GAI214

De MLOps à FMOps : Comment passer vos solutions d'IA générative à l'échelle

Ségolène Dessertine-Panhard (elle/elle)
Senior Manager au Generative AI Innovation Center
Amazon Web Services

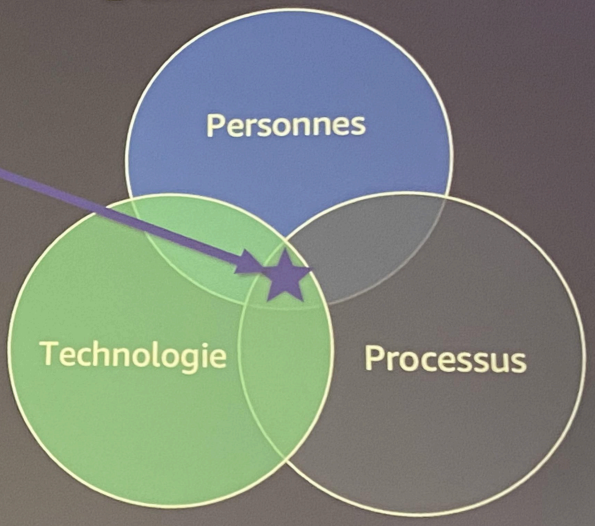
Patrick Sard (il/lui)
Architecte de solutions senior
Amazon Web Services

Qu'est-ce que le MLOps ?

ML + Ops
Apprentissage automatique
+ opérations

La combinaison de personnes, de processus et de technologies afin déployer et mettre en production des solutions de machine learning de façon efficace et fiable.

Définition du MLOps



Bénéfices attendus du MLOps

STANDARDISATION DES OPÉRATIONS ET DE L'INFRASTRUCTURE DE VOS SOLUTIONS D'IA

	Objectif	Métrique technique	Avant MLOPS	Résultats attendus avec MLOps	Valeur ajoutée
1	Efficacité d'exécution	Délai de rentabilisation (de l'idée à la production)	jusqu'à 12 mois	< 3 mois	Améliorez le rapport vitesse/valeur par 4x
2	Simplification des processus	Délai de mise en production des cas d'usage d'IA	3 à 6 mois	< 2 semaines	Gain de productivité x 8

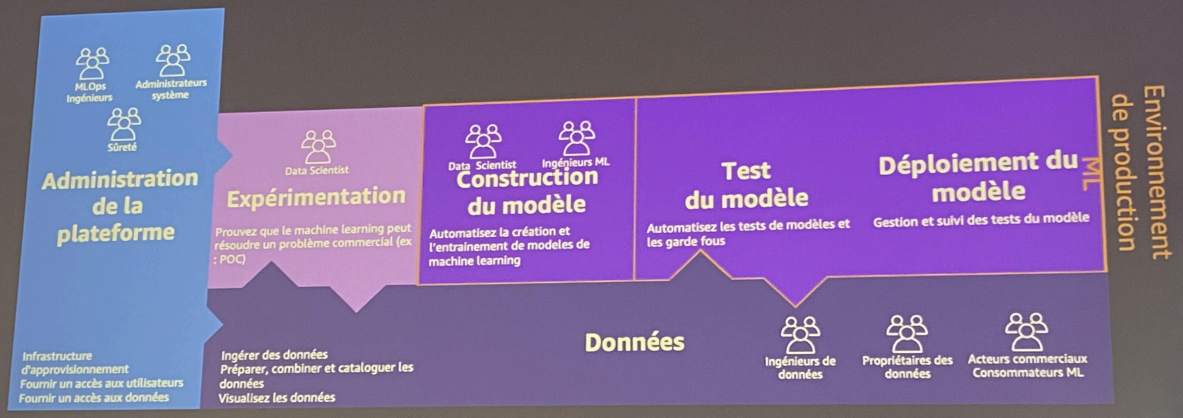
Bénéfices attendus du MLOps

STANDARDISATION DES OPÉRATIONS ET DE L'INFRASTRUCTURE DE VOS SOLUTIONS D'IA

	Objectif	Métrique technique	Avant MLOPS	Résultats attendus avec MLOps	Valeur ajoutée
1	Efficacité d'exécution	Délai de rentabilisation (de l'idée à la production)	jusqu'à 12 mois	< 3 mois	Améliorez le rapport vitesse/valeur par 4x
2	Simplification des processus	Délai de mise en production des cas d'usage d'IA	3 à 6 mois	< 2 semaines	Gain de productivité x 8
3	Standardisation de l'infrastructure, des données et du code	% de réutilisation des pipelines existantes	s.o.	> 85 %	Mettez l'accent sur l'innovation en augmentant la réutilisabilité de 85 %
4	Standardisation de l'intégration des nouvelles équipes et des cas d'utilisation du ML	Délai de mise en place de nouvelles infrastructures MLOps et de projets de ML	40 jours	< 1 heure	Accélérez l'adoption du ML dans tous les domaines d'activité
5	Garantie de normes de sécurité élevées	Exécution des solutions ML sans accès à Internet dans un espace cloud privatif	s.o.	Pas d'Internet	Vos données sont en sécurité dans votre espace cloud privatif

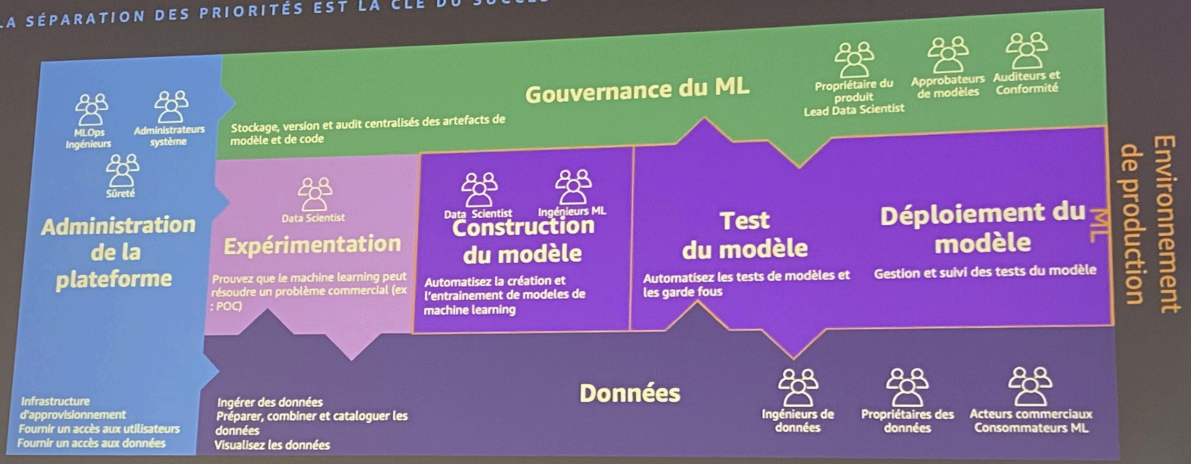
Acteurs et processus du MLOps

LA SÉPARATION DES PRIORITÉS EST LA CLÉ DU SUCCÈS



Acteurs et processus du MLOps

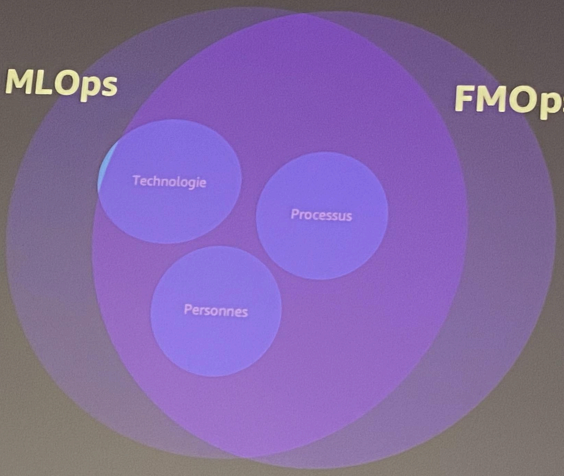
LA SÉPARATION DES PRIORITÉS EST LA CLÉ DU SUCCÈS



Définitions clés

Opérations d'apprentissage automatique
Productionnalisez efficacement les solutions de machine learning

MLOps



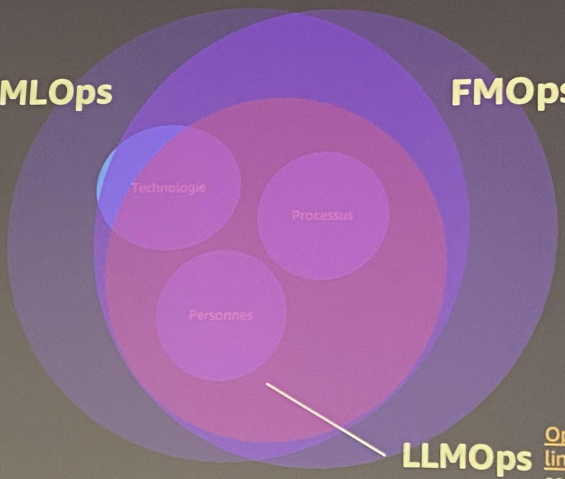
FMOps

Opérations des modèles de fondation : permet de produire des solutions d'IA génératives (texte-texte/image/vidéo/audio/…)

Définitions clés

Opérations d'apprentissage automatique
Productionnalisez efficacement les solutions de machine learning

MLOps



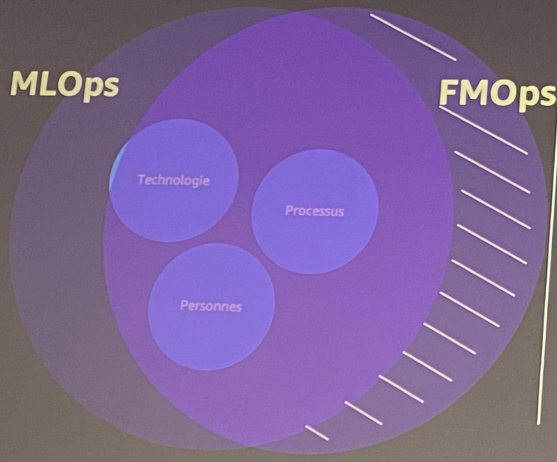
FMOps

Opérations des modèles de fondation : permet de produire des solutions d'IA génératives (texte-texte/image/vidéo/audio/…)

LLMOps

Opérations des grands modèles linguistiques Produisez des solutions basées sur les grands modèles de language

Différences entre le MLops et le FMOps



Profils et processus :
Fournisseurs, fine-tuners et consommateurs

Sélection et personnalisation du FM en fonction d'un contexte spécifique
- Fine-tuning, fine-tuning avec efficacité des paramètres, prompt engineering
- Modèle propriétaire ou open source selon l'application

Évaluation et monitoring des modèles fine tunes
Feedback humain, vitesse d'exécution, toxicité/biais...

