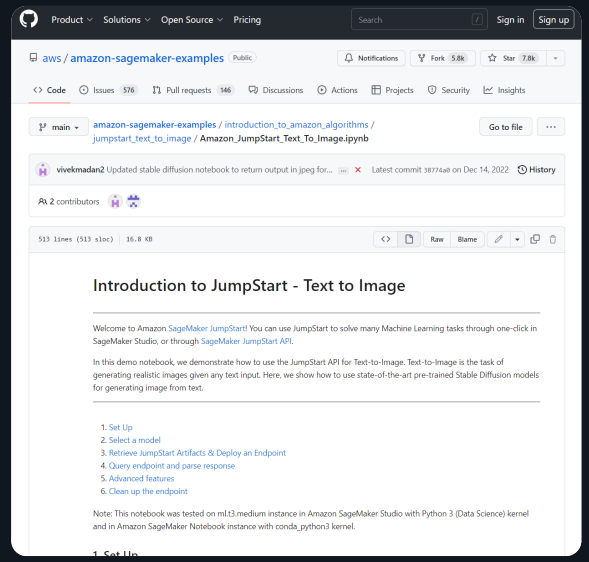
3 種類の方法で生成系 AI を SageMaker で開始

SageMaker Studio (JumpStart)



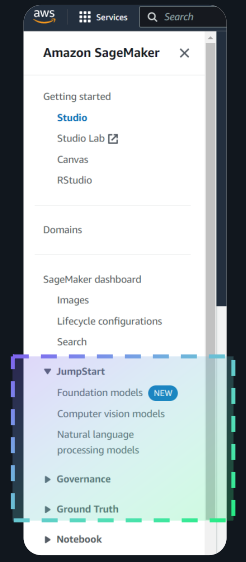
マウスポチポチでモデルを
ホストして生成開始

Notebooks



Shift + Enter 連打で
ホストして生成開始
(コードがある場合)

AWS console (JumpStart)



モデルをホストせずに
console から生成開始

Amazon SageMaker JumpStart

ML Hub から基盤モデルや
ビルトインアルゴリズムを
数クリックでデプロイ

Machine learning hub

400 以上の事前学習済の基盤モデルやソリューション、
サンプルノートブックを閲覧

モデルもトレーニングコードも推論コードも不要

自前でコードを用意せずに推論したり、
独自のデータセットで Fine Tune 可能

UI も API も

マウス操作によるモデルのデプロイだけでなく、
同様のことをする API の準備もあり、試用からシステム統合まで簡単

セキュリティ

VPC エンドポイントで閉域アクセスも可能

モデルやデータのコラボレーション

デプロイしたモデルや、ノートブックは認証認可の元
誰でもアクセス可能

Foundation modelの幅広い選択肢

AMAZON SAGEMAKER JUMPSTARTから提供中

Meta AI

Models

Llama 2 7B, 13B, 70B

Tasks

Question answering
Chat
Summarization
Paraphrasing
Sentiment analysis
Text generation

AI21 labs

Models

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

Tasks

Text generation
Long-form generation
Summarization
Paraphrasing
Chat
Information extraction
Question answering
Classification

