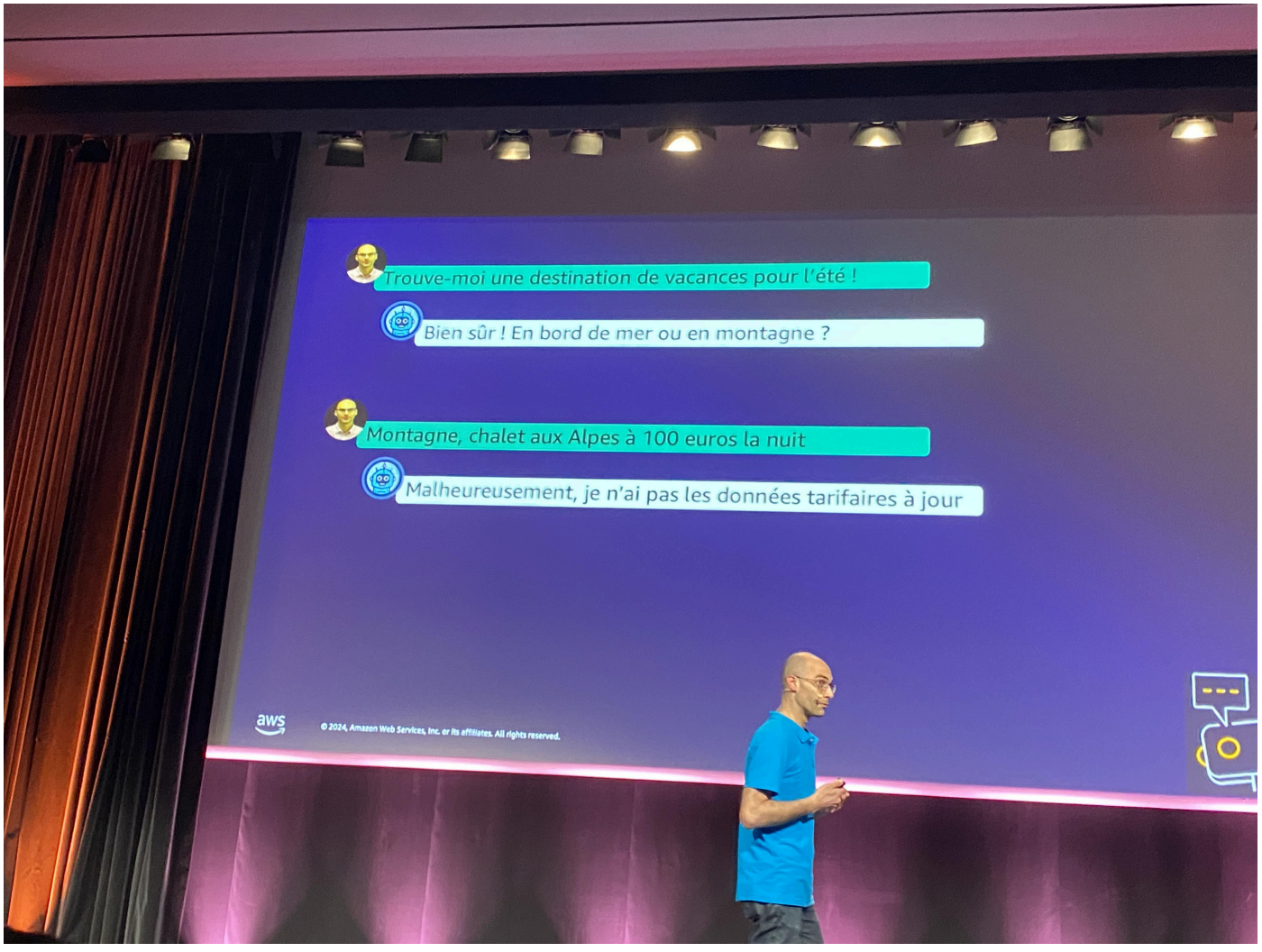
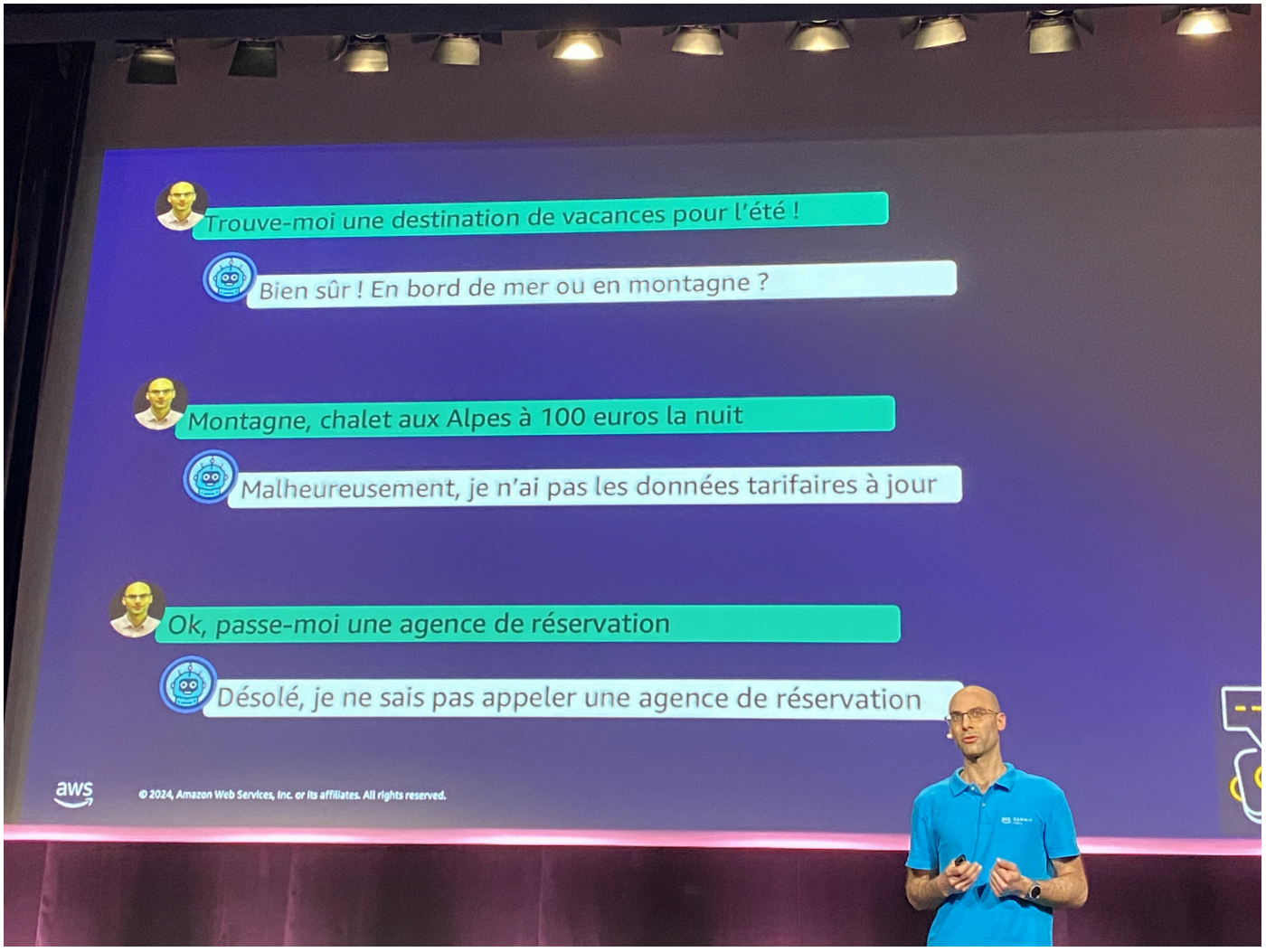

aws
SUMMIT
PARIS


Paris – 3 avril 2024
aws
SUMMIT
Améliorez vos applications d'IA générative
avec RAG sur Amazon Bedrock
aws
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.


Trouve-moi une destination de vacances pour l'été !
Bien sûr ! En bord de mer ou en montagne ?
aws
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.


Trouve-moi une destination de vacances pour l'été !
Bien sûr ! En bord de mer ou en montagne ?
Montagne, chalet aux Alpes à 100 euros la nuit
Malheureusement, je n'ai pas les données tarifaires à jour
aws
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.


Trouve-moi une destination de vacances pour l'été !
Bien sûr ! En bord de mer ou en montagne ?
Montagne, chalet aux Alpes à 100 euros la nuit
Malheureusement, je n'ai pas les données tarifaires à jour
Ok, passe-moi une agence de réservation
Désolé, je ne sais pas appeler une agence de réservation
aws
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.