Déploiement
Des données et des modèles, confidentialité des données, mutualisation, coût, latence et précision

Technologie
MLOps données et couches applicatives

Les profils utilisateurs de l'IA générative

Types de profils	Fournisseurs	Fine Tuners	Consommateurs
	Pré-entrainent les FM à partir de zéro et les fournissent sous forme de produit aux fine tuners et aux consommateurs.	Personnalisent les FM pré-entraînés des fournisseurs avec leurs propres données et effectuent des inférences, tout en fournissant un accès aux consommateurs.	Interagissent avec les services d'IA générative du fournisseur ou du fine tuner à l'aide d'instructions textuelles ou d'une interface visuelle pour effectuer les actions souhaitées.
Compétences	Approche approfondie de bout en bout du machine learning, du NLP et de la data science, labélisation des données.	Solide expertise du ML de bout en bout et connaissance du domaine pour le fine tuning, y compris en prompt engineering	Aucune expertise en ML n'est requise. Se concentrent sur le prompt engineering et le RAG

Les profils utilisateurs de l'IA générative




peut également être

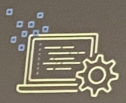
peut devenir

Types de profils	Fournisseurs	Fine Tuners	Consommateurs
	Pré-entrainent les FM à partir de zéro et les fournissent sous forme de produit aux fine tuners et aux consommateurs.	Personnalisent les FM pré-entraînés des fournisseurs avec leurs propres données et effectuent des inférences, tout en fournissant un accès aux consommateurs.	Interagissent avec les services d'IA générative du fournisseur ou du fine tuner à l'aide d'instructions textuelles ou d'une interface visuelle pour effectuer les actions souhaitées.
Compétences	Approche approfondie de bout en bout du machine learning, du NLP et de la data science, labélisation des données.	Solide expertise du ML de bout en bout et connaissance du domaine pour le fine tuning, y compris en prompt engineering	Aucune expertise en ML n'est requise. Se concentrent sur le prompt engineering et le RAG

Productionalisation de grands modèles en tirant parti du ML et des opérations (FMOps)

Productionnalisation des applications d'IA générative en tirant parti du FMOps

Processus d'IA générative — Consommateurs



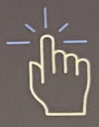
Sélectionnez, évaluez et utilisez le FM et adaptez au contexte

Utilisation de chained modèles et de techniques de prompt engineering pour s'adapter au contexte (si nécessaire).
Exposition de la solution aux utilisateurs finaux

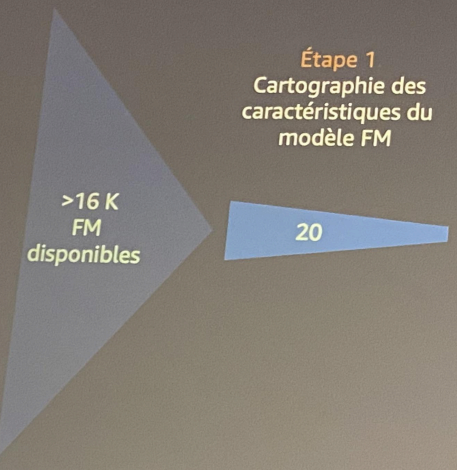
→

Entrées/sorties et évaluation

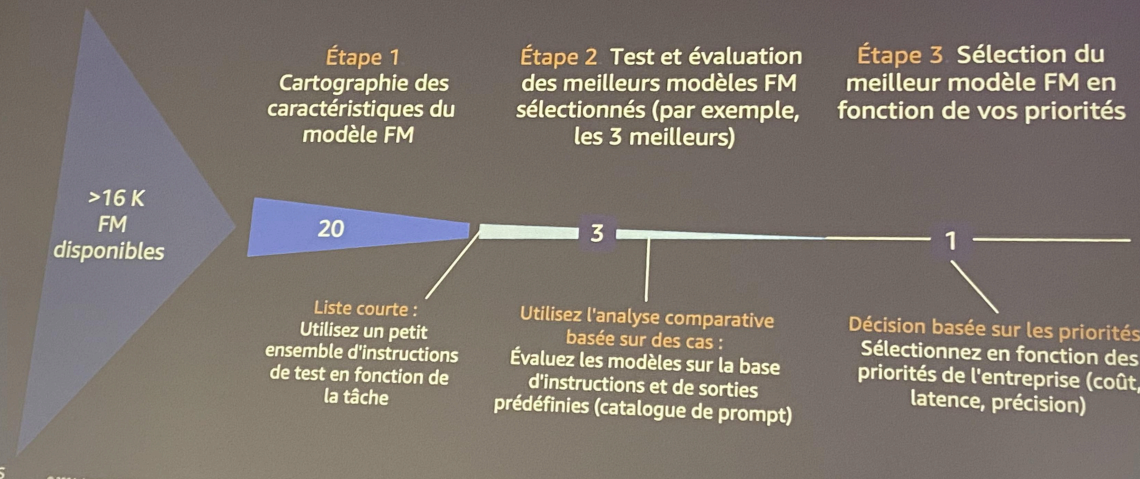
Interaction avec les solutions d'IA générative.
Améliorations des résultats et du prompt engineering grâce a l' évaluation



Sélection du modèle FM - Consommateurs



Sélection du modèle FM - Consommateurs

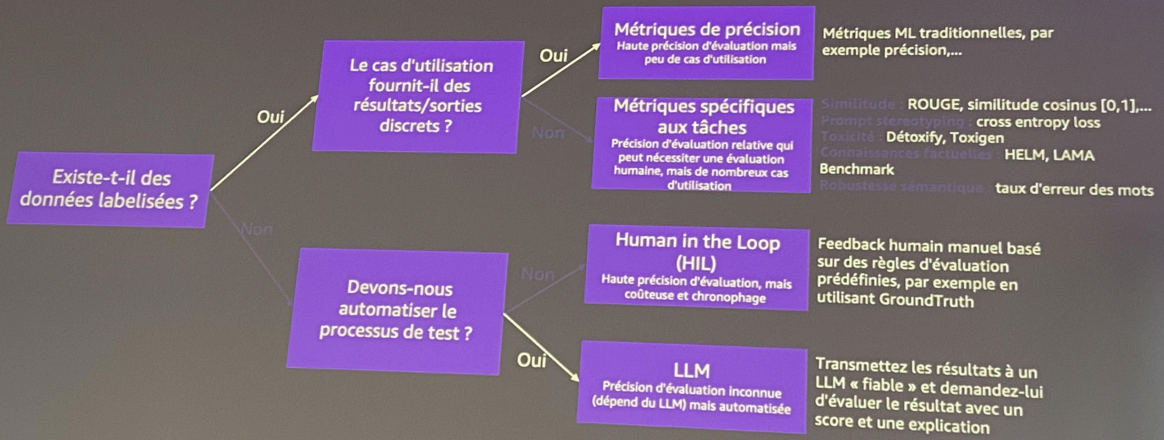


Étape 1: Découvrez les principales fonctionnalités FM

Matrice des caractéristiques d'un modèle FM

- Modèle FM propriétaire ou open source
- Licence commerciale
- Fine tunable
- Vitesse
- Nombre de paramètres
- Fenêtre contextuelle (#tokens)
- Entraînement (instructions, code, données Internet)
- Qualité
- Compétences internes

Étape 2 : Évaluez les meilleurs modèles FM



Étape 3 : Sélectionnez le meilleur FM en fonction de vos priorités

Modèle	Vitesse
FM1	⚡⚡
FM2	⚡
FM3	⚡

Vitesse — Aucune priorité

Faible latence, petit modèle, faible précision, moindre coût

P1 : Précision — Précision

Coût — P0 : moindre coût

Modèle	Score d'évaluation	Feedback HIL/LLM
FM1	5/5	<Feedback summary>
FM2	4/5	<Feedback summary>
FM3	3/5	<Feedback summary>

Vitesse inférieure, modèle plus grand, précision supérieure, coût plus élevé

Étape 3 : Sélectionnez le meilleur FM en fonction de vos priorités

Modèle	Vitesse
FM1	⚡⚡
FM2	⚡
FM3	⚡

Vitesse — Aucune priorité

Sélection du modèle : FM2

Faible latence, petit modèle, faible précision, moindre coût

P1 : Précision — Précision

Coût — P0 : moindre coût

Modèle	Score d'évaluation	Feedback HIL/LLM
FM1	5/5	<Feedback summary>
FM2	4/5	<Feedback summary>
FM3	3/5	<Feedback summary>

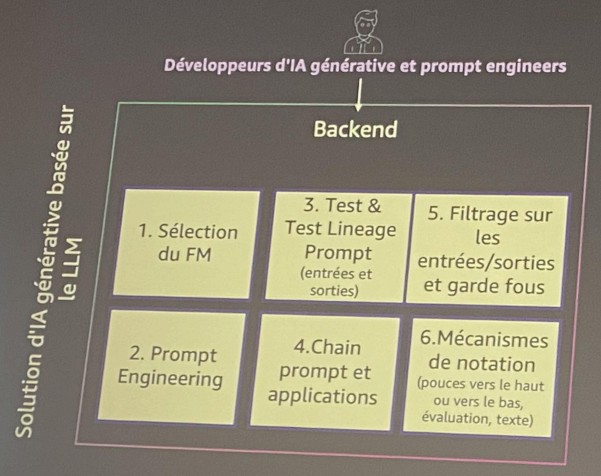
Vitesse inférieure, modèle plus grand, précision supérieure, coût plus élevé

Modèle	Coût
FM1	$$$$
FM2	$
FM3	$$$

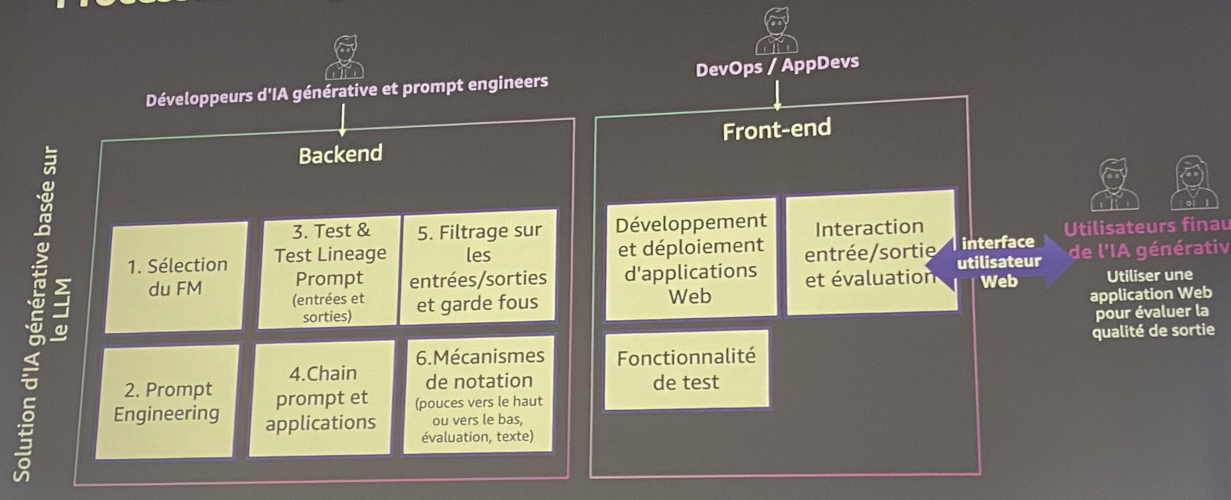
aws © 2024, Amazon Web Services, Inc. ou ses filiales. Tous droits réservés.

20

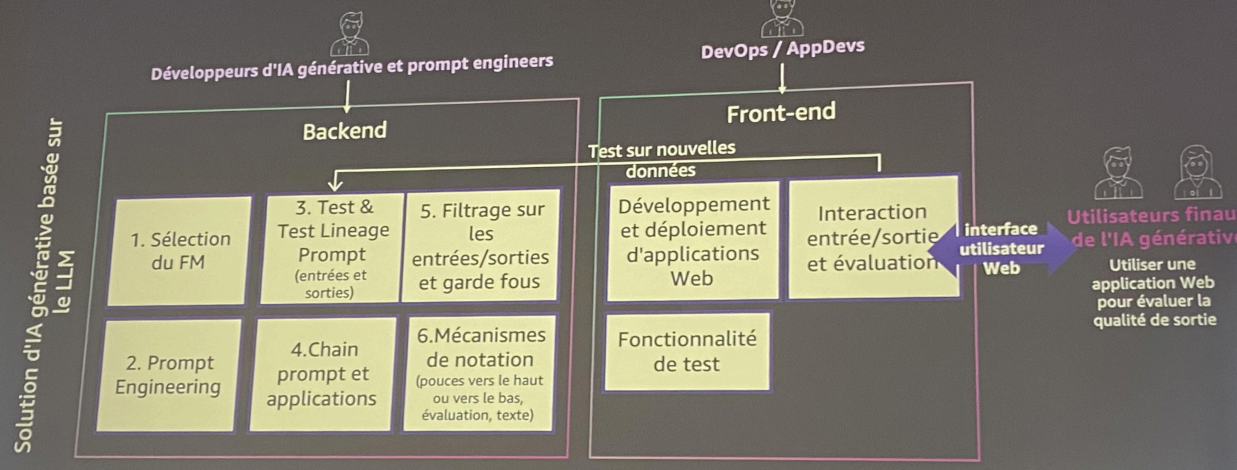
Processus d'IA générative pour le LLM- Consommateurs



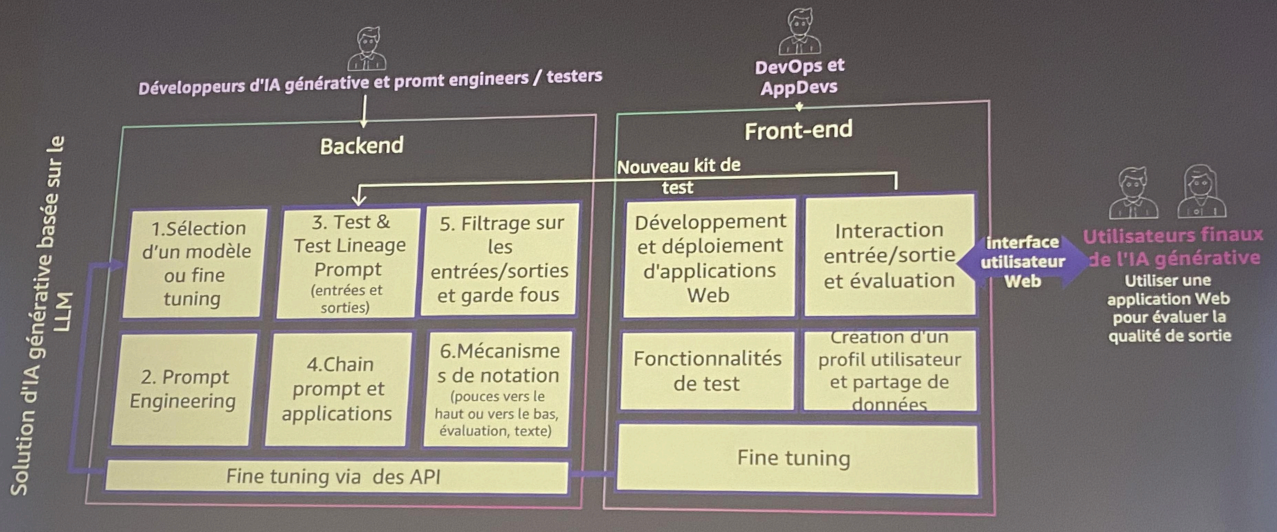
Processus d'IA générative pour le LLM- Consommateurs