co:here

Models

Cohere Command XL

Tasks

Text generation
Information extraction
Question answering
Summarization

Models

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

Tasks

Machine translation
Question answering
Summarization
Annotation
Data generation

Features

Fine-tuning on FLAN T5 models, GPT-6B, Falcon-7B

stability.ai

Models

Stable Diffusion XL
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

alexa

Models

AlexaTM 20B

Tasks

Machine translation
Question answering
Summarization
Annotation
Data generation

aws

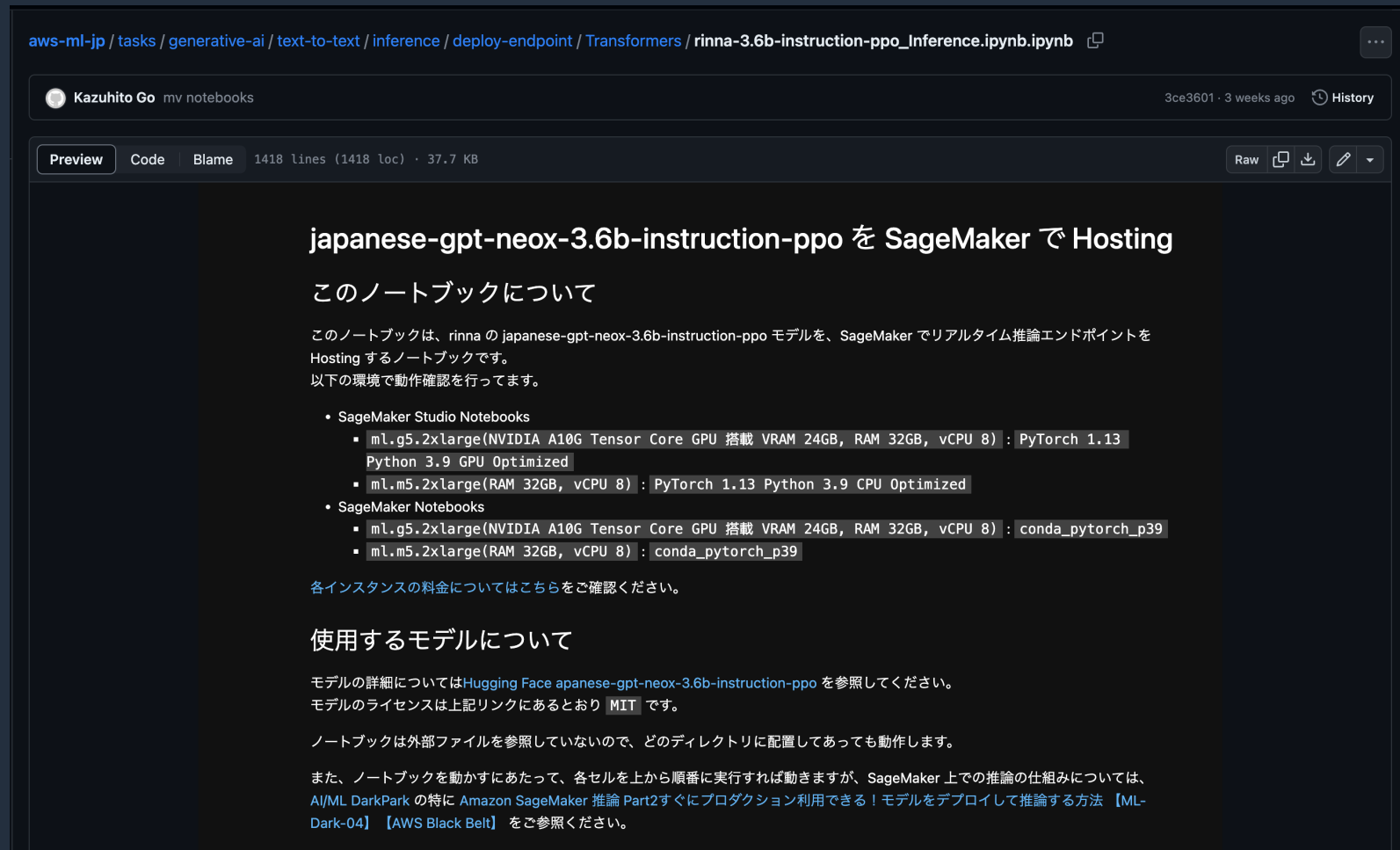
SageMaker JumpStart できなくても Jump Start！

SageMaker JumpStart が肩代わりしていたことを SageMaker の API で再現

1. モデルと推論コードを用意して tar.gz に
2. 実行環境を定義
3. デプロイ

ノートブックは
GitHubで公開中

https://github.com/aws-samples/aws-ml-jp



公開されているモデルでは要件を満たさないことも

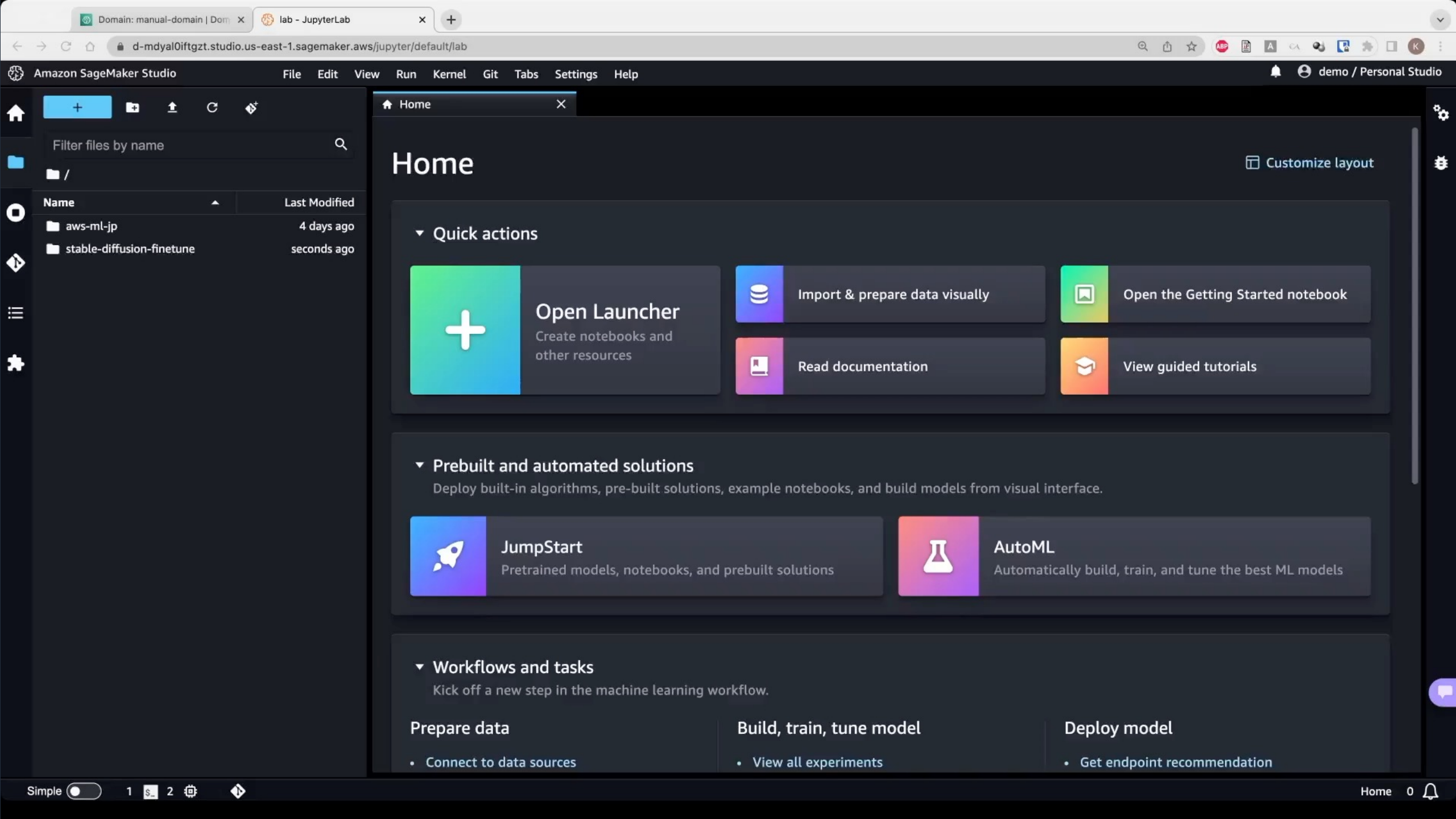
- 素の Stable Diffusion では G くんの画像を生成できない
- 素の japanese-gpt-neo-x-3.6b-ppo(rinna) では下の句を読めない
- 学習していないタスク
- 知らないドメイン知識
- 多言語対応
- 画風や口調