SESSION GAI302
Améliorez vos applications d'IA générative avec RAG sur Amazon Bedrock
Hugues Gendre
Chief Information Officer
UCPA
Merieme Ezzaouia
Solutions Architect
AWS
Nizar Kheir
Senior Solutions Architect
AWS
aws
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

SESSION GAI302

Améliorez vos applications d'IA générative avec RAG sur Amazon Bedrock

Hugues Gendre
Chief Information Officer
UCPA

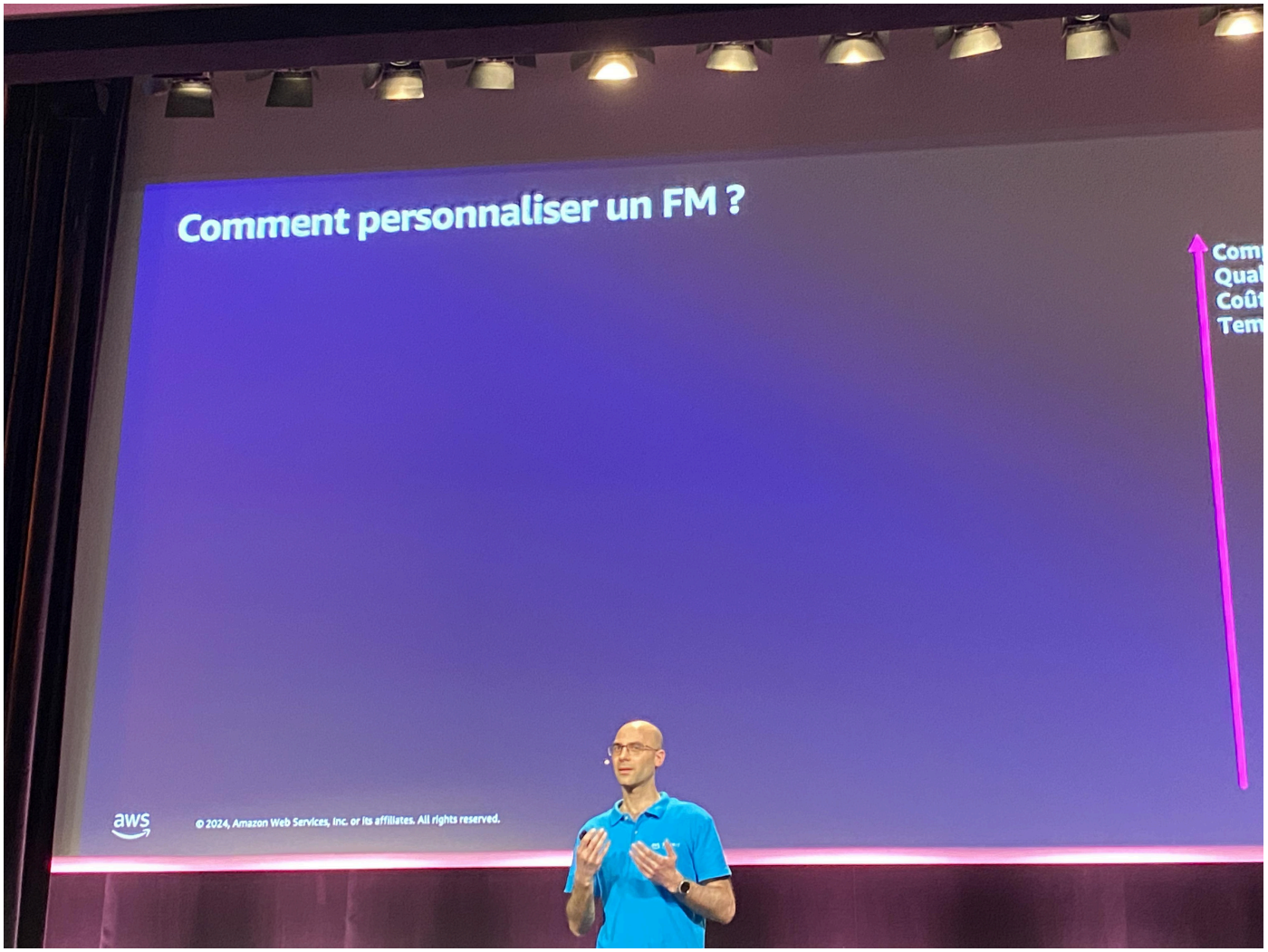
Merieme Ezzaouia
Solutions Architect
AWS

Nizar Kheir
Senior Solutions Architect
AWS

Pourquoi personnaliser un *Foundation Model* (FM) ?

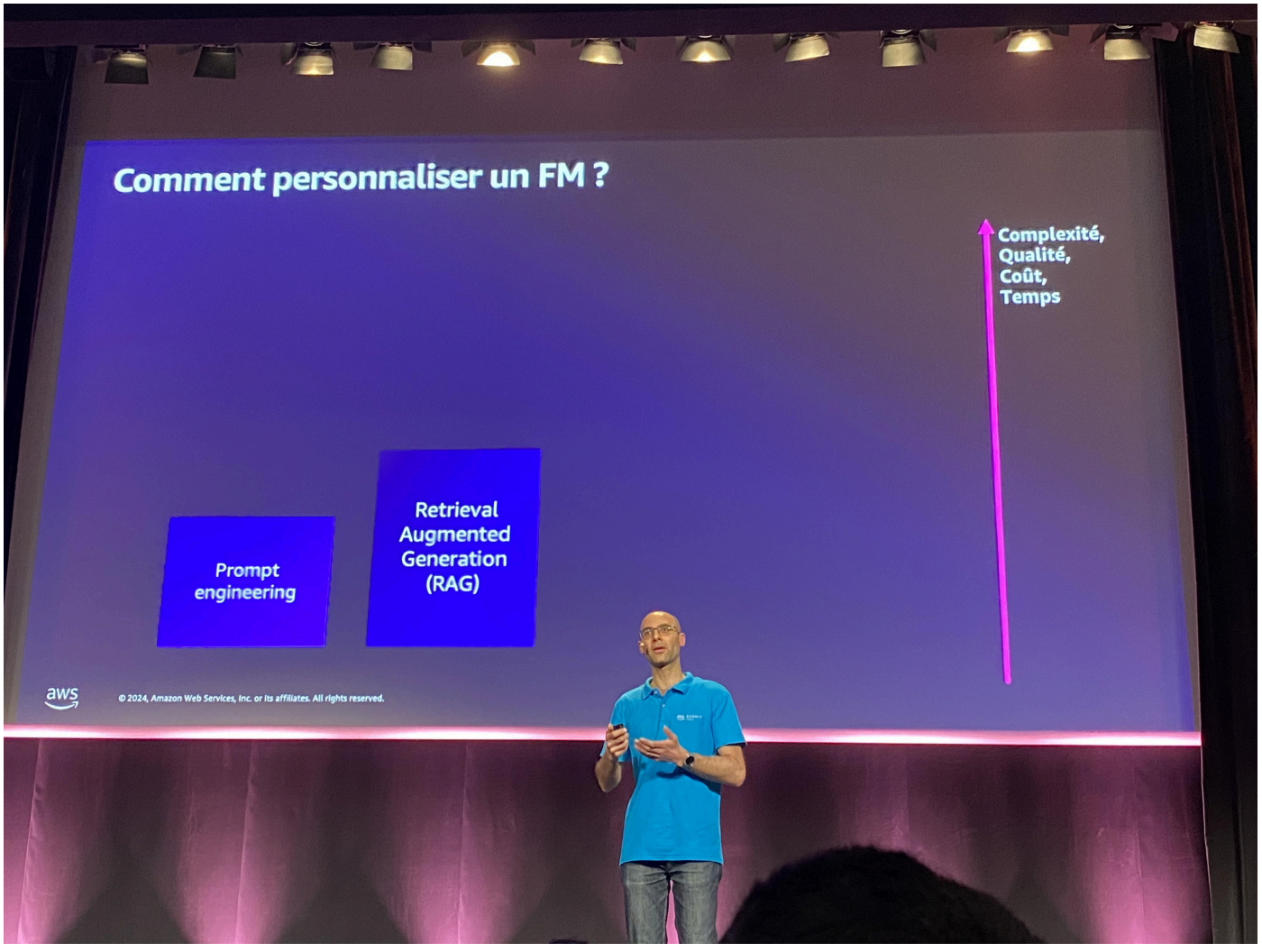
"Là où il y a de la fumée de données, il y a un feu d'opportunités pour les entreprises"