Processus d'IA générative pour le LLM- Consommateurs



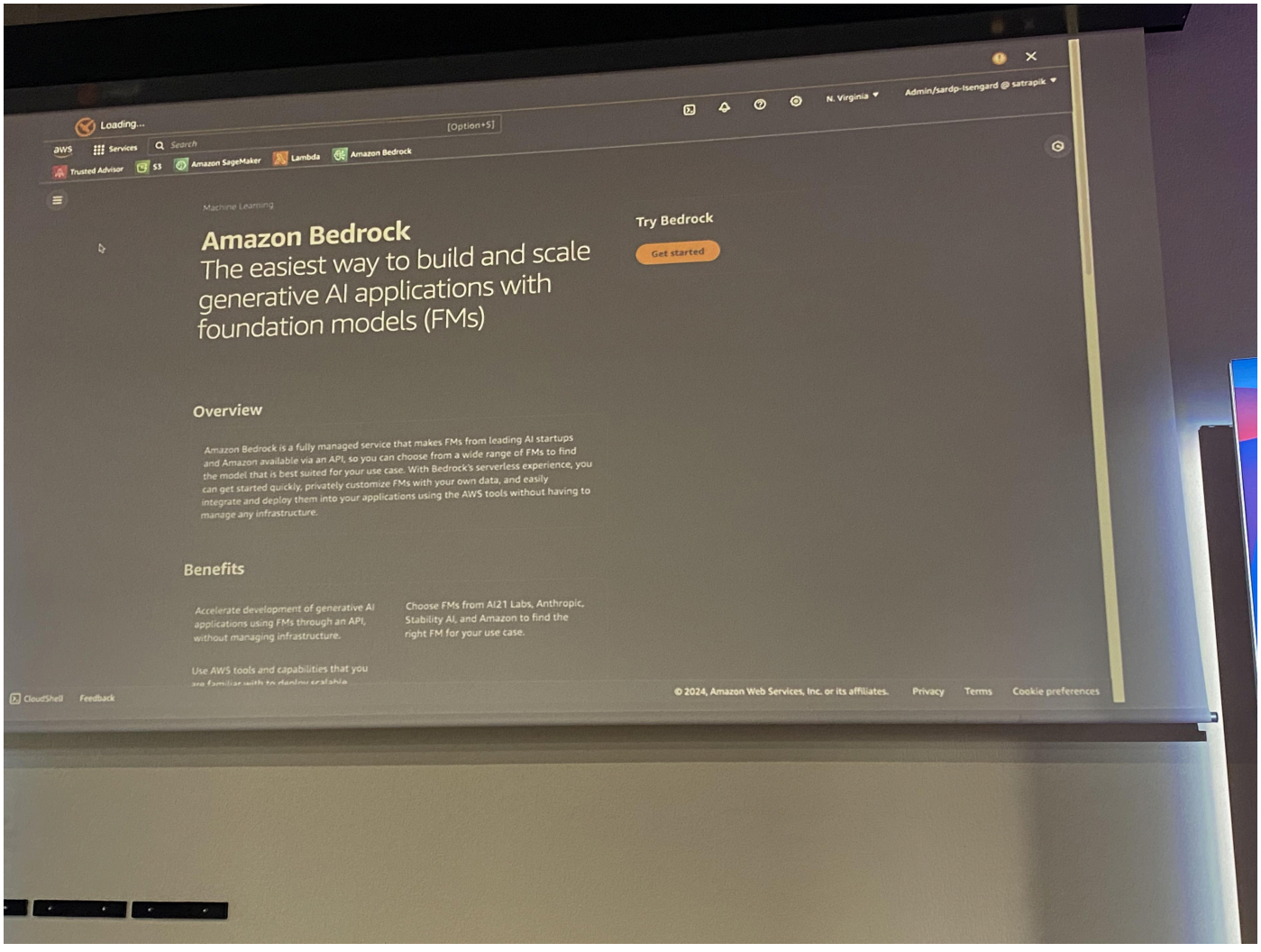
Processus d'IA générative pour le LLM- Consommateurs