etc...

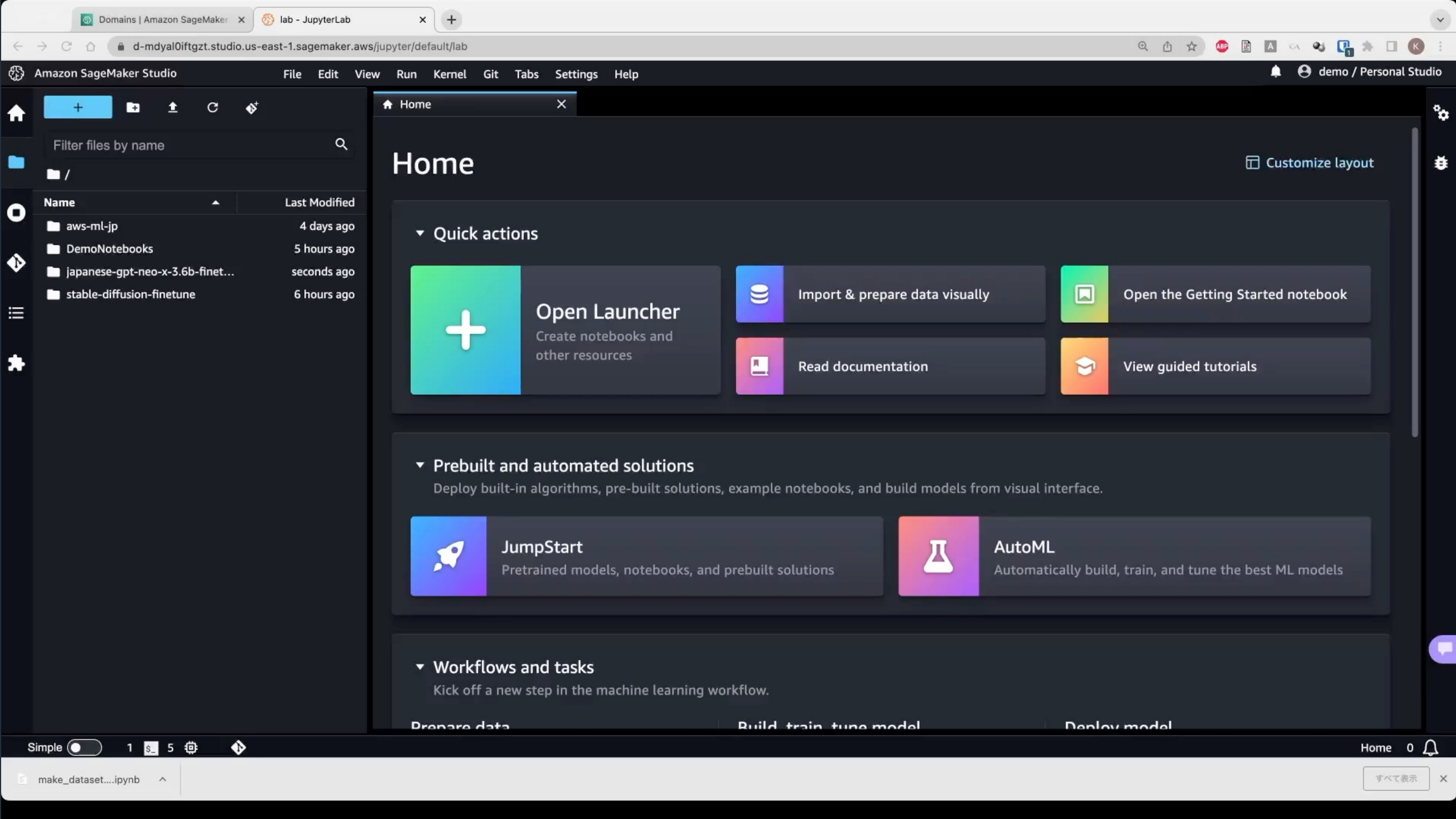
Fine Tune !