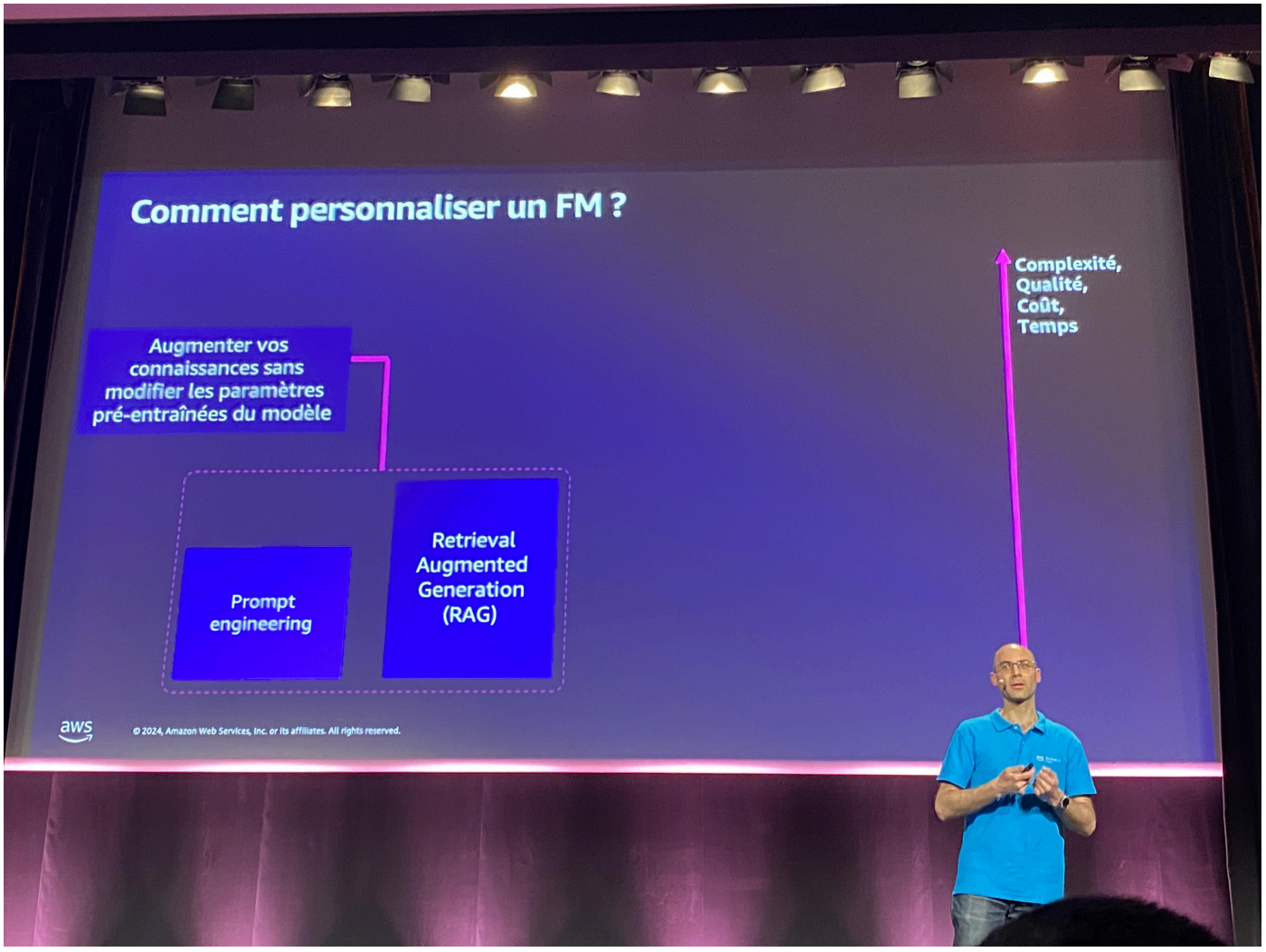
T. Redman

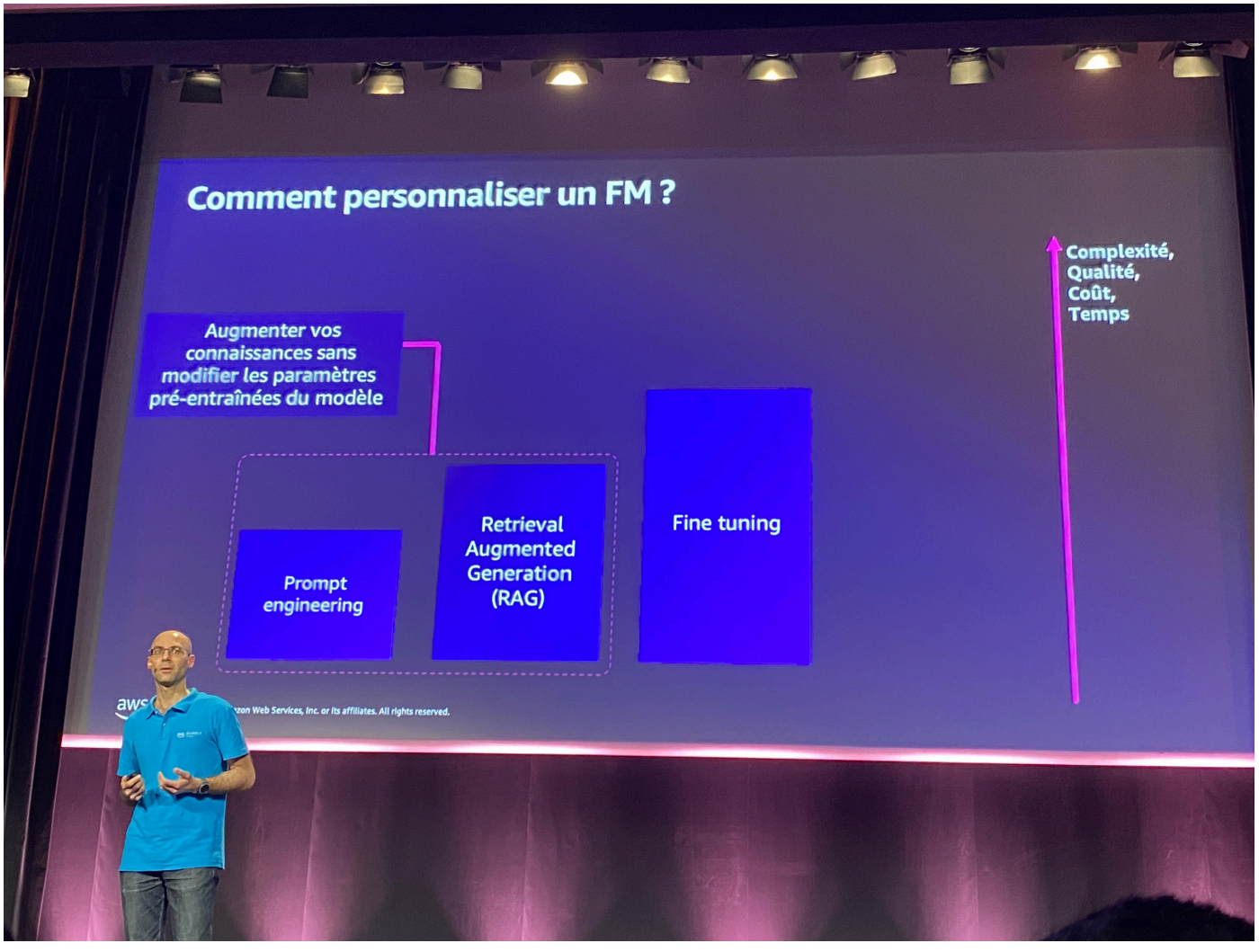

Comment personnaliser un FM ?
Comp
Qual
Coût
Tem
aws
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

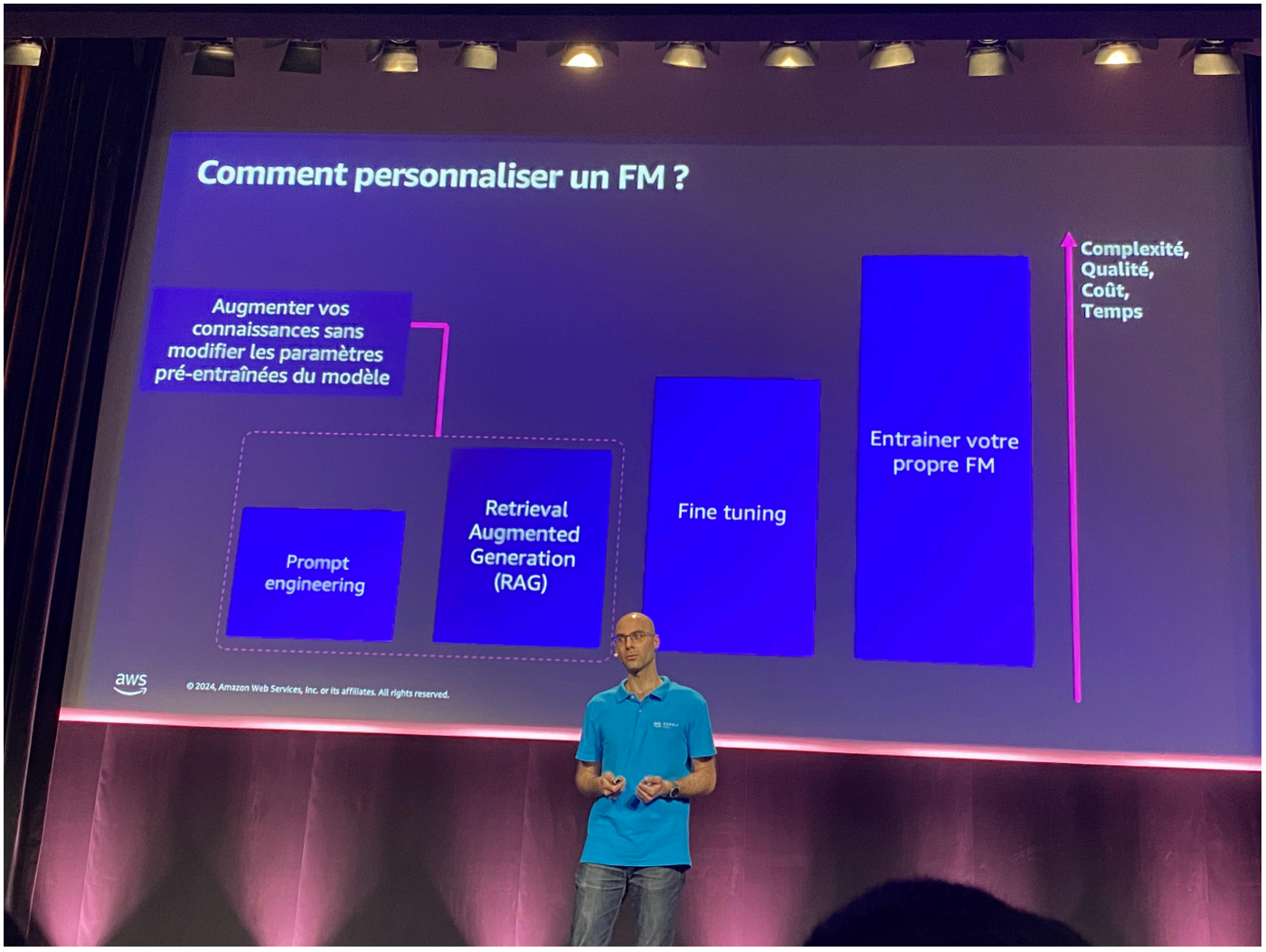

Comment personnaliser un FM ?
Complexité,
Qualité,
Coût,
Temps
Prompt engineering
aws
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.