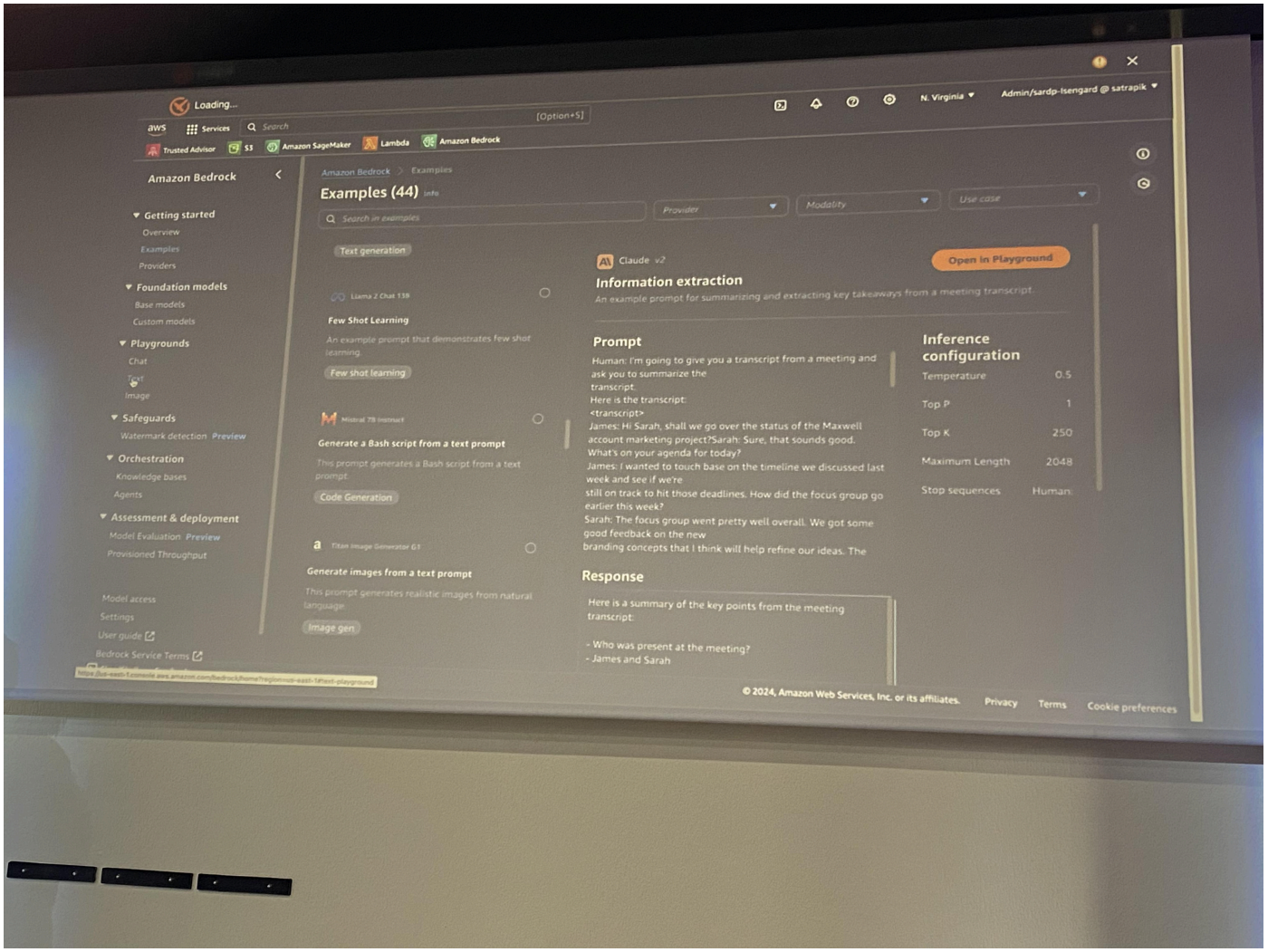


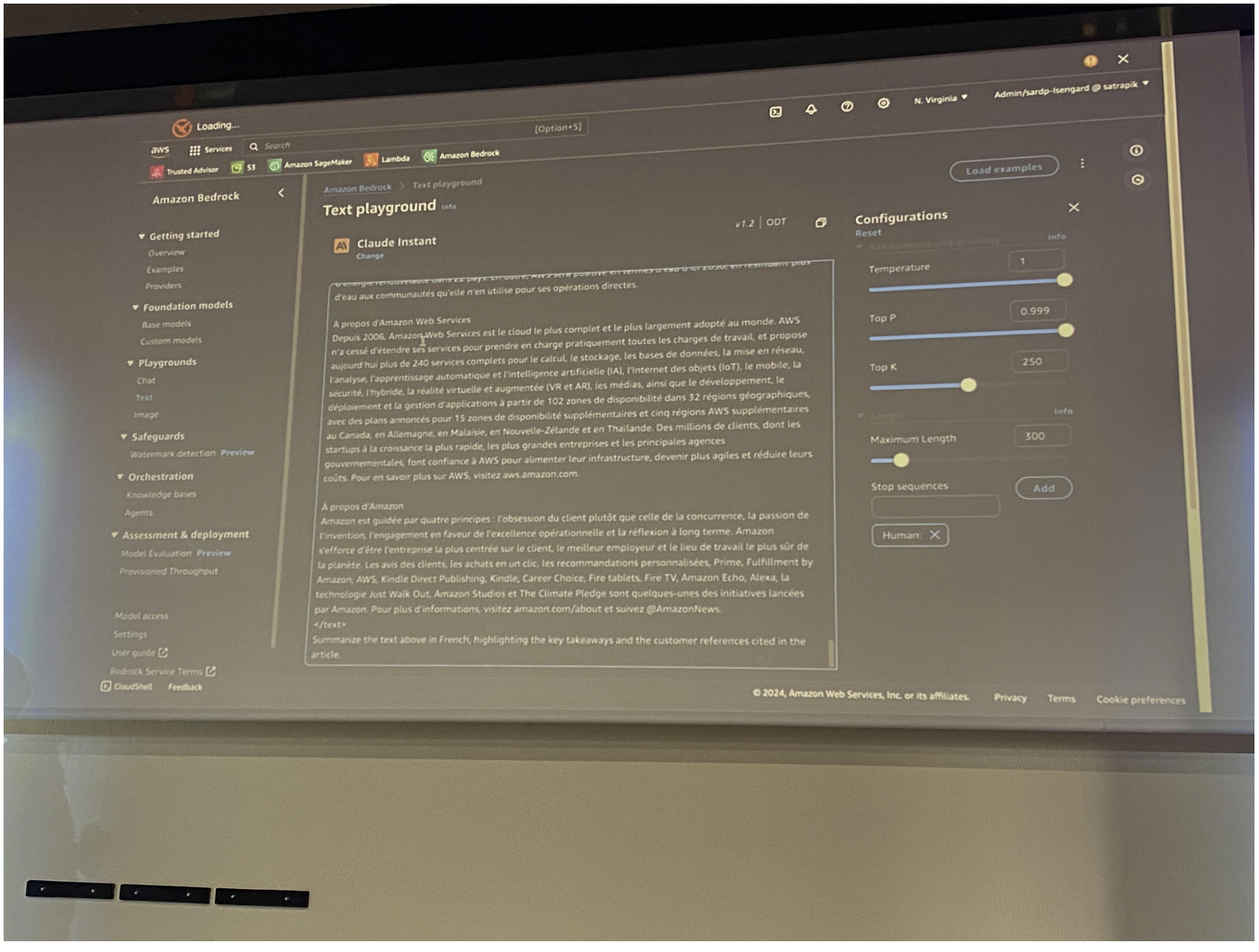
Amazon Bedrock

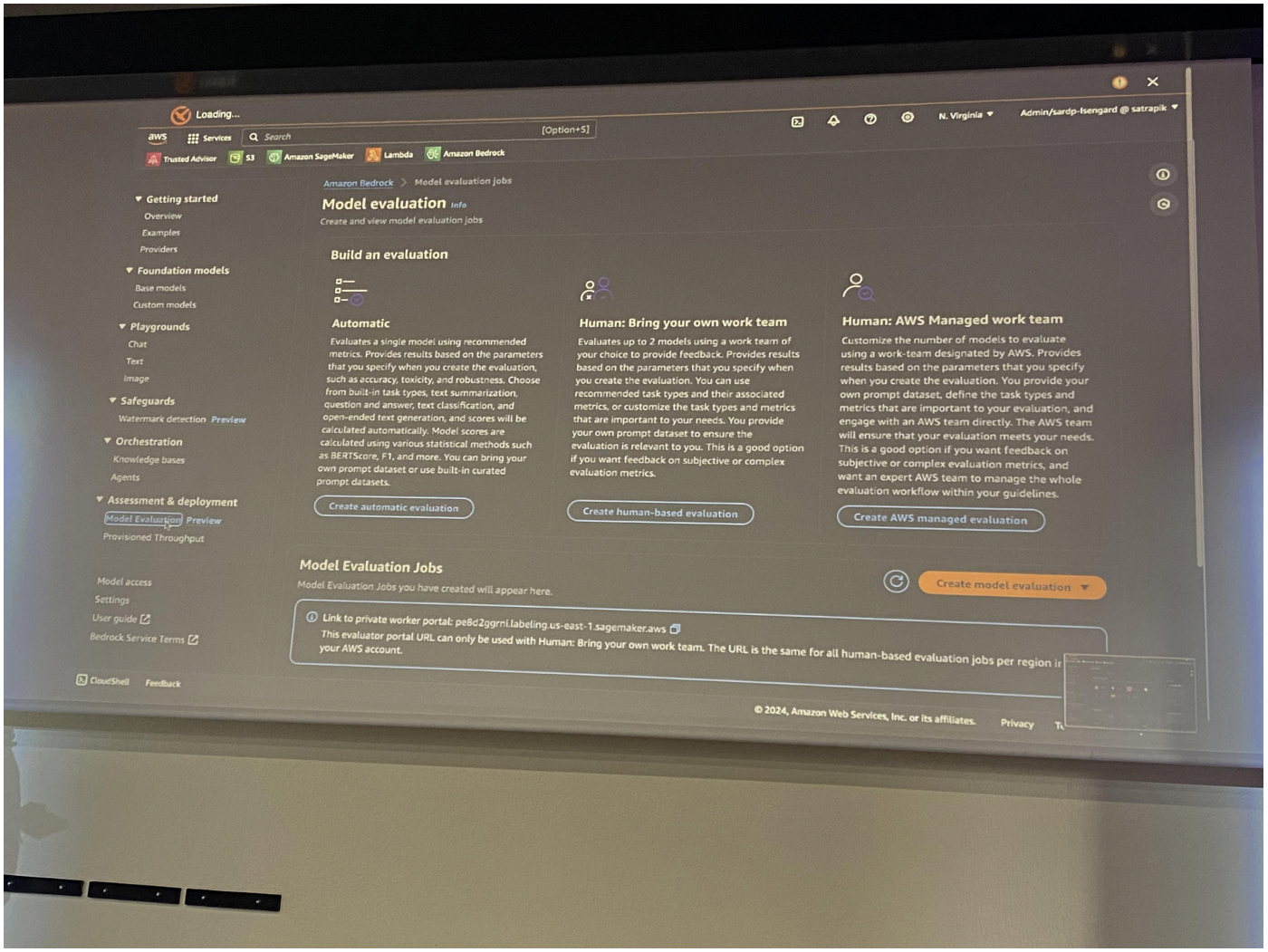
Le moyen le plus simple de créer et de faire évoluer des applications d'IA générative avec des modèles de base à l'aide d'une simple API

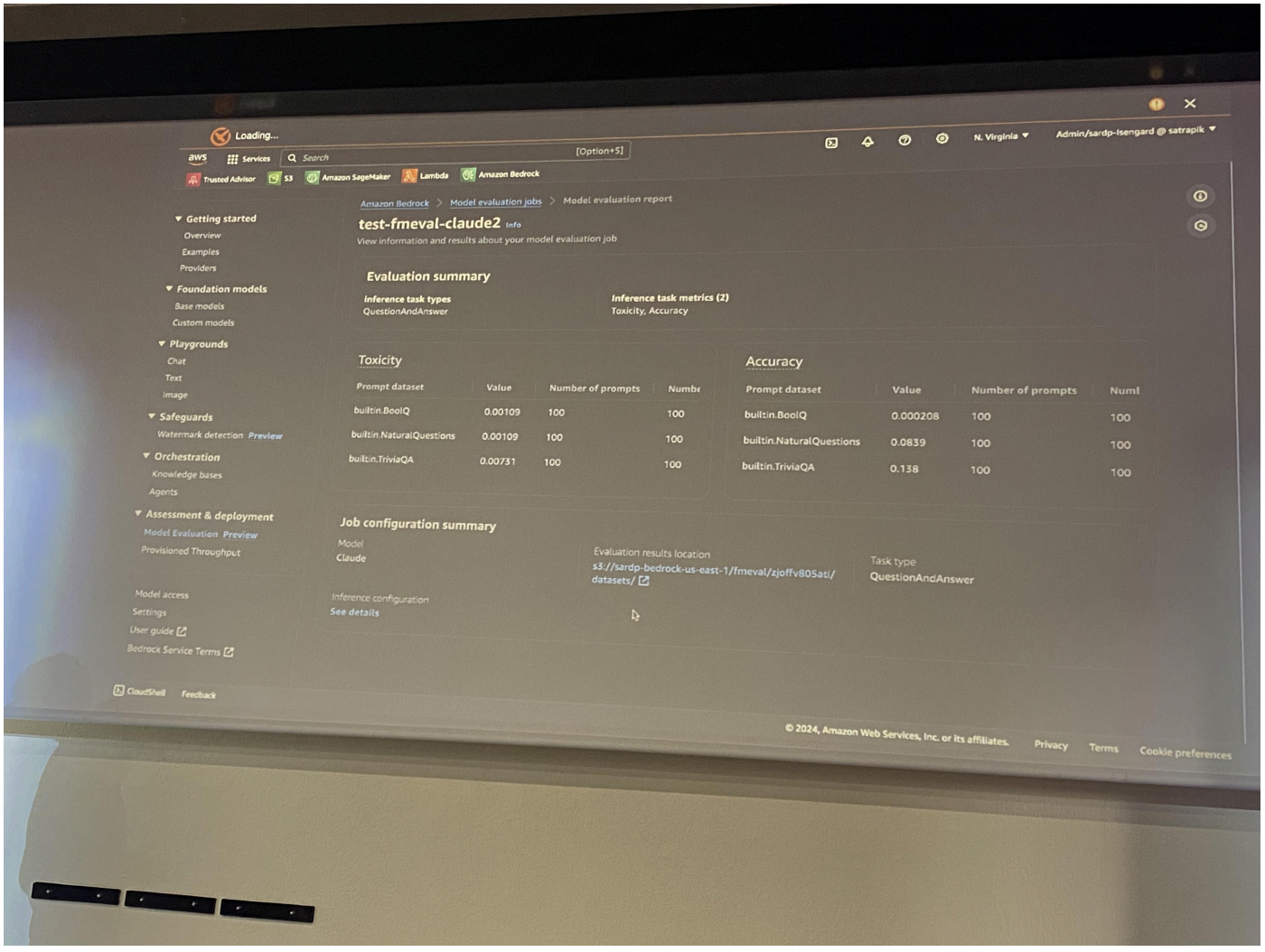
- Choix des principaux FM via une seule API
- Personnalisation du modèle (fine tuning)
- (RAG) à l'aide des agents Amazon Bedrock et de la knowledge base
- Application plus fiable en utilisant Amazon Bedrock Guardrails
- Sécurité, confidentialité et sûreté




Loading...
aws
Services
Search
[Option+S]
Trusted Advisor
S3
Amazon SageMaker
Lambda
Amazon Bedrock
N. Virginia
Admin/sardp-lsengard @ satrapik
Amazon Bedrock
Getting started
Overview
Examples
Providers
Foundation models
Base models
Custom models
Playgrounds
Chat
Text
Image
Safeguards
Watermark detection Preview
Orchestration
Knowledge bases
Agents
Assessment & deployment
Model Evaluation Preview
Provisioned Throughput
Model access
Settings
User guide
Bedrock Service Terms
Amazon Bedrock > Examples
Examples (44) Info
Search in examples
Provider
Modality
Use case
Text generation
Llama 2 Chat 13B
Few Shot Learning
An example prompt that demonstrates few shot learning.
Few shot learning
Mistral 7B Instruct
Generate a Bash script from a text prompt
This prompt generates a Bash script from a text prompt.
Code Generation
Titan Image Generator G1
Generate images from a text prompt
This prompt generates realistic images from natural language.
Image gen
Claude v2
Open in Playground
Information extraction
An example prompt for summarizing and extracting key takeaways from a meeting transcript.
Prompt
Human: I'm going to give you a transcript from a meeting and ask you to summarize the
transcript.
Here is the transcript:
<transcript>
James: Hi Sarah, shall we go over the status of the Maxwell account marketing project?Sarah: Sure, that sounds good. What's on your agenda for today?
James: I wanted to touch base on the timeline we discussed last week and see if we're
still on track to hit those deadlines. How did the focus group go earlier this week?
Sarah: The focus group went pretty well overall. We got some good feedback on the new
branding concepts that I think will help refine our ideas. The
Response
Here is a summary of the key points from the meeting transcript:
- Who was present at the meeting?
- James and Sarah
Inference configuration
Temperature 0.5
Top P 1
Top K 250
Maximum Length 2048
Stop sequences Human:
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences


Loading...
aws
Services
Search
[Option+S]
Trusted Advisor
S3
Amazon SageMaker
Lambda
Amazon Bedrock
N. Virginia
Admin/sardp-lsengard @ satrapik
Amazon Bedrock
Getting started
Overview
Examples
Providers
Foundation models
Base models
Custom models
Playgrounds
Chat
Text
Image
Safeguards
Watermark detection Preview
Orchestration
Knowledge bases
Agents
Assessment & deployment
Model Evaluation Preview
Provisioned Throughput
Model access
Settings
User guide
Bedrock Service Terms
CloudShell
Feedback
Amazon Bedrock > Text playground
Text playground Info
Load examples
Claude Instant
Change
v1.2 | ODT
d'eau aux communautés qu'elle n'en utilise pour ses opérations directes.
À propos d'Amazon Web Services
Depuis 2006, Amazon Web Services est le cloud le plus complet et le plus largement adopté au monde. AWS n'a cessé d'étendre ses services pour prendre en charge pratiquement toutes les charges de travail, et propose aujourd'hui plus de 240 services complets pour le calcul, le stockage, les bases de données, la mise en réseau, l'analyse, l'apprentissage automatique et l'intelligence artificielle (IA), l'Internet des objets (IoT), le mobile, la sécurité, l'hybride, la réalité virtuelle et augmentée (VR et AR), les médias, ainsi que le développement, le déploiement et la gestion d'applications à partir de 102 zones de disponibilité dans 32 régions géographiques, avec des plans annoncés pour 15 zones de disponibilité supplémentaires et cinq régions AWS supplémentaires au Canada, en Allemagne, en Malaisie, en Nouvelle-Zélande et en Thaïlande. Des millions de clients, dont les startups à la croissance la plus rapide, les plus grandes entreprises et les principales agences gouvernementales, font confiance à AWS pour alimenter leur infrastructure, devenir plus agiles et réduire leurs coûts. Pour en savoir plus sur AWS, visitez aws.amazon.com.
À propos d'Amazon
Amazon est guidée par quatre principes : l'obsession du client plutôt que celle de la concurrence, la passion de l'invention, l'engagement en faveur de l'excellence opérationnelle et la réflexion à long terme. Amazon s'efforce d'être l'entreprise la plus centrée sur le client, le meilleur employeur et le lieu de travail le plus sûr de la planète. Les avis des clients, les achats en un clic, les recommandations personnalisées, Prime, Fulfillment by Amazon, AWS, Kindle Direct Publishing, Kindle, Career Choice, Fire tablets, Fire TV, Amazon Echo, Alexa, la technologie Just Walk Out, Amazon Studios et The Climate Pledge sont quelques-unes des initiatives lancées par Amazon. Pour plus d'informations, visitez amazon.com/about et suivez @AmazonNews.
</text>
Summarize the text above in French, highlighting the key takeaways and the customer references cited in the article.
Configurations
Reset
Temperature 1
Top P 0.999
Top K 250
Maximum Length 300
Stop sequences
Add
Human:
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences


Loading...
aws
Services
Search
[Option+5]
N. Virginia
Admin/sardp-lsengard @ satrapik
Trusted Advisor
S3
Amazon SageMaker
Lambda
Amazon Bedrock
Amazon Bedrock > Model evaluation jobs
Getting started
Overview
Examples
Providers
Foundation models
Base models
Custom models
Playgrounds
Chat
Text
Image
Safeguards
Watermark detection Preview
Orchestration
Knowledge bases
Agents
Assessment & deployment
Model Evaluation Preview
Provisioned Throughput
Model access
Settings
User guide
Bedrock Service Terms
Model evaluation Info
Create and view model evaluation jobs
Build an evaluation
Automatic
Evaluates a single model using recommended metrics. Provides results based on the parameters that you specify when you create the evaluation, such as accuracy, toxicity, and robustness. Choose from built-in task types, text summarization, question and answer, text classification, and open-ended text generation, and scores will be calculated automatically. Model scores are calculated using various statistical methods such as BERTScore, F1, and more. You can bring your own prompt dataset or use built-in curated prompt datasets.
Create automatic evaluation
Human: Bring your own work team
Evaluates up to 2 models using a work team of your choice to provide feedback. Provides results based on the parameters that you specify when you create the evaluation. You can use recommended task types and their associated metrics, or customize the task types and metrics that are important to your needs. You provide your own prompt dataset to ensure the evaluation is relevant to you. This is a good option if you want feedback on subjective or complex evaluation metrics.
Create human-based evaluation
Human: AWS Managed work team
Customize the number of models to evaluate using a work-team designated by AWS. Provides results based on the parameters that you specify when you create the evaluation. You provide your own prompt dataset, define the task types and metrics that are important to your evaluation, and engage with an AWS team directly. The AWS team will ensure that your evaluation meets your needs. This is a good option if you want feedback on subjective or complex evaluation metrics, and want an expert AWS team to manage the whole evaluation workflow within your guidelines.
Create AWS managed evaluation
Model Evaluation Jobs
Model Evaluation Jobs you have created will appear here.
Create model evaluation
Link to private worker portal: pe8d2ggrnl.labeling.us-east-1.sagemaker.aws
This evaluator portal URL can only be used with Human: Bring your own work team. The URL is the same for all human-based evaluation jobs per region in your AWS account.
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy


Loading...
aws
Services
Search
[Option+5]
N. Virginia
Admin/sardp-lsengard @ satrapik
Trusted Advisor
S3
Amazon SageMaker
Lambda
Amazon Bedrock
Amazon Bedrock > Model evaluation jobs > Model evaluation report
Getting started
Overview
Examples
Providers
Foundation models
Base models
Custom models
Playgrounds
Chat
Text
Image
Safeguards
Watermark detection Preview
Orchestration
Knowledge bases
Agents
Assessment & deployment
Model Evaluation Preview
Provisioned Throughput
Model access
Settings
User guide
Bedrock Service Terms
test-fmeval-claude2 Info
View information and results about your model evaluation job
Evaluation summary
Inference task types
QuestionAndAnswer
Inference task metrics (2)
Toxicity, Accuracy
Toxicity
Prompt dataset | Value | Number of prompts | Numb
builtin.BoolQ | 0.00109 | 100 | 100
builtin.NaturalQuestions | 0.00109 | 100 | 100
builtin.TriviaQA | 0.00731 | 100 | 100
Accuracy
Prompt dataset | Value | Number of prompts | Numl
builtin.BoolQ | 0.000208 | 100 | 100
builtin.NaturalQuestions | 0.0839 | 100 | 100
builtin.TriviaQA | 0.138 | 100 | 100
Job configuration summary
Model
Claude
Evaluation results location
s3://sardp-bedrock-us-east-1/fmeval/zjoffv80Sati/datasets/
Task type
QuestionAndAnswer
Inference configuration
See details
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

Le parcours des Fine-Tuners

Processus d'IA générative - Fine-Tuners

Labélisation des données

Un humain impliqué dans la boucle pour annoter des centaines voire des milliers de données

Fine tuning

Personnalisation pour des domaines spécifiques

Processus d'IA générative - Fine-Tuners

Labélisation des données

Un humain impliqué dans la boucle pour annoter des centaines voire des milliers de données

Fine tuning

Personnalisation pour des domaines spécifiques

Déploiement et prompt engineering

Compromis entre coût, précision et latence

Chaîne de modèles

Conception et prompt engineering

Filtrer les demandes de saisie et les résultats à l'aide d'embeddings

Processus d'IA générative - Fine-Tuners

Labélisation des données

Un humain impliqué dans la boucle pour annoter des centaines voire des milliers de données

Fine tuning

Personnalisation pour des domaines spécifiques

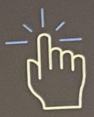
Déploiement et prompt engineering

Compromis entre coût, précision et latence

Chaîne de modèles

Conception et prompt engineering

Filtrer les demandes de saisie et les résultats à l'aide d'embeddings



Surveillance

Feedback et évaluation humaine, similitude des résultats, taux de toxicité, nouvelles méthodes en cours de recherche

Processus d'IA générative - Fine-Tuners

Labélisation des données

Un humain impliqué dans la boucle pour annoter des centaines voire des milliers de données

Fine tuning

Personnalisation pour des domaines spécifiques

Déploiement et prompt engineering

Compromis entre coût, précision et latence

Chaîne de modèles

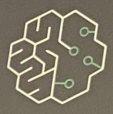
Conception et prompt engineering

Filtrer les demandes de saisie et les résultats à l'aide d'embeddings

Surveillance

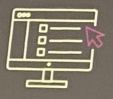
Feedback et évaluation humaine, similitude des résultats, taux de toxicité, nouvelles méthodes en cours de recherche

Les fonctionnalités de Sagemaker pour l'IA génératives



Amazon SageMaker Gen AI pour les développeurs

Les data scientists et les développeurs du machine learning pré-entrainent, personnalisent et déploient des machines virtuelles avec des contrôles précis pour les cas d'utilisation de l'IA générative.

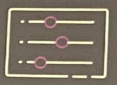


Naviguez

Parcourez plus de 730 FM publics et plus de 20 FM propriétaires sur SageMaker Jumpstart

Evaluer

Évaluez plusieurs FM avec SageMaker Clarify Model Evaluation afin de sélectionner celui qui convient le mieux à votre cas d'utilisation



Personnaliser

Personnalisez vos modèles grâce à SageMaker Training à l'aide de votre propre jeu de données.

Déployer

Déployez le modèle avec SageMaker Inference et exécutez l'inférence

Les fonctionnalités de Sagemaker pour l'IA générative

Amazon SageMaker Gen AI pour les développeurs

Les data scientists et les développeurs du machine learning pré-entrainent, personnalisent et déploient des machines virtuelles avec des contrôles précis pour les cas d'utilisation de l'IA générative.

Naviguez → **Evaluer** → **Personnaliser** → **Déployer**

Naviguez — Parcourez plus de 730 FM publics et plus de 20 FM propriétaires sur SageMaker Jumpstart

Evaluer — Évaluez plusieurs FM avec SageMaker Clarify Model Evaluation afin de sélectionner celui qui convient le mieux à votre cas d'utilisation

Personnaliser — Personnalisez vos modèles grâce à SageMaker Training à l'aide de votre propre jeu de données.

Déployer — Déployez le modèle avec SageMaker Inference et exécutez l'inférence

EXPÉRIENCE

Personnalisez et évaluez les modèles de manière itérative avec SageMaker Experiments

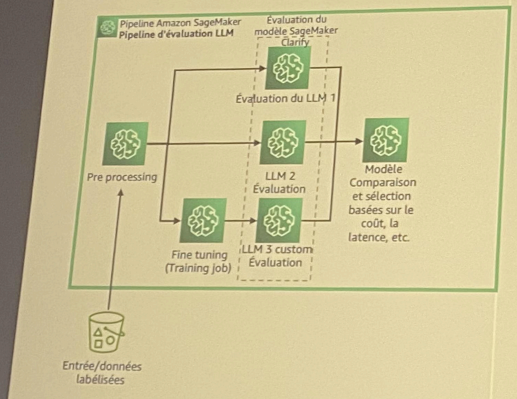
Industrialisez avec les MLops et la surveillance

Automatisez la sélection, l'évaluation et le déploiement des modèles avec SageMaker Pipelines

Exemple d'un cycle de vie FMOps

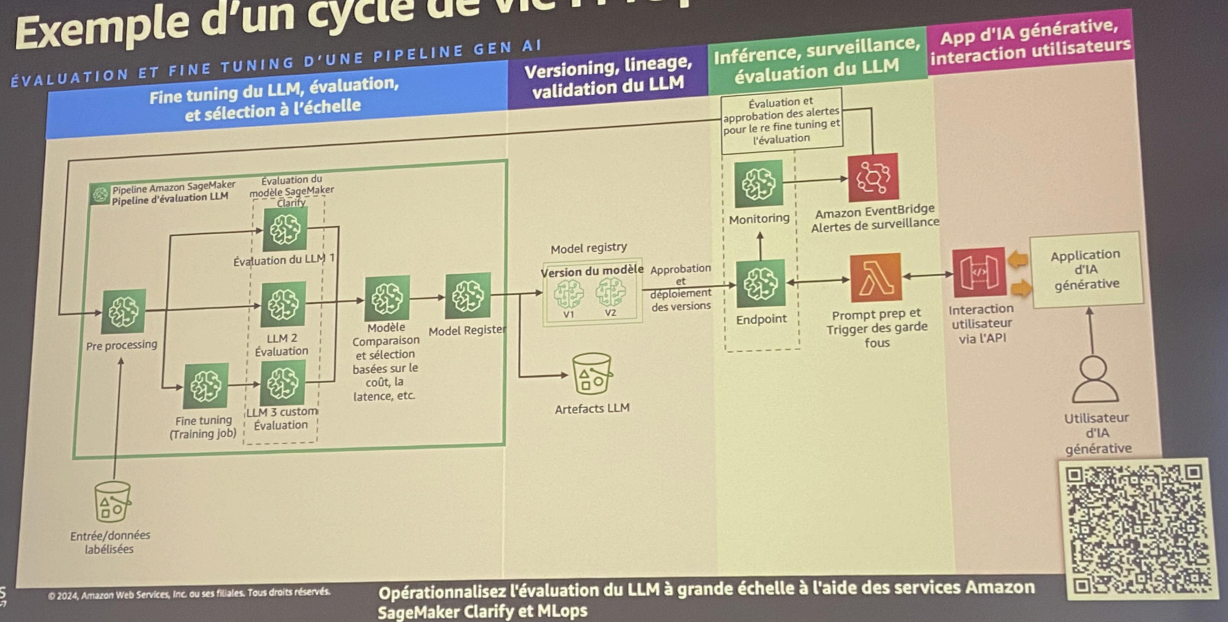
ÉVALUATION ET FINE TUNING D'UNE PIPELINE GEN AI

Fine tuning du LLM, évaluation, et sélection à l'échelle	Versioning, lineage, validation du LLM	Inférence, surveillance, évaluation du LLM	App d'IA générative, interaction utilisateurs