※Fine Tune とは学習済モデルを少量のデータで再学習し、モデルを微調整すること

Demo2-1.
Text to Image で仕事をやっているアピールする G くん


Domain: manual-domain | Dom
lab - JupyterLab
d-mdyal0iftgzt.studio.us-east-1.sagemaker.aws/jupyter/default/lab
Amazon SageMaker Studio
File Edit View Run Kernel Git Tabs Settings Help
demo / Personal Studio
Filter files by name
/
Name
Last Modified
aws-ml-jp
4 days ago
stable-diffusion-finetune
seconds ago
Home
Home
Customize layout
Quick actions
Open Launcher
Create notebooks and other resources
Import & prepare data visually
Open the Getting Started notebook
Read documentation
View guided tutorials
Prebuilt and automated solutions
Deploy built-in algorithms, pre-built solutions, example notebooks, and build models from visual interface.
JumpStart
Pretrained models, notebooks, and prebuilt solutions
AutoML
Automatically build, train, and tune the best ML models
Workflows and tasks
Kick off a new step in the machine learning workflow.
Prepare data
Connect to data sources
Build, train, tune model
View all experiments
Deploy model
Get endpoint recommendation
Simple
1
2
Home
0

Demo2-2：Text to Text で連歌を雅に嗜む G くん


Domains | Amazon SageMaker
lab - JupyterLab
d-mdyal0iftgzt.studio.us-east-1.sagemaker.aws/jupyter/default/lab
Amazon SageMaker Studio
File Edit View Run Kernel Git Tabs Settings Help
demo / Personal Studio
Filter files by name
/
Name
Last Modified
aws-ml-jp
4 days ago
DemoNotebooks
5 hours ago
japanese-gpt-neo-x-3.6b-finet...
seconds ago
stable-diffusion-finetune
6 hours ago
Home
Home
Customize layout
Quick actions
Open Launcher
Create notebooks and other resources
Import & prepare data visually
Open the Getting Started notebook
Read documentation
View guided tutorials
Prebuilt and automated solutions
Deploy built-in algorithms, pre-built solutions, example notebooks, and build models from visual interface.
JumpStart
Pretrained models, notebooks, and prebuilt solutions
AutoML
Automatically build, train, and tune the best ML models
Workflows and tasks
Kick off a new step in the machine learning workflow.
Simple
1
5
Home
0
make_dataset....ipynb

SageMaker で Fine Tune すれば要求を満たせる

Fine Tune：モデルを小数のデータセットで再学習してモデルの目的を達成させる

Stable Diffusion (Text to Image) の例

1. プロンプトと画像を複数用意

- **instance_prompt:**"呉和仁という人間の写真"
- **class_prompt:**"人間の写真"

2. S3にアップロード

3. SageMaker JumpStart で Stable Diffusion を選択して S3 の URI をして [Train] < ｶﾁｯ

japanese-gpt-neo-x-3.6b (Text to Text) の例

1. 質問とコンテキストと回答を多数用意

- **instruction:**"input は短歌の上の句です。下の句を詠んでください。"
- **input:**"データあるうまい言葉に騙される"
- **output:**"実態いつもガベージコレクション"
- **category:**"open_qa"