Comment personnaliser un FM ?
Complexité,
Qualité,
Coût,
Temps
Prompt engineering
aws
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

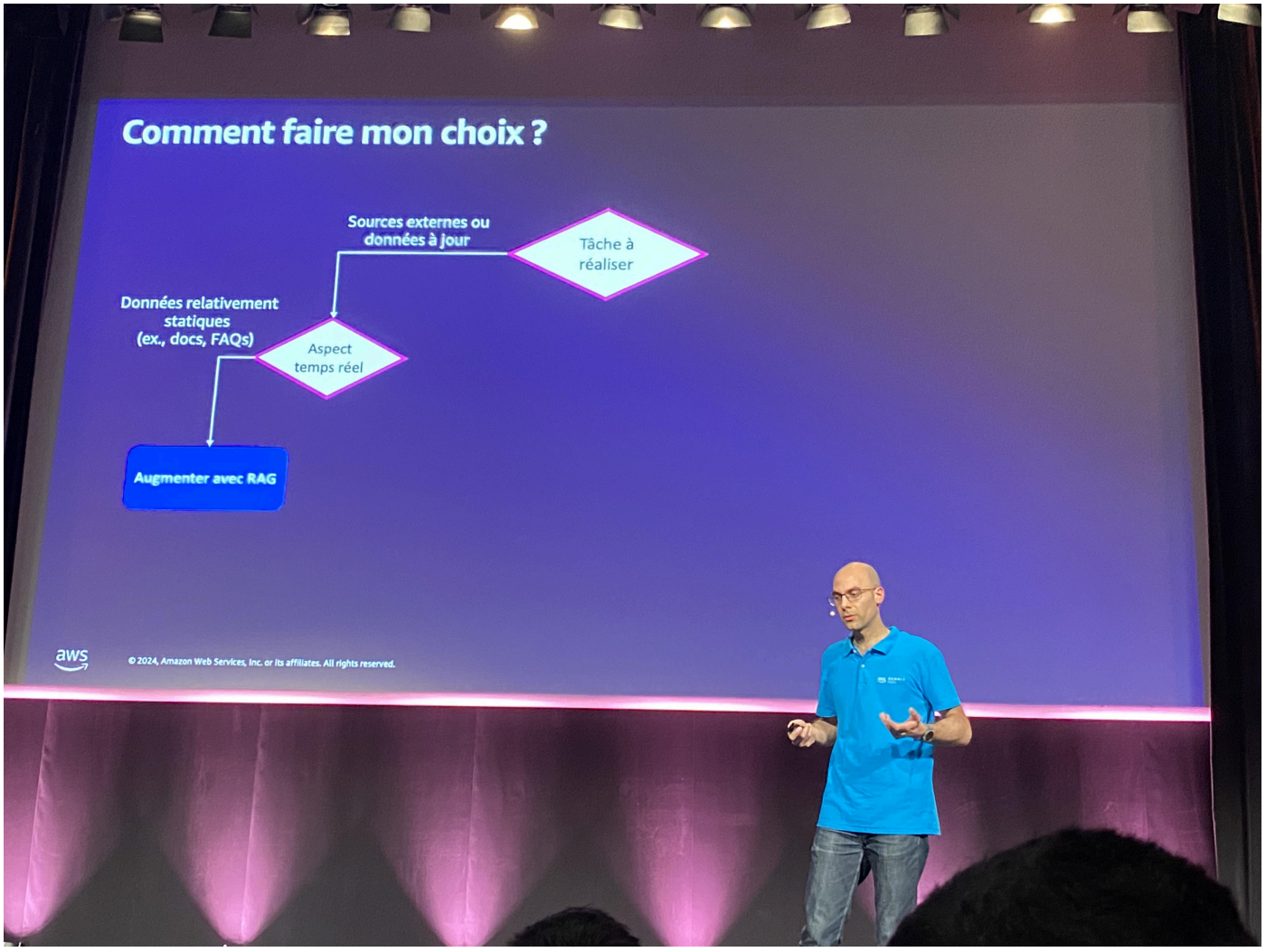

Comment personnaliser un FM ?
Complexité,
Qualité,
Coût,
Temps
Prompt engineering
Retrieval Augmented Generation (RAG)
aws
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

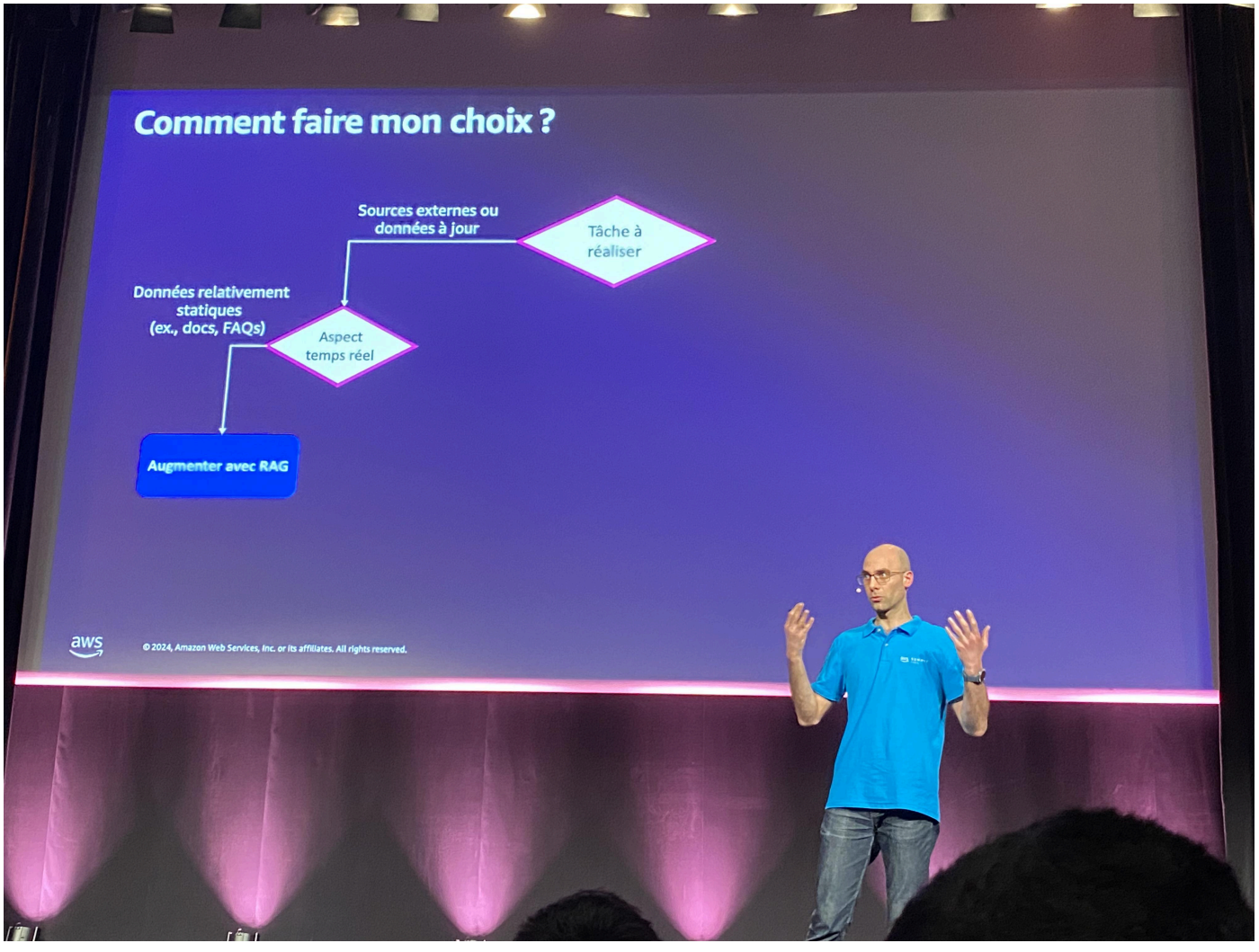

Comment personnaliser un FM ?
Complexité,
Qualité,
Coût,
Temps
Augmenter vos connaissances sans modifier les paramètres pré-entraînées du modèle
Prompt engineering
Retrieval Augmented Generation (RAG)
aws
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.


Comment personnaliser un FM ?
Complexité,
Qualité,
Coût,
Temps
Augmenter vos connaissances sans modifier les paramètres pré-entraînées du modèle
Prompt engineering
Retrieval Augmented Generation (RAG)
Fine tuning
aws
zon Web Services, Inc. or its affiliates. All rights reserved.

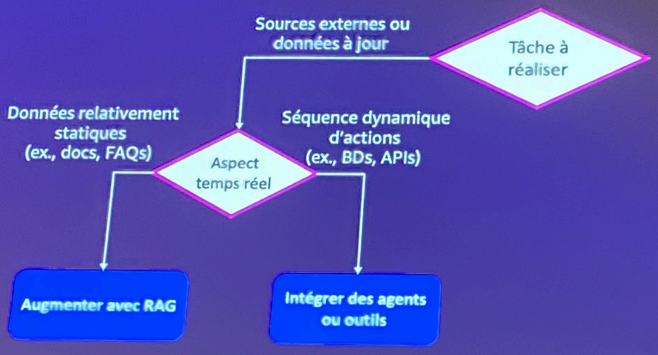

Comment personnaliser un FM ?
Augmenter vos connaissances sans modifier les paramètres pré-entraînées du modèle
Prompt engineering
Retrieval Augmented Generation (RAG)
Fine tuning
Entrainer votre propre FM
Complexité, Qualité, Coût, Temps
aws
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.


Comment faire mon choix ?
Tâche à réaliser
aws
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

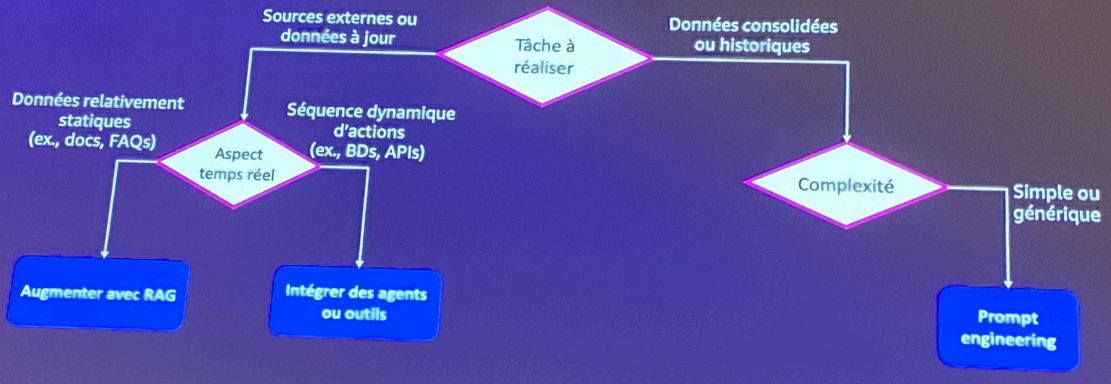

Comment faire mon choix ?
Sources externes ou données à jour
Tâche à réaliser
Données relativement statiques (ex., docs, FAQs)
Aspect temps réel
Augmenter avec RAG
aws
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.