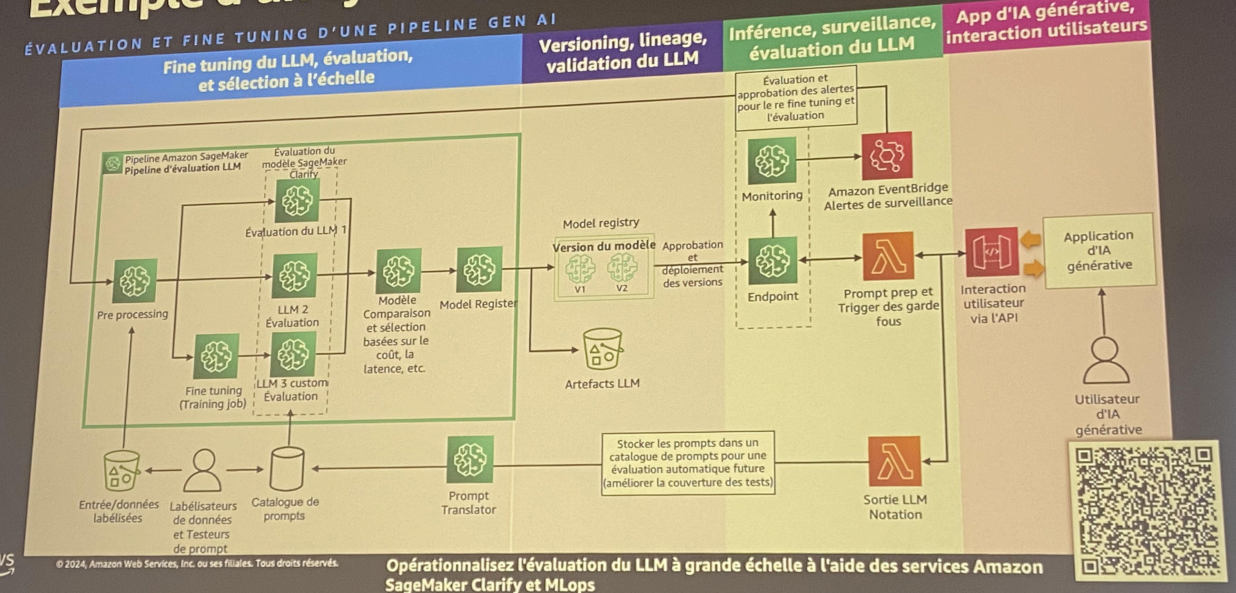


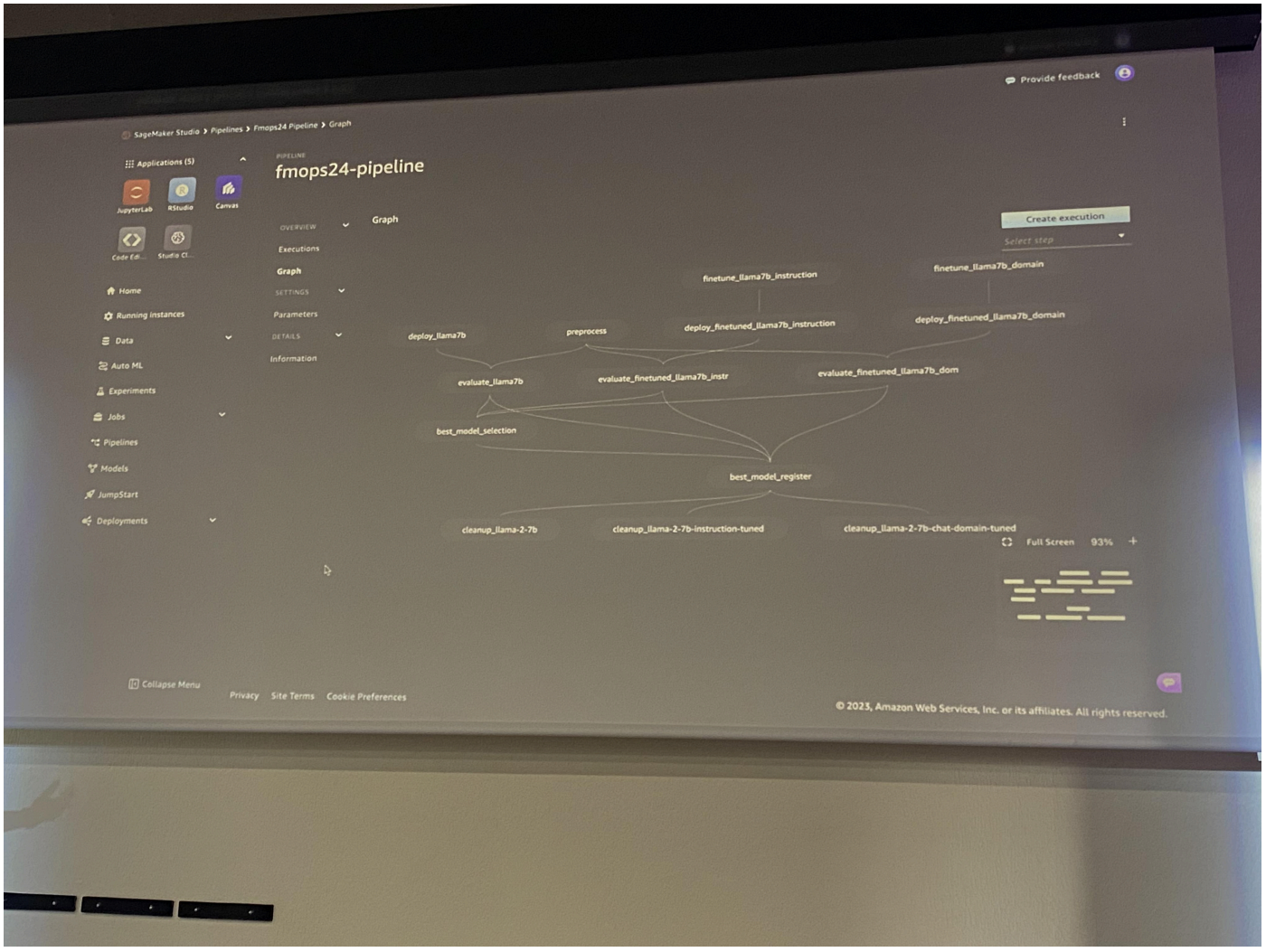
Opérationnalisez l'évaluation du LLM à grande échelle à l'aide des services Amazon SageMaker Clarify et MLops