2. S3 にアップロード

3. ノートブックのファイルパスを書き換えて Shift + Enter を連打

SageMaker で Fine Tune すれば要求を満たせる

Fine Tune：モデルを小数のデータセットで再学習してモデルの目的を達成させる

Stable Diffusion (Text to Image) の例

Before

G くんとは程遠い
顔を生成

After

G くんに似た顔を生成

japanese-gpt-neo-x-3.6b (Text to Text) の例

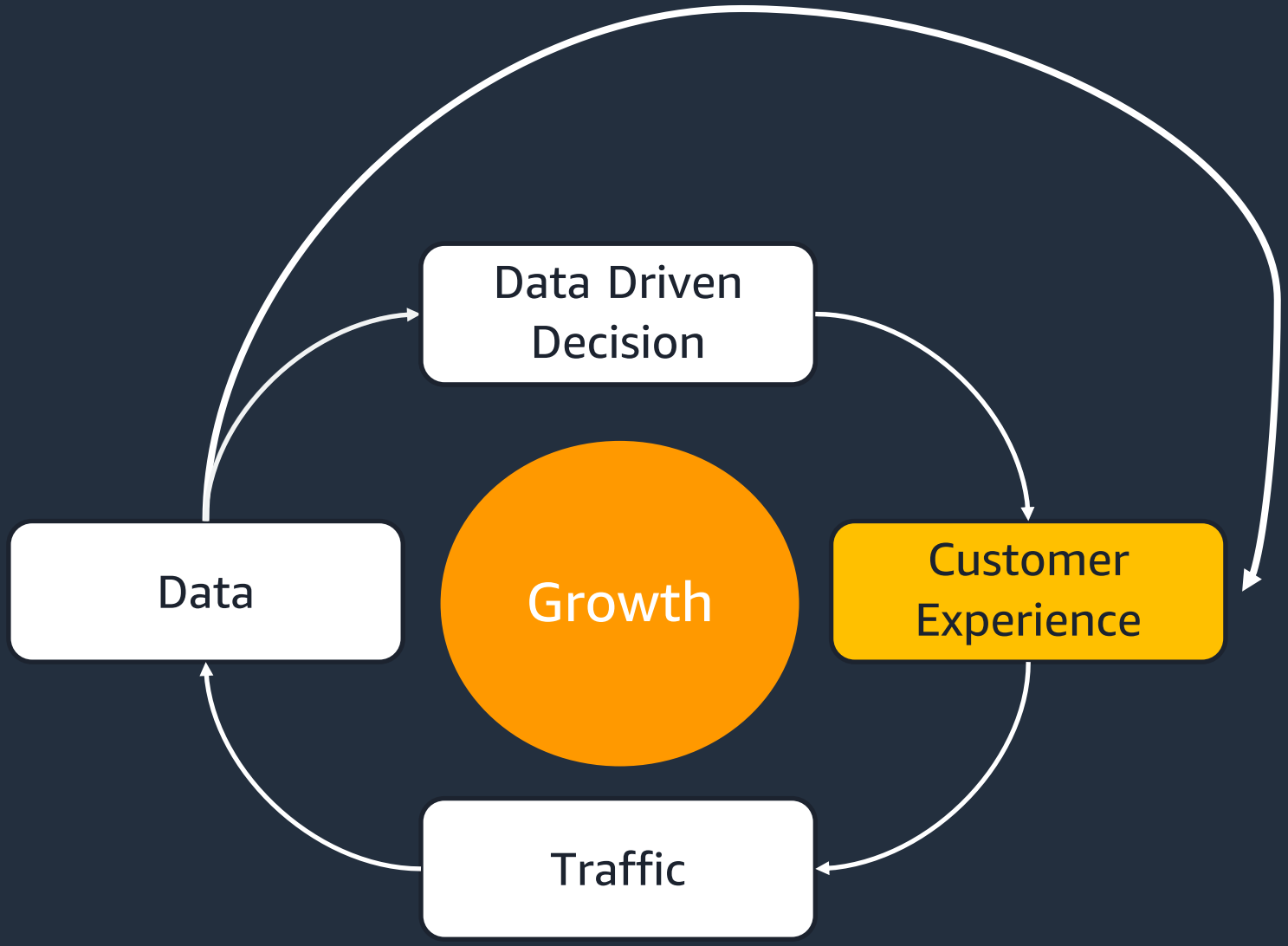
Before

この世をば わが世と思ふ 望月の
↓