Comment faire mon choix ?
Sources externes ou données à jour
Tâche à réaliser
Données relativement statiques (ex., docs, FAQs)
Aspect temps réel
Augmenter avec RAG
aws
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

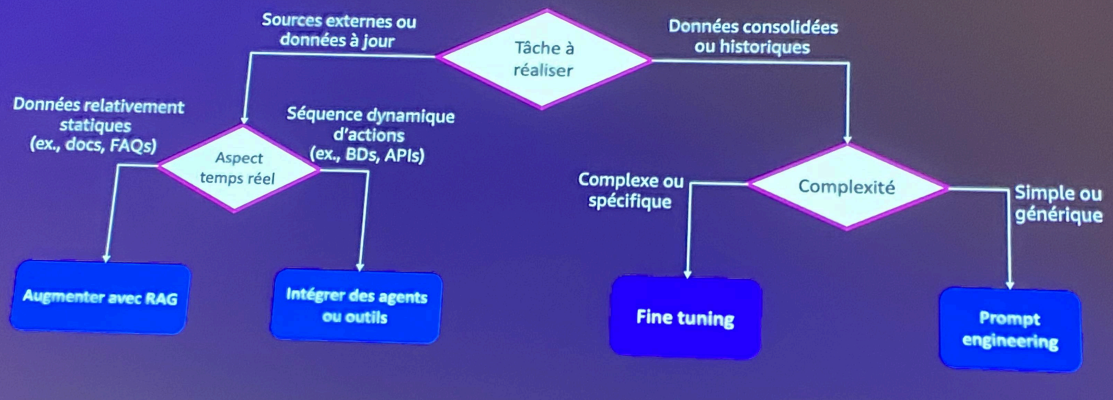
Comment faire mon choix ?



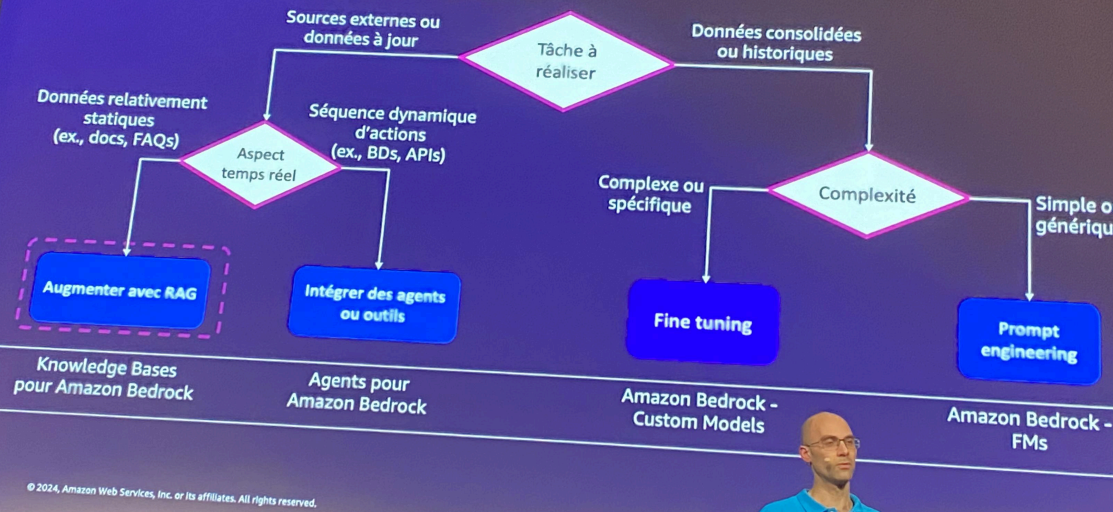
Comment faire mon choix ?

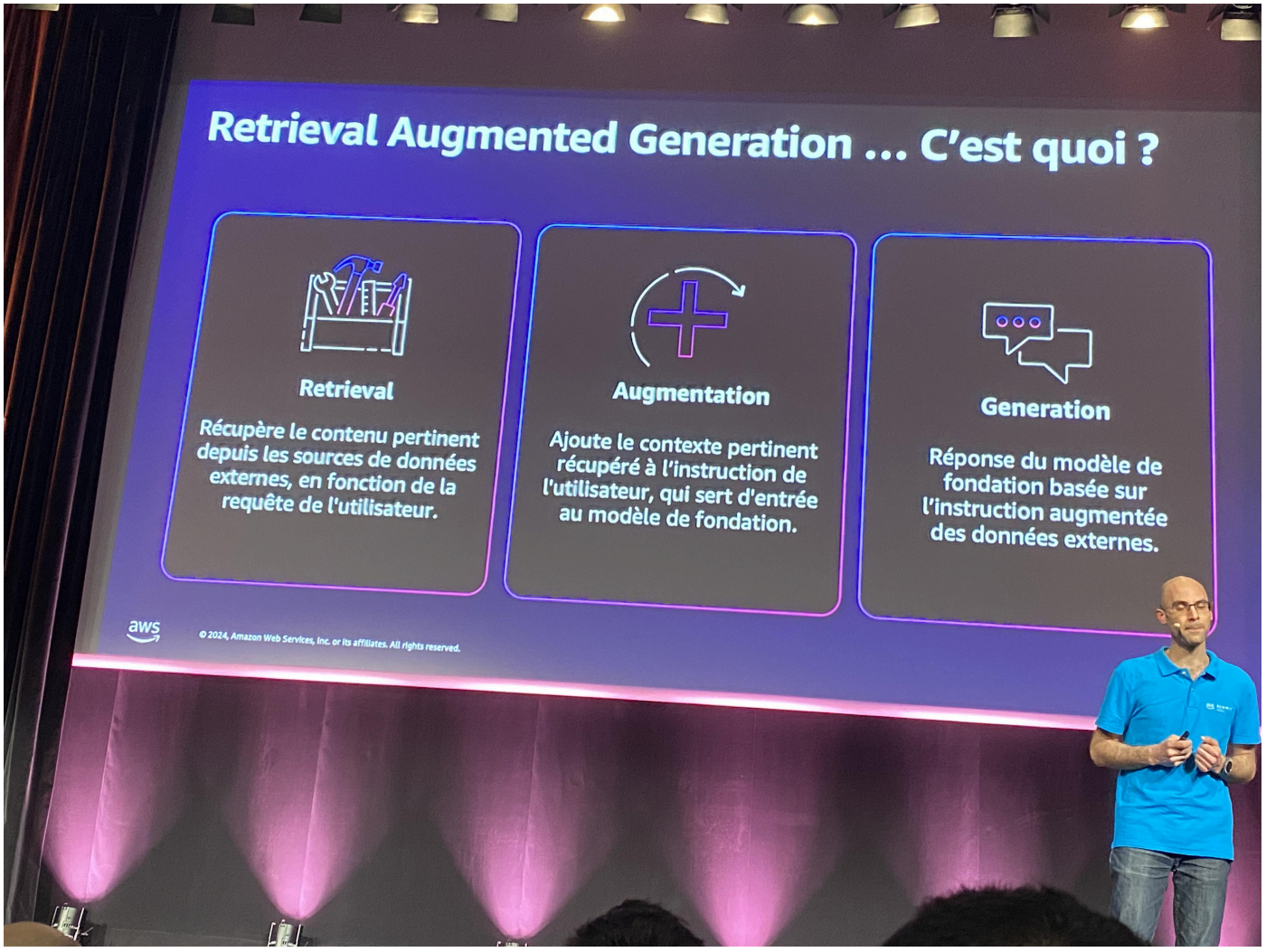


Comment faire mon choix ?



Comment faire mon choix ?




Retrieval Augmented Generation … C'est quoi ?
Retrieval
Récupère le contenu pertinent depuis les sources de données externes, en fonction de la requête de l'utilisateur.
Augmentation
Ajoute le contexte pertinent récupéré à l'instruction de l'utilisateur, qui sert d'entrée au modèle de fondation.
Generation
Réponse du modèle de fondation basée sur l'instruction augmentée des données externes.
aws
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.