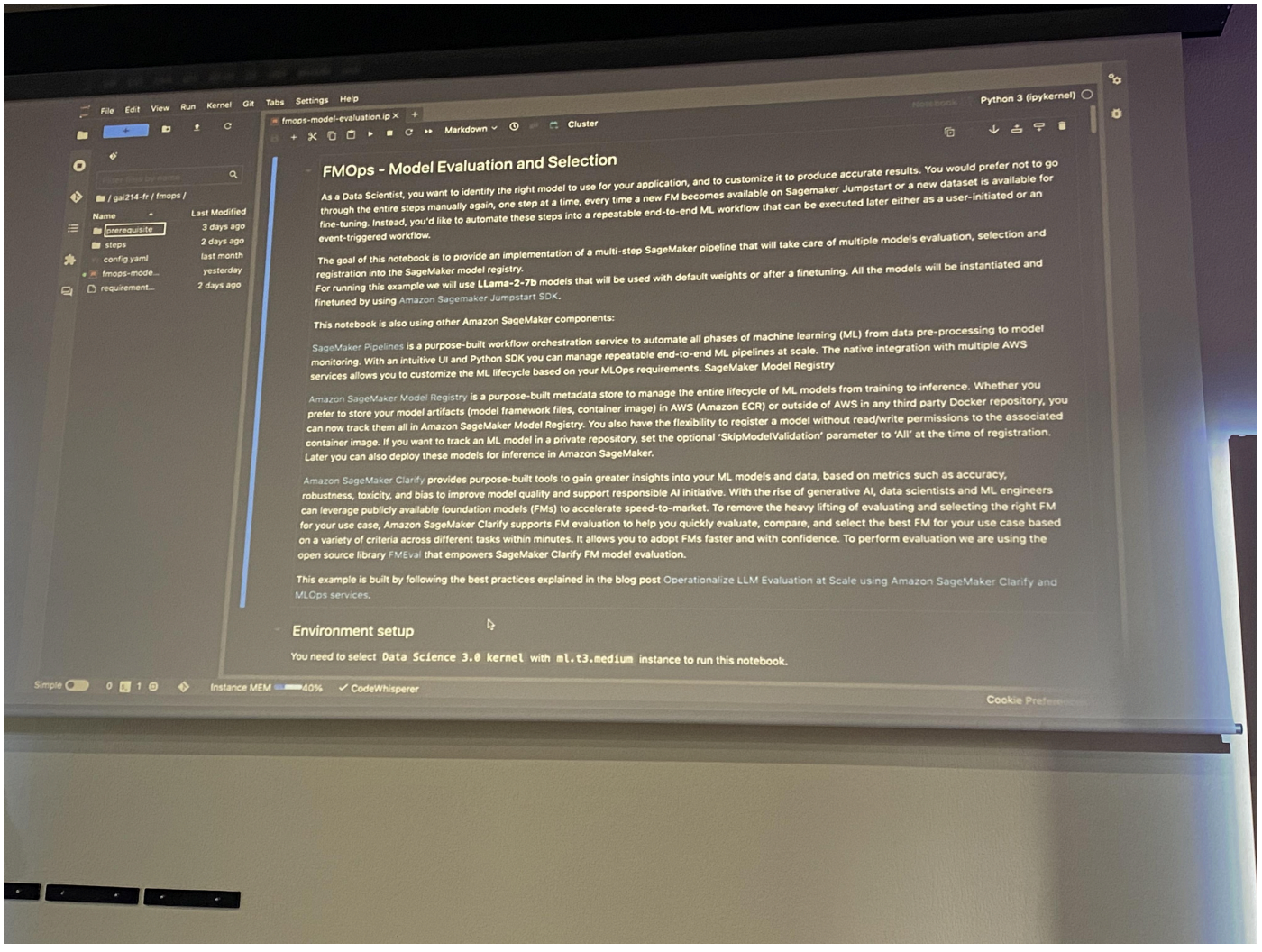
Exemple d'un cycle de vie FMOps

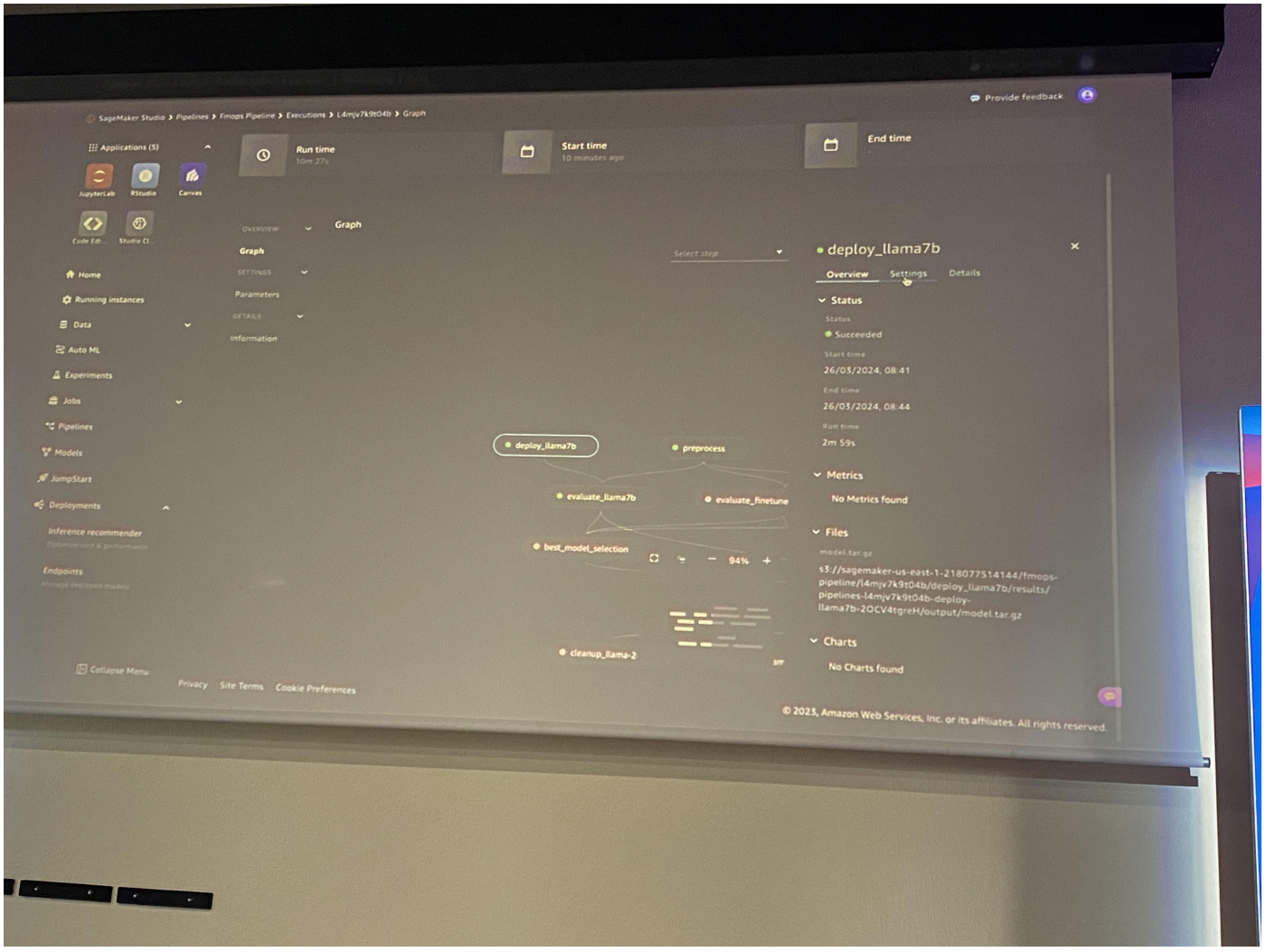


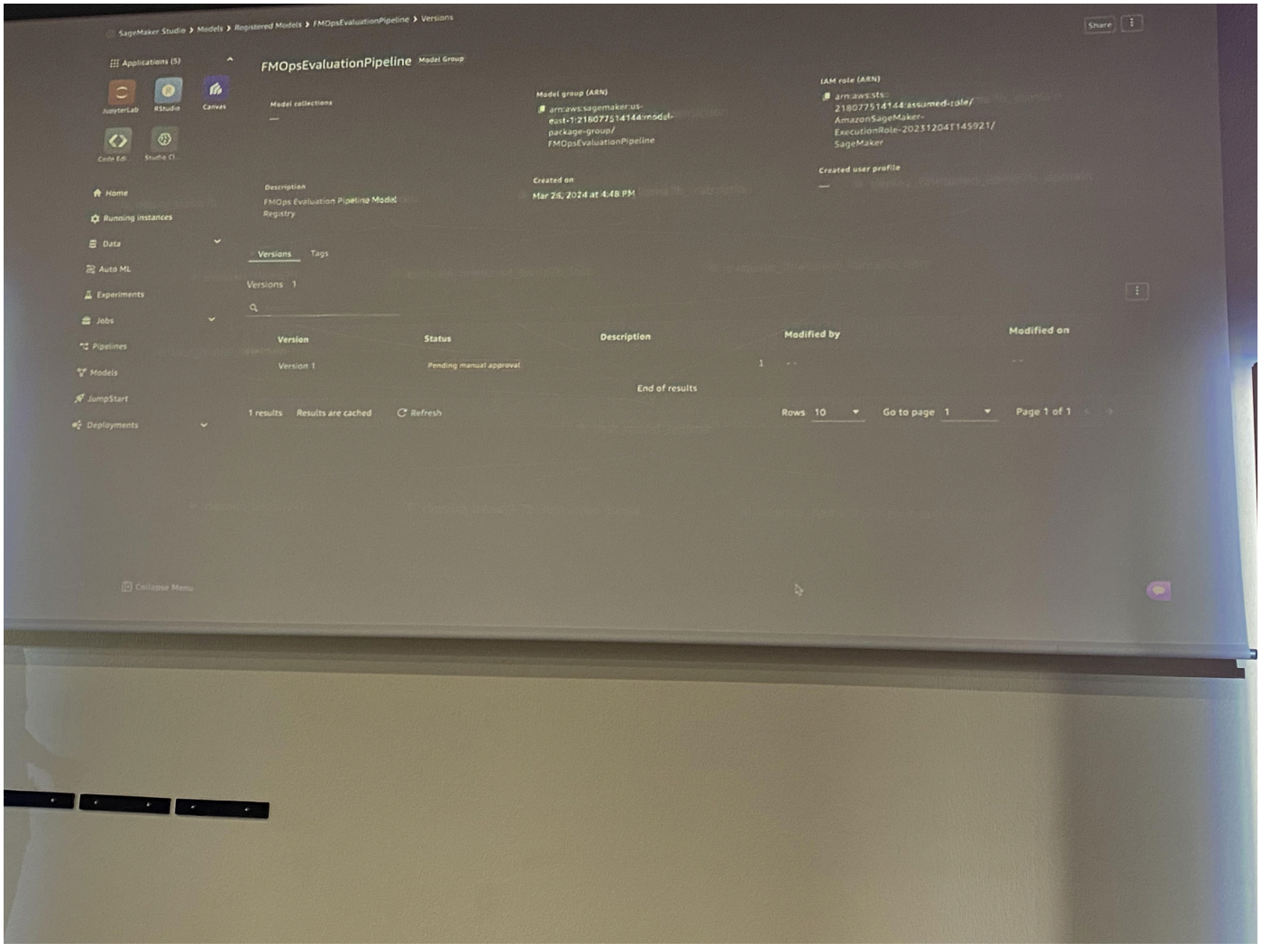
Exemple d'un cycle de vie FMOps

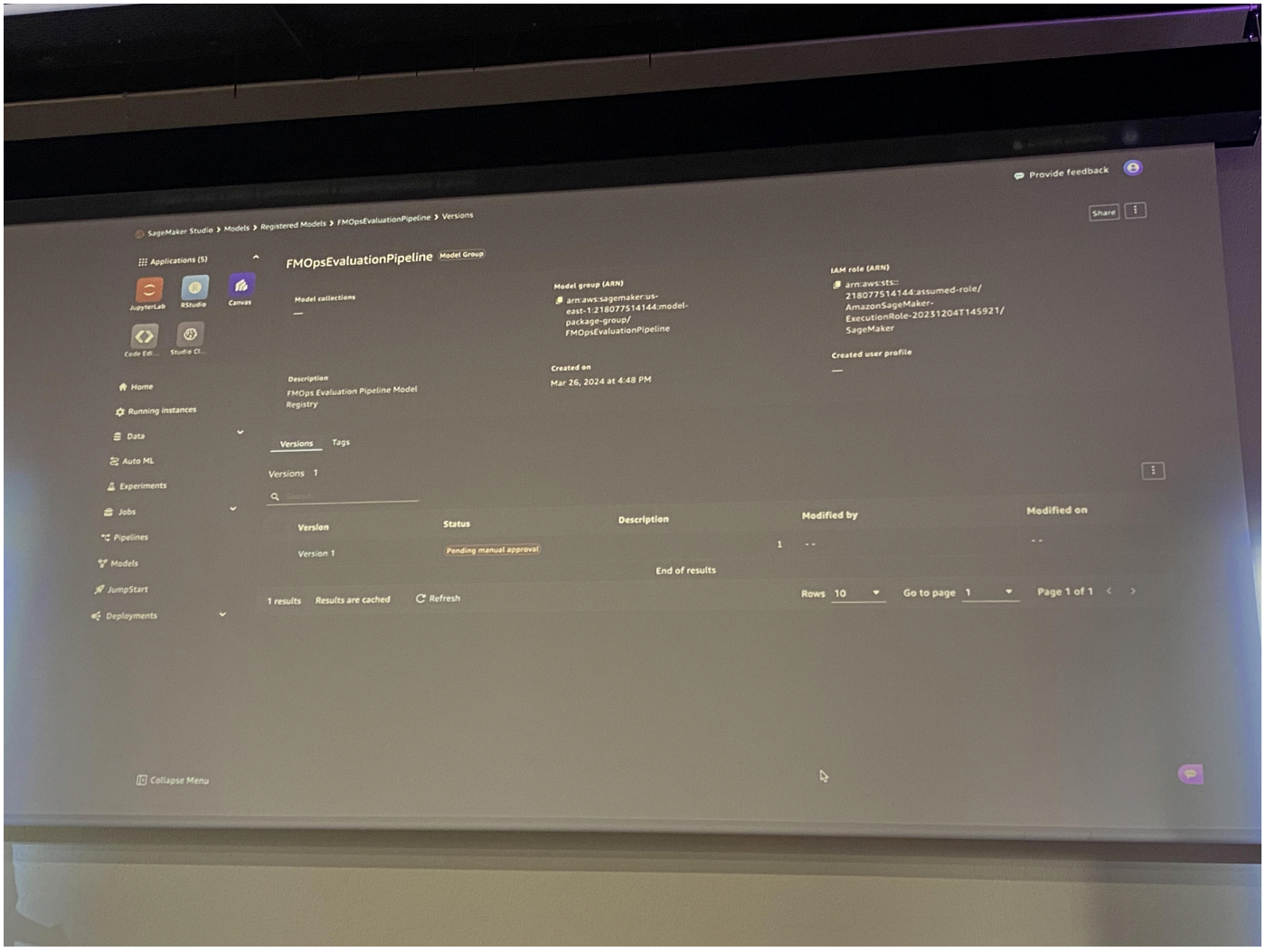


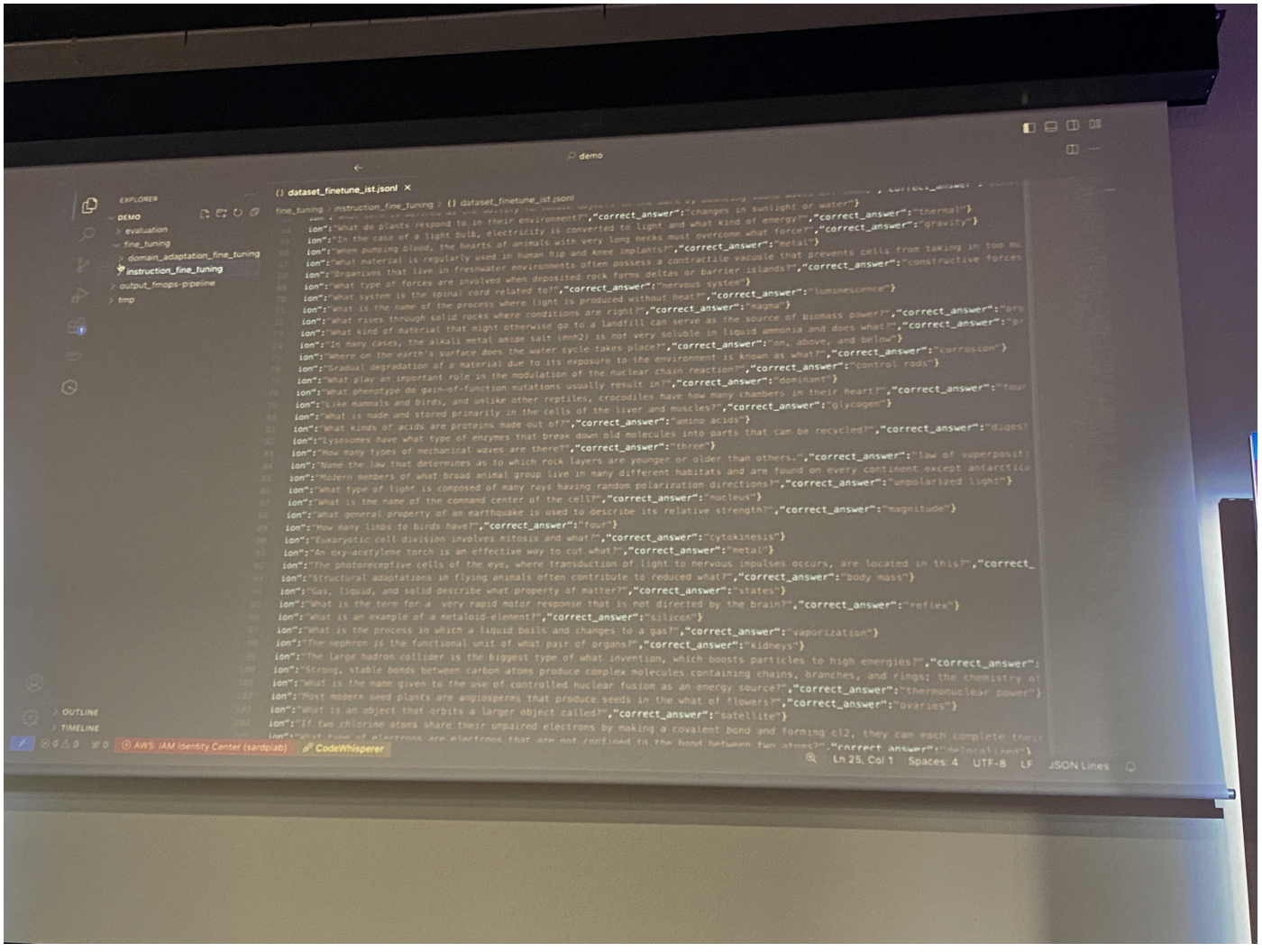

Provide feedback
SageMaker Studio > Pipelines > fmops24 Pipeline > Graph
Applications (5)
JupyterLab
RStudio
Canvas
Code Edi...
Studio Cl...
PIPELINE
fmops24-pipeline
OVERVIEW
Executions
Graph
SETTINGS
Parameters
DETAILS
Information
Graph
Create execution
Select step
Home
Running instances
Data
Auto ML
Experiments
Jobs
Pipelines
Models
JumpStart
Deployments
finetune_llama7b_instruction
finetune_llama7b_domain
deploy_llama7b
preprocess
deploy_finetuned_llama7b_instruction
deploy_finetuned_llama7b_domain
evaluate_llama7b
evaluate_finetuned_llama7b_instr
evaluate_finetuned_llama7b_dom
best_model_selection
best_model_register
cleanup_llama-2-7b
cleanup_llama-2-7b-instruction-tuned
cleanup_llama-2-7b-chat-domain-tuned
Full Screen
93%
Collapse Menu
Privacy
Site Terms
Cookie Preferences
© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.