望月の 満ち欠けの 影を 見つつ 我が身は 有りけり
5・7・5・7・7 が全く守られていない

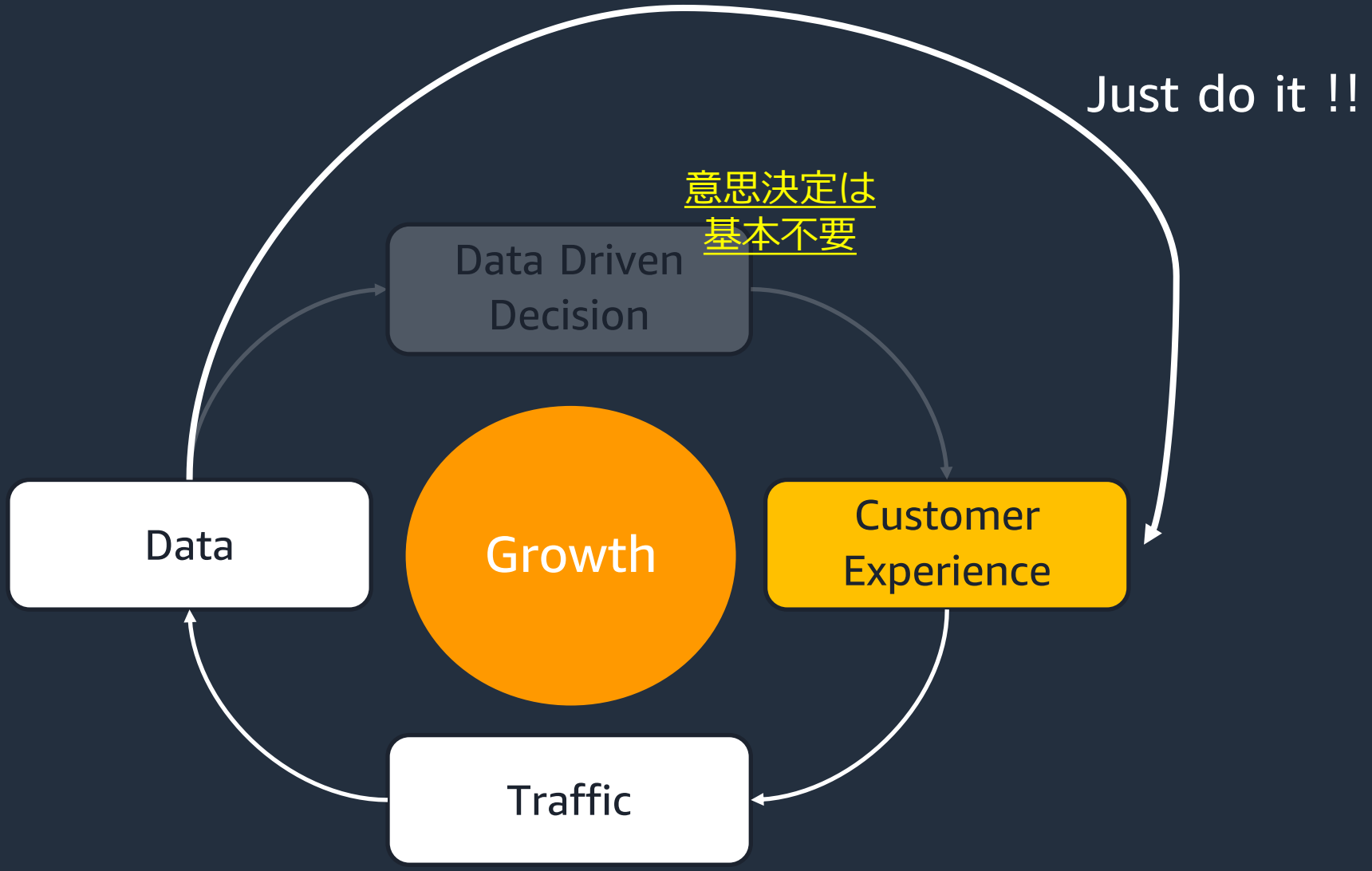
After

ダイエット 腹減る夜中 耐え難き
↓
こよひは月も みでねさめぬ
5・7・5・7・7 がほぼ守られている

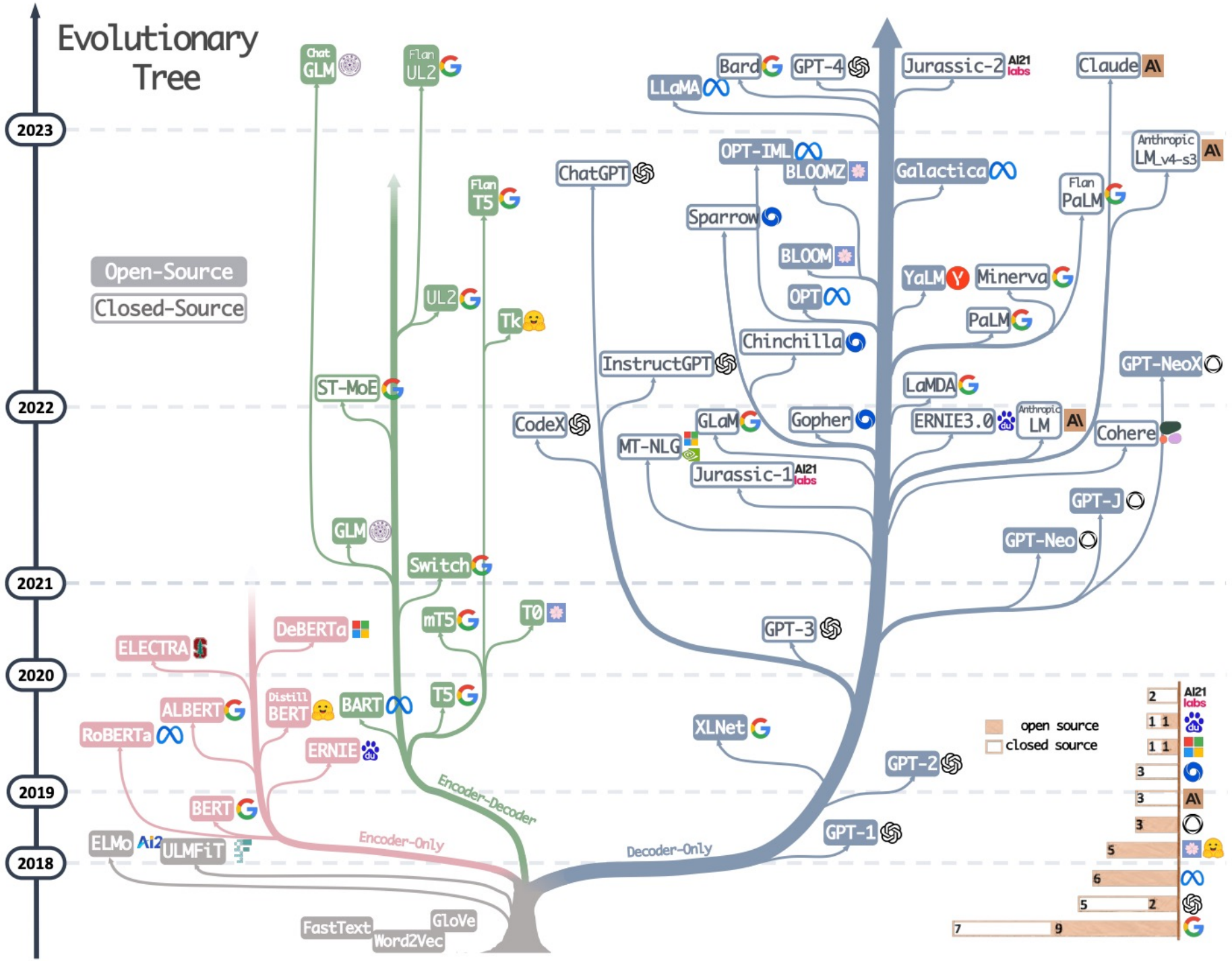
差別化要素はデータへ



差別化要素はデータへ



モデルは
コモディティ化
して選べる立場に



Source:
https://github.com/Mooler0410/LLMsPracticalGuide

aws

2023/5~7 のモデルの公開