Cas d'utilisation du RAG
Qualité des connaissances
Réponses contextualisées
Recherche personnalisée
Synthèse en temps reel
aws
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Types de recherche (*retrieval*)

Basé sur des règles

Données non structurées, supports documentaires

ex., recherche par mots clés

Données structurées

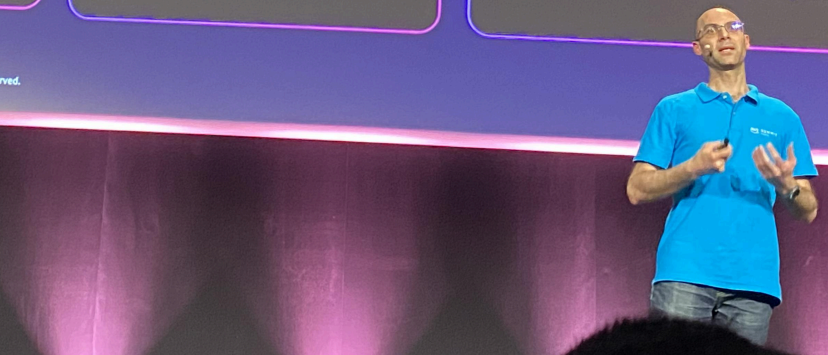
Bases de données transactionnelles, APIs externes

ex., Select customers from All_orders where order == 'XYZ'

Recherche sémantique

Analyse sémantique à travers des modèles de text embeddings

Paris → Metro, Tour Eiffel, Champs Elysées



Types de recherche (*retrieval*)

Basé sur des règles

Données non structurées, supports documentaires

ex., recherche par mots clés

Données structurées

Bases de données transactionnelles, APIs externes

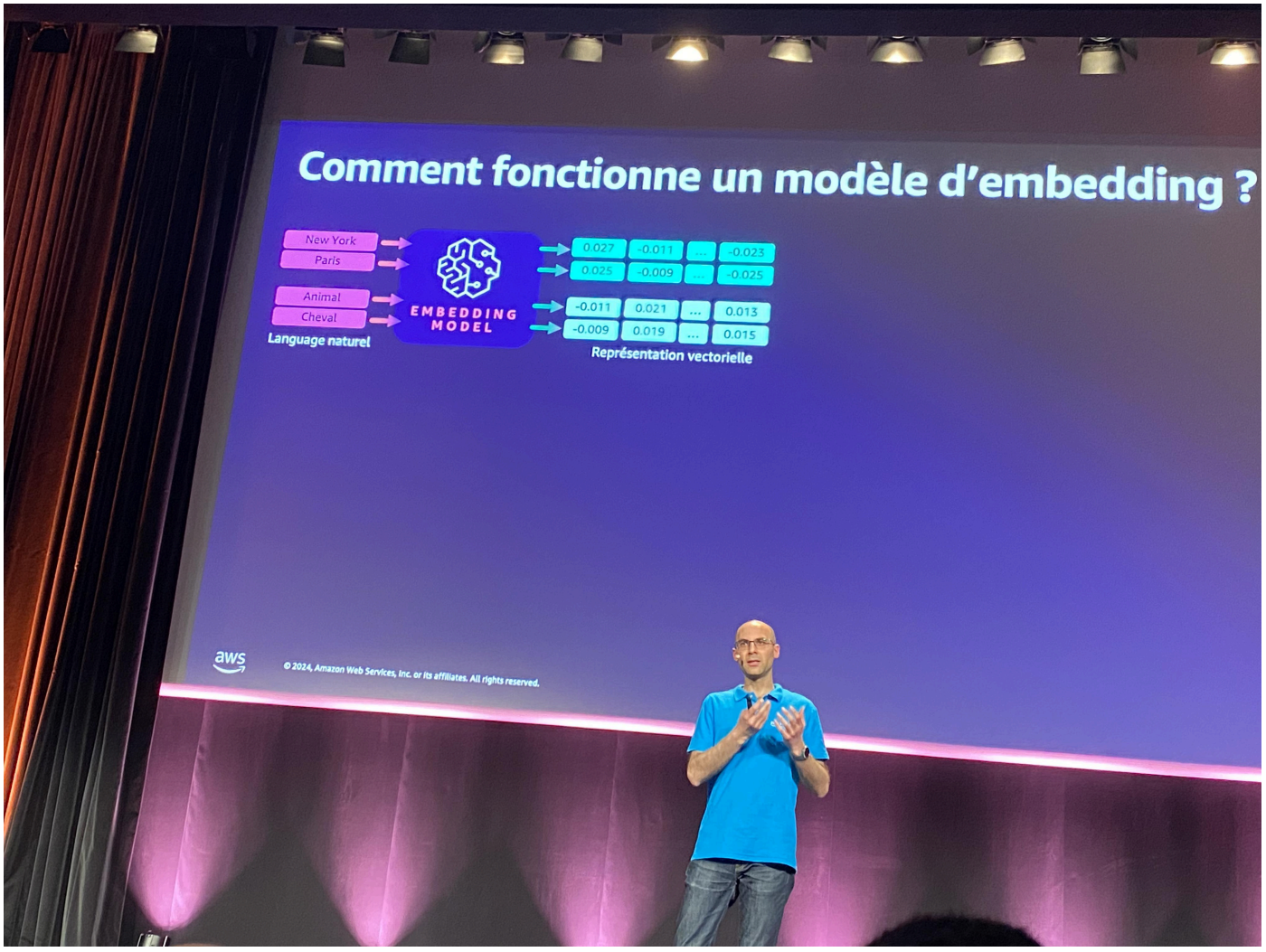
ex., Select customers from All_orders where order == 'XYZ'