File Edit View Run Kernel Git Tabs Settings Help
fmops-model-evaluation.ip
Markdown
Cluster
Python 3 (ipykernel)
/ gai214-fr / fmops /
Name
Last Modified
prerequisite
3 days ago
steps
2 days ago
config.yaml
last month
fmops-mode...
yesterday
requirement...
2 days ago
FMOps - Model Evaluation and Selection
As a Data Scientist, you want to identify the right model to use for your application, and to customize it to produce accurate results. You would prefer not to go through the entire steps manually again, one step at a time, every time a new FM becomes available on Sagemaker Jumpstart or a new dataset is available for fine-tuning. Instead, you'd like to automate these steps into a repeatable end-to-end ML workflow that can be executed later either as a user-initiated or an event-triggered workflow.
The goal of this notebook is to provide an implementation of a multi-step SageMaker pipeline that will take care of multiple models evaluation, selection and registration into the SageMaker model registry.
For running this example we will use LLama-2-7b models that will be used with default weights or after a finetuning. All the models will be instantiated and finetuned by using Amazon Sagemaker Jumpstart SDK.
This notebook is also using other Amazon SageMaker components:
SageMaker Pipelines is a purpose-built workflow orchestration service to automate all phases of machine learning (ML) from data pre-processing to model monitoring. With an intuitive UI and Python SDK you can manage repeatable end-to-end ML pipelines at scale. The native integration with multiple AWS services allows you to customize the ML lifecycle based on your MLOps requirements. SageMaker Model Registry
Amazon SageMaker Model Registry is a purpose-built metadata store to manage the entire lifecycle of ML models from training to inference. Whether you prefer to store your model artifacts (model framework files, container image) in AWS (Amazon ECR) or outside of AWS in any third party Docker repository, you can now track them all in Amazon SageMaker Model Registry. You also have the flexibility to register a model without read/write permissions to the associated container image. If you want to track an ML model in a private repository, set the optional 'SkipModelValidation' parameter to 'All' at the time of registration. Later you can also deploy these models for inference in Amazon SageMaker.
Amazon SageMaker Clarify provides purpose-built tools to gain greater insights into your ML models and data, based on metrics such as accuracy, robustness, toxicity, and bias to improve model quality and support responsible AI initiative. With the rise of generative AI, data scientists and ML engineers can leverage publicly available foundation models (FMs) to accelerate speed-to-market. To remove the heavy lifting of evaluating and selecting the right FM for your use case, Amazon SageMaker Clarify supports FM evaluation to help you quickly evaluate, compare, and select the best FM for your use case based on a variety of criteria across different tasks within minutes. It allows you to adopt FMs faster and with confidence. To perform evaluation we are using the open source library FMEval that empowers SageMaker Clarify FM model evaluation.
This example is built by following the best practices explained in the blog post Operationalize LLM Evaluation at Scale using Amazon SageMaker Clarify and MLOps services.
Environment setup
You need to select Data Science 3.0 kernel with ml.t3.medium instance to run this notebook.
Simple
Instance MEM
40%
CodeWhisperer
Cookie Preferences


SageMaker Studio > Pipelines > Fmops Pipeline > Executions > l4mjv7k9t04b > Graph
Provide feedback
Run time
10m 27s
Start time
10 minutes ago
End time
Graph
deploy_llama7b
Overview
Settings
Details
Status
Succeeded
26/03/2024, 08:41
26/03/2024, 08:44
2m 59s
Metrics
No Metrics found
Files
Charts
No Charts found
deploy_llama7b
preprocess
evaluate_llama7b
evaluate_finetune
best_model_selection
cleanup_llama-2
Collapse Menu
Privacy
Site Terms
Cookie Preferences
© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.


SageMaker Studio > Models > Registered Models > FMOpsEvaluationPipeline > Versions
Share
FMOpsEvaluationPipeline
Model Group
Model collections
Model group (ARN)
Created on
Mar 26, 2024 at 4:48 PM
IAM role (ARN)
Created user profile
Description
FMOps Evaluation Pipeline Model Registry
Versions
Tags
Versions 1
Version
Status
Description
Modified by
Modified on
Version 1
Pending manual approval
End of results
1 results
Results are cached
Refresh
Rows 10
Go to page 1
Page 1 of 1
Collapse Menu


SageMaker Studio > Models > Registered Models > FMOpsEvaluationPipeline > Versions
Provide feedback
Share
Applications (5)
JupyterLab
RStudio
Canvas
Code Edi...
Studio Cl...
FMOpsEvaluationPipeline
Model Group
Model collections
—
Model group (ARN)
arn:aws:sagemaker:us-east-1:218077514144:model-package-group/FMOpsEvaluationPipeline
IAM role (ARN)
arn:aws:sts::218077514144:assumed-role/AmazonSageMaker-ExecutionRole-20231204T145921/SageMaker
Created user profile
—
Created on
Mar 26, 2024 at 4:48 PM
Description
FMOps Evaluation Pipeline Model Registry
Home
Running instances
Data
Auto ML
Experiments
Jobs
Pipelines
Models
JumpStart
Deployments
Versions
Tags
Versions 1
Search
Version
Status
Description
Modified by
Modified on
Version 1
Pending manual approval
1
- -
- -
End of results
1 results
Results are cached
Refresh
Rows 10
Go to page 1
Page 1 of 1
Collapse Menu


demo
EXPLORER
DEMO
evaluation
fine_tuning
domain_adaptation_fine_tuning
instruction_fine_tuning
output_fmops-pipeline
tmp
dataset_finetune_ist.jsonl
fine_tuning > instruction_fine_tuning > {} dataset_finetune_ist.jsonl
OUTLINE
TIMELINE
AWS: IAM Identity Center (sardplab)
CodeWhisperer
Ln 25, Col 1
Spaces: 4
UTF-8
LF
JSON Lines

Le parcours des fournisseurs

Processus d'IA générative - Fournisseurs

LE PRÉ-ENTRAINEMENT D'UN FM N'EST PAS UNE MINCE AFFAIRE

Labélisation massive des données

Des annotateurs humains pour labéliser des milliards de données

Processus d'IA générative - Fournisseurs

LE PRÉ-ENTRAINEMENT D'UN FM N'EST PAS UNE MINCE AFFAIRE

Labélisation massive des données

Des annotateurs humains pour labéliser des milliards de données

Des semaines, voire des mois d'entrainement

Distribuez l'entraînement des modèles de fondation

Processus d'IA générative - Fournisseurs

LE PRÉ-ENTRAINEMENT D'UN FM N'EST PAS UNE MINCE AFFAIRE

Labélisation massive des données

Des annotateurs humains pour labéliser des milliards de données

Des semaines, voire des mois d'entrainement

Distribuez l'entraînement des modèles de fondation

Déploiement

Déployez des modèles FM propriétaires ou open source, les artefacts et le code

Processus d'IA générative - Fournisseurs

LE PRÉ-ENTRAINEMENT D'UN FM N'EST PAS UNE MINCE AFFAIRE

Labélisation massive des données

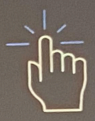
Des annotateurs humains pour labéliser des milliards de données

Des semaines, voire des mois d'entrainement

Distribuez l'entraînement des modèles de fondation

Déploiement

Déployez des modèles FM propriétaires ou open source, les artefacts et le code



Classements

Les FM sont évalués par des organisations tierces

Création de FM à l'aide d'Amazon SageMaker HyperPod

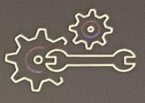
DÉVELOPPEZ ET ENTRAÎNEZ LES FM EN CONTINU PENDANT DES SEMAINES ET DES MOIS

Environnement résilient

Les clusters auto-réparables réduisent la durée de l'entraînement jusqu'à 20 %

Simplifiez l' entrainement distribué

Les bibliothèques d'entrainement distribuées de SageMaker améliorent les performances jusqu'à 20 %



Utilisation optimisée des ressources

Contrôle de l'environnement informatique et planification de la charge de travail

Création de FM à l'aide d'Amazon SageMaker HyperPod

DÉVELOPPEZ ET ENTRAÎNEZ LES FM EN CONTINU PENDANT DES SEMAINES ET DES MOIS

Environnement résilient

Les clusters auto-réparables réduisent la durée de l'entraînement jusqu'à 20 %

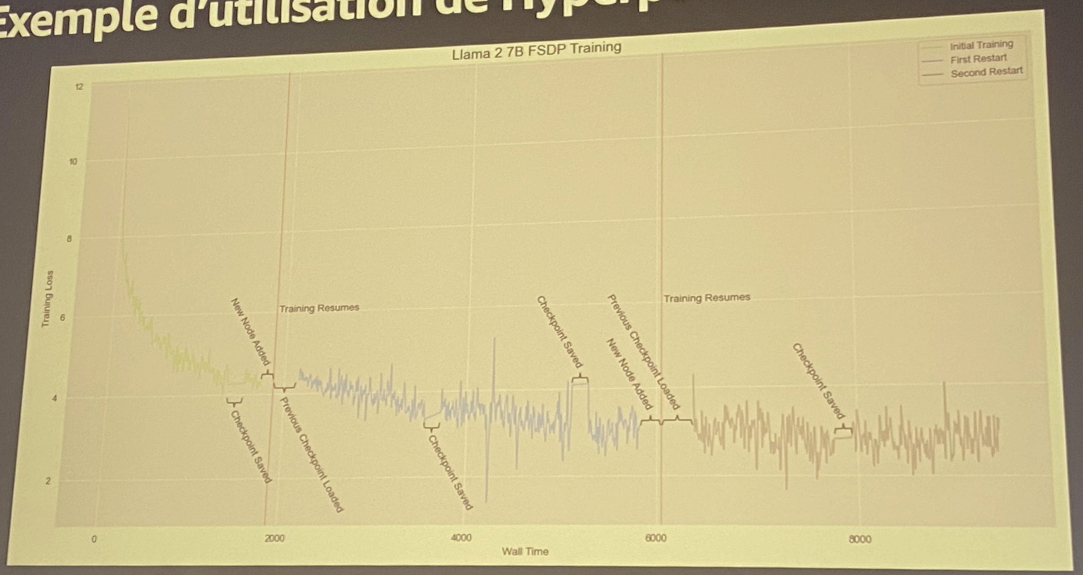
Simplifiez l' entrainement distribué

Les bibliothèques d'entrainement distribuées de SageMaker améliorent les performances jusqu'à 20 %

Utilisation optimisée des ressources

Contrôle de l'environnement informatique et planification de la charge de travail

Exemple d'utilisation de Hyperpod



Artificial failures injected via DCGM Error Injection
https://docs.nvidia.com/datacenter/dcgm/latest/user-guide/dcgm-error-injection.html

Conclusion et ressources

En résumé

- MLOps et FMOps : même objectif, c.à.d. mettre en place des solutions fiables, performantes et réutilisables
- On conserve les bonnes pratiques classiques de software et le Well-Architected Framework pour les composants de l'application d'IA Générative
- Pratiques similaires DevOps et MLOps pour les optimisations du déploiement même si nuances dans l'implémentation
- MAIS **différences majeures** aussi de par la nature générative et l'utilisation de FM lors des phases d'expérimentation, de développement et de production

Commencez dès aujourd'hui !!!

1. Déterminez le bon cas d'usage
2. Prenez en main Amazon Bedrock et itérez itérez itérez !
3. Contactez votre équipe de compte pour une nomination au Generative AI Innovation Center

Ressources sur le FMOps

- Lien vers le Gen AI Innovation Center
  https://aws.amazon.com/fr/generative-ai/innovation-center/?th=tile&tile=start-your-ai-journey&p=2
- FMOPS/LLMops : opérationnalisez l'IA générative et ses différences avec le MLOps
  https://aws.amazon.com/blogs/machine-learning/fmops-llmops-operationalize-generative-ai-and-differences-with-mlops
- Opérationnalisez l'évaluation du LLM à grande échelle à l'aide d'Amazon SageMaker Clarify et des services MLOps
  https://aws.amazon.com/blogs/machine-learning/operationalize-llm-evaluation-at-scale-using-amazon-sagemaker-clarify-and-mlops-services
- Créez un service SaaS interne avec suivi des coûts et de l'utilisation pour les modèles de base sur Amazon Bedrock
  https://aws.amazon.com/blogs/machine-learning/build-an-internal-saas-service-with-cost-and-usage-tracking-for-foundation-models-on-amazon-bedrock/


Merci !
Ségolène Dessertine-Panhard (elle/elle)
linkedin.com/in/dr-segolène-dessertine-panhard-41416010/
Patrick Sard (il/lui)
linkedin.com/in/patrick-sard/
Veuillez remplir le questionnaire de session dans l'application mobile
© 2024, Amazon Web Services, Inc. ou ses filiales. Tous droits réservés.
aws SUMMIT