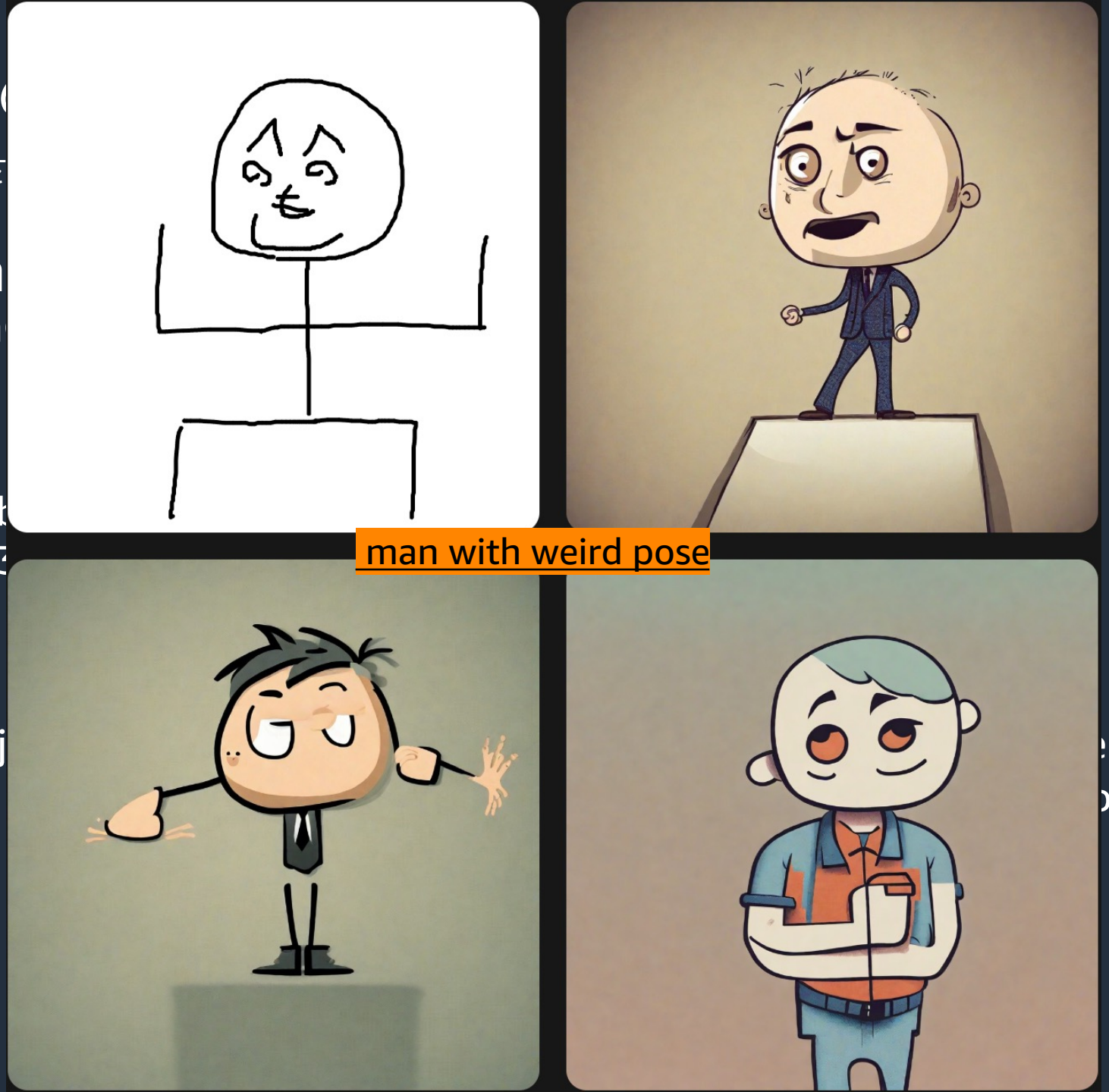
朝目覚めたら公開されていたが日常茶飯事

- CyberAgent
  - 5/18 OpenCALM
- MosaicML
  - 5/5 mpt-7b
  - 6/22 mpt-30b
- rinna
  - 5/17, 6/9 japanese-gpt-neo-x-3.6b
- TII
  - 6/5 falcon-7b
  - 6/5 falcon-40b
- Anthropic
  - 7/11 Claude 2
- Stability AI
  - 7/14 Stable Doodle
  - 7/27 Stable Diffusion XL 1.0
- Meta
  - 7/19 Llama 2