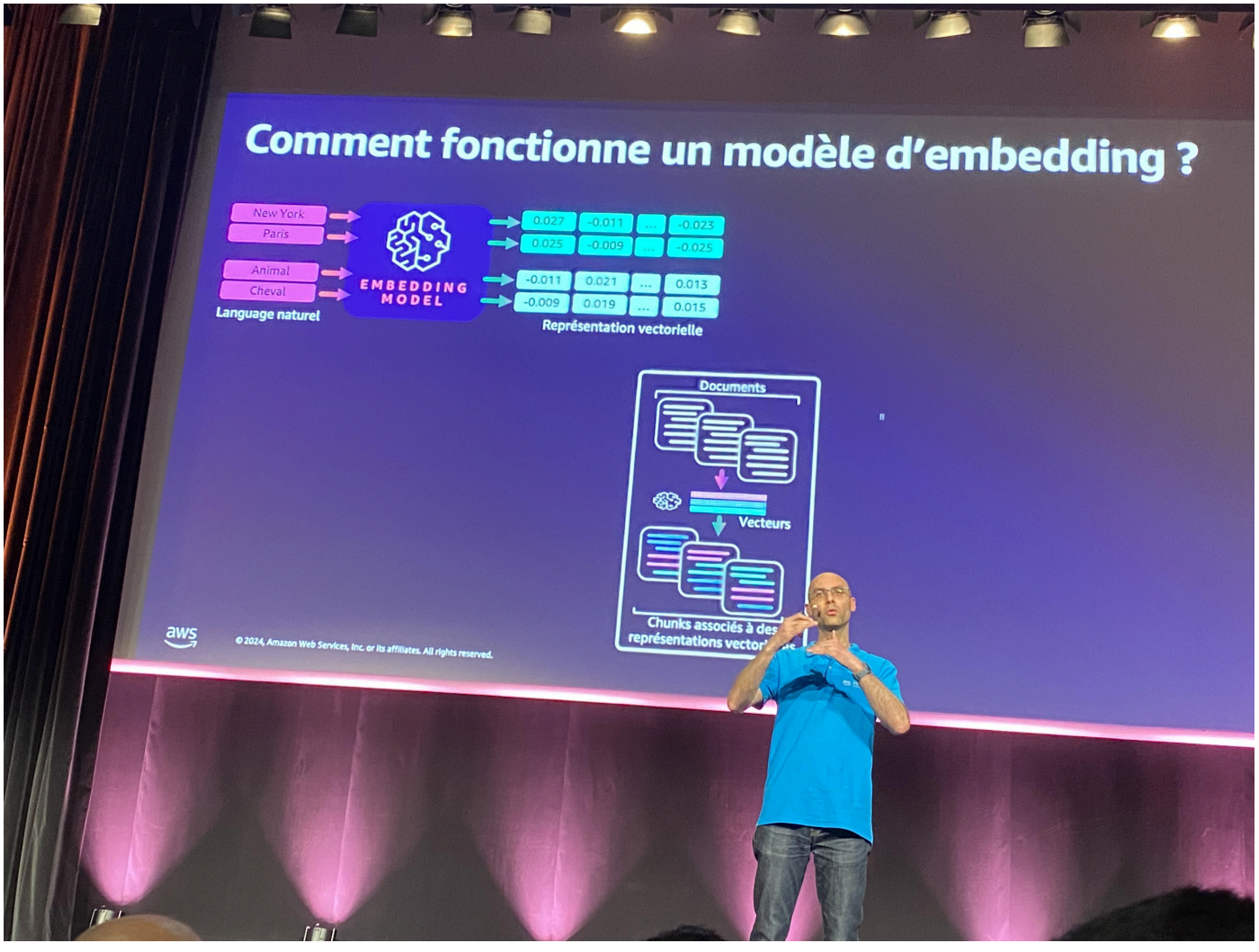
Recherche sémantique

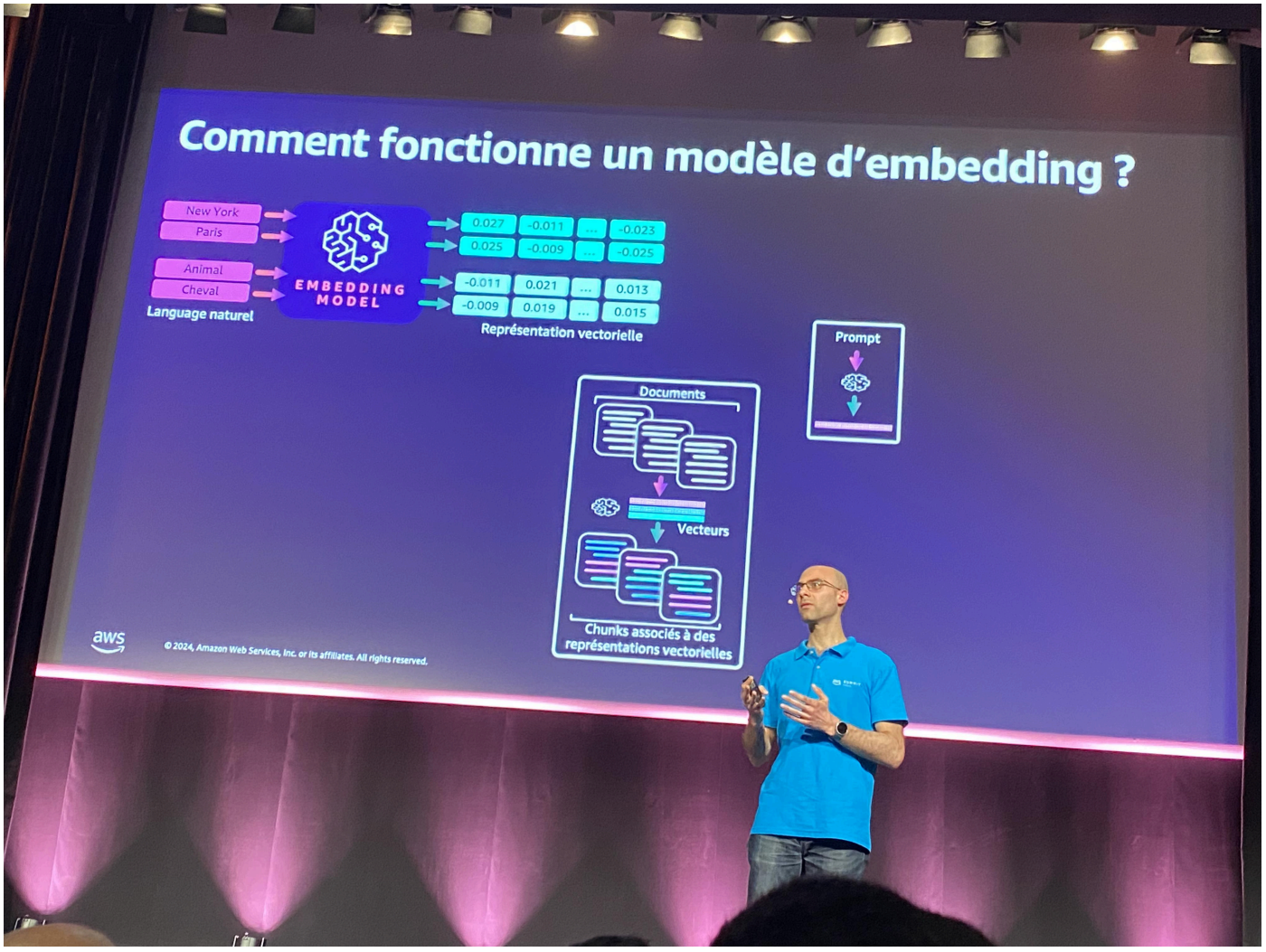
Analyse sémantique à travers des modèles de text embeddings

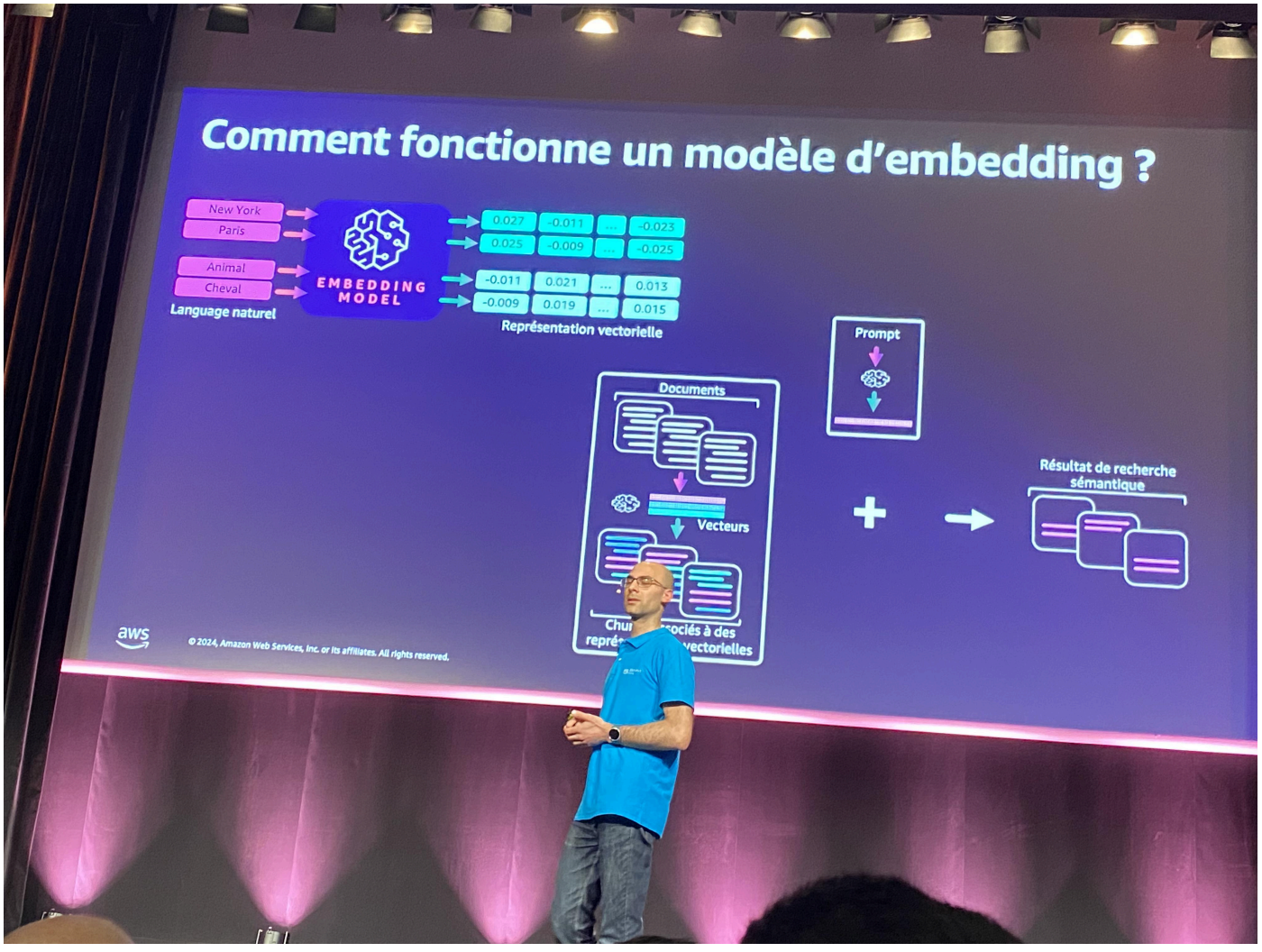
Paris → Metro, Tour Eiffel, Champs Elysées

aws © 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.


Comment fonctionne un modèle d'embedding ?
New York
Paris
Animal
Cheval
Language naturel
EMBEDDING MODEL
0.027 -0.011 ... -0.023
0.025 -0.009 ... -0.025
-0.011 0.021 ... 0.013
-0.009 0.019 ... 0.015
Représentation vectorielle
aws
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.


Comment fonctionne un modèle d'embedding ?
New York
Paris
Animal
Cheval
Language naturel
EMBEDDING MODEL
0.027 -0.011 ... -0.023
0.025 -0.009 ... -0.025
-0.011 0.021 ... 0.013
-0.009 0.019 ... 0.015
Représentation vectorielle
Documents
Vecteurs
Chunks associés à des
aws
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.


Comment fonctionne un modèle d'embedding ?
New York
Paris
Animal
Cheval
Language naturel
EMBEDDING MODEL
0.027 -0.011 ... -0.023
0.025 -0.009 ... -0.025
-0.011 0.021 ... 0.013
-0.009 0.019 ... 0.015
Représentation vectorielle
Prompt
Documents
Vecteurs
Chunks associés à des représentations vectorielles
aws
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.


Comment fonctionne un modèle d'embedding ?
New York
Paris
Animal
Cheval
Language naturel
EMBEDDING MODEL
0.027 -0.011 ... -0.023
0.025 -0.009 ... -0.025
-0.011 0.021 ... 0.013
-0.009 0.019 ... 0.015
Représentation vectorielle
Prompt
Documents
Vecteurs
Résultat de recherche sémantique
aws
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Titan text embeddings model

Amazon Titan Text Embeddings V2.0

Traduit les entrées textuelles (mots, phrases) en représentations numériques (embeddings). Comparer les embeddings produit des réponses contextuelles plus pertinentes que la correspondance de mots.

Max Tokens: 8,000
Output Vectors: 1,536
Language: Multilingual (25 languages)