2023/5~7 (

朝目覚めたら公開さ

- CyberAgen
  - 5/18 Open
- MosaicML
  - 5/5 mpt-7b
  - 6/22 mpt-3
- rinna
  - 5/17, 6/9 j



on XL 1.0

2023/5/7 のモデルの公開

朝目覚めたら公開されていたが日常茶飯事

- CyberAg
  - 5/18 O
- MosaicM
  - 5/5 mp
  - 6/22 m
- rinna
  - 5/17, 6

[9]:
```
for dialog in dialogs:
    payload = {
        "inputs": [dialog],
        "parameters": {"max_new_tokens": 256, "top_p": 0.9, "temperature"
    }
    result = query_endpoint(payload)[0]
    for msg in dialog:
        print(f"{msg['role'].capitalize()}: {msg['content']}\n")
    print(f"> {result['generation']['role'].capitalize()}: {result['gener
    print("\n==================================\n")
```

```
System: 上の句を与えるので下の句を詠んでください

User: おもしろきこともなき世をおもしろく

> Assistant:  Sure! Here's a possible completion of the haiku:

笑顔で歩くやすらぎの道

==================================
```

Amazon Bedrock が出てもデータは必要

- Amazon Bedrock は SageMaker が行っていたモデルのデプロイが不要になったサーバーレスサービス
- API にプロンプトを投げるだけ
- インスタンスやモデルのメンテナンスが不要
- Fine Tune が可能

集めたデータは無駄にならないので
すぐに集め始めましょう！

第三歩目は？

- モデルをゼロから作る選択肢も
- モデルを公開して収益化も可能
- 数億オーダーの予算確保から

Provider

お金（数億円オーダー）

HW/ミドルウェア/
インフラの知識

ML の知識

データ

Tuner

データ

Consumer

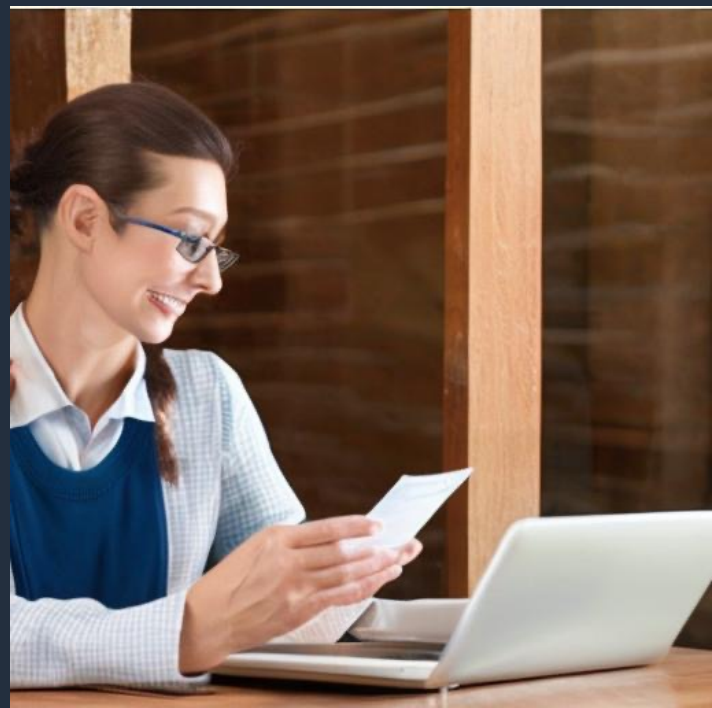
やる気

やる気

やる気

まとめ

- Amazon SageMaker JumpStart を使えば
  モデルのデプロイも Fine Tune もすぐに可能
- Amazon SageMaker JumpStart にないモデルも公開ノートブックを使えば、
  モデルのデプロイも Fine Tune もすぐに可能
- まずは、Consumer としての第一歩を今日すぐにでも踏み出し、
  第二歩のTunerを目指す
- データは今から集める仕組みを検討しましょう



Stable Diffusion で生成した
Consideration of data collection mechanisms
(データ収集メカニズムの検討)

Next Action / Materials

デモで紹介したノートブック他
https://github.com/aws-samples/aws-ml-jp/

- tasks/generative-ai/text-to-text/inference/deploy-endpoint/Transformers/rinna-3.6b-instruction-ppo_Inference.ipynb.ipynb
- tasks/generative-ai/text-to-text/fine-tuning/instruction-tuning/Transformers/Rinna_Neox_LoRA_ja.ipynb
- tasks/generative-ai/text-to-image/inference/stable-diffusion-webui

本日の Text to Text のブログ (8/2 公開予定なのでまだ開けません)
https://aws.amazon.com/jp/builders-flash/202308/generative-ai-renga/

Black Belt AI ML Dark Part (SageMaker の解説)
https://www.youtube.com/playlist?list=PLAOq15s3RbuL32mYUphPDoeWKUiEUhcug

Black Belt AI ML Light Part(機械学習モデルをプロダクトで活用するためのプロセスを解説)
https://www.youtube.com/playlist?list=PLAOq15s3RbuJ81DBtH66tQL2_9H519ODQ

Thank you!