Model ID: amazon.titan-embed-g1-text-02

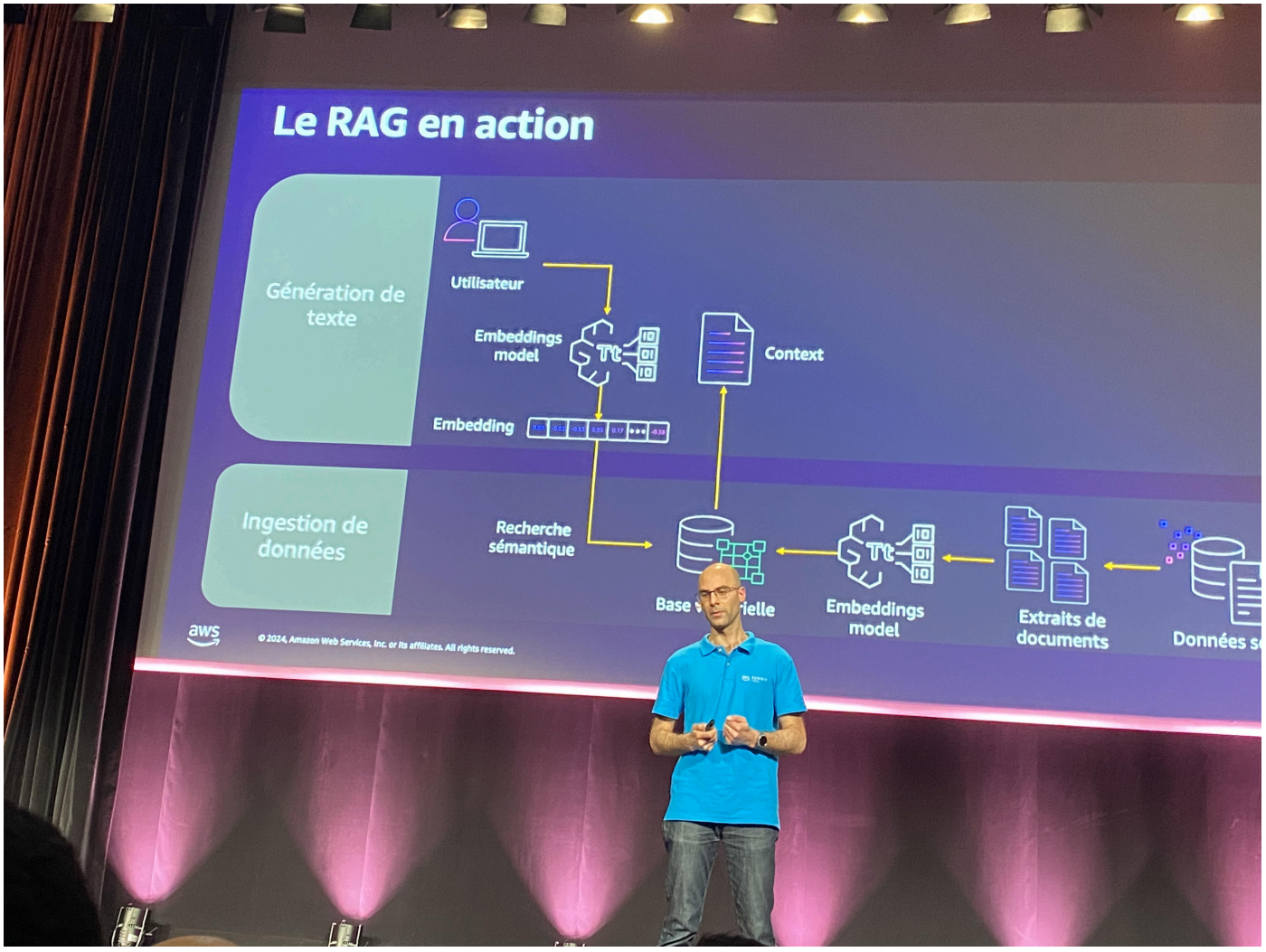
Points forts

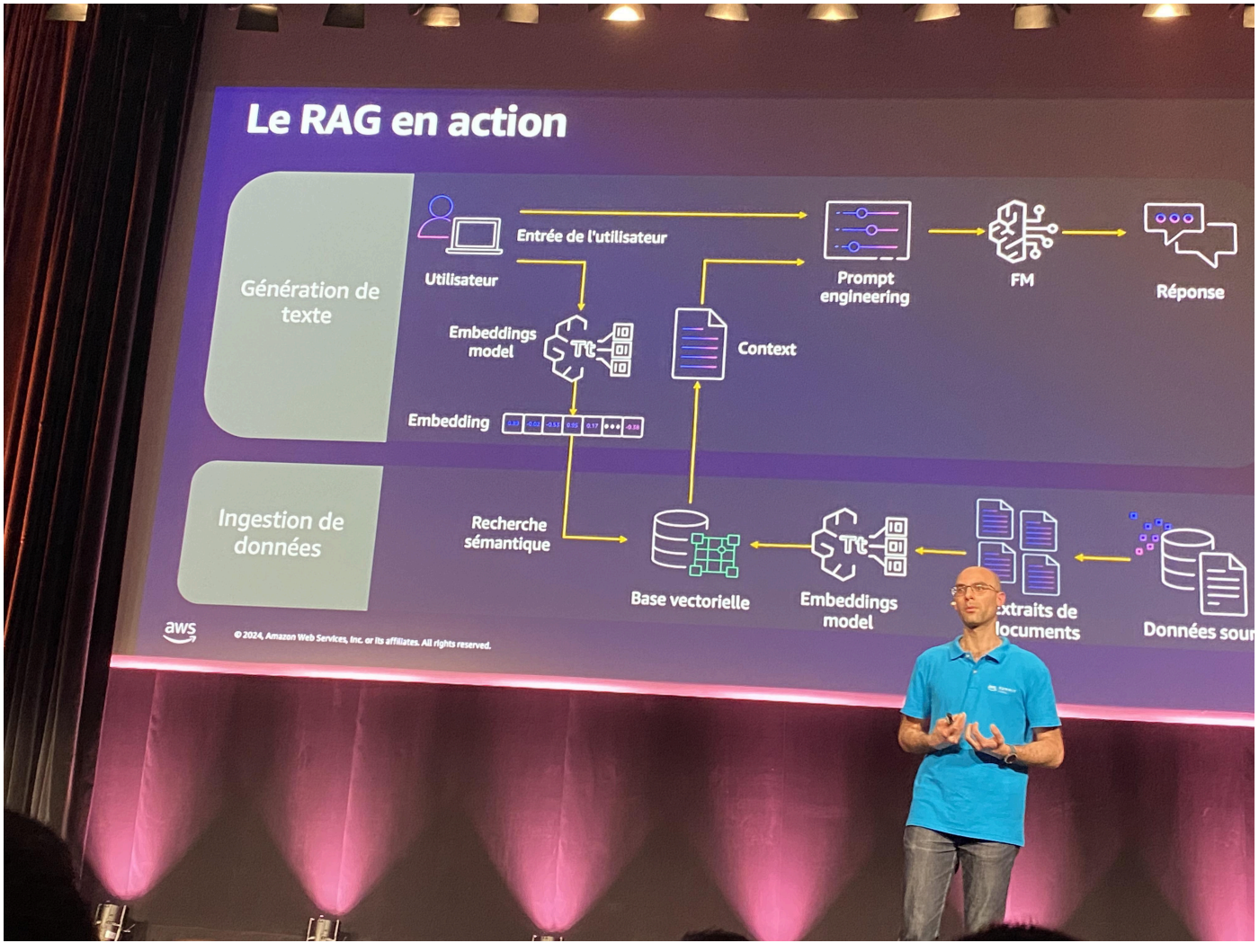
- Titan Text Embeddings propose des embeddings rapides, efficientes, de haute performance, et précis dans 25 langues.
- Optimisé pour les tâches de récupération de texte, de similarité sémantique et de clustering.
- Les applications de ce modèle incluent la recherche sémantique et la personnalisation

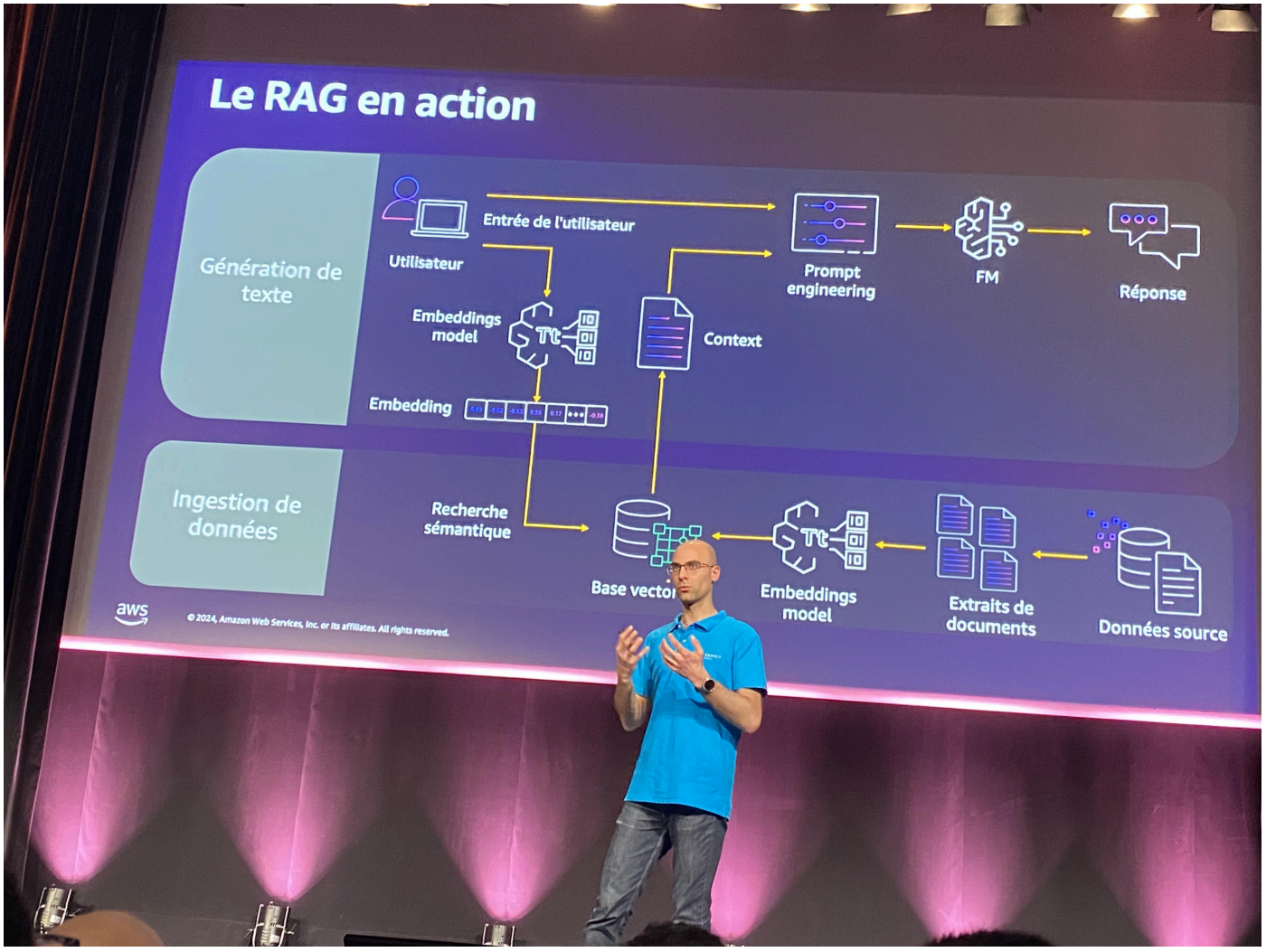
Le RAG en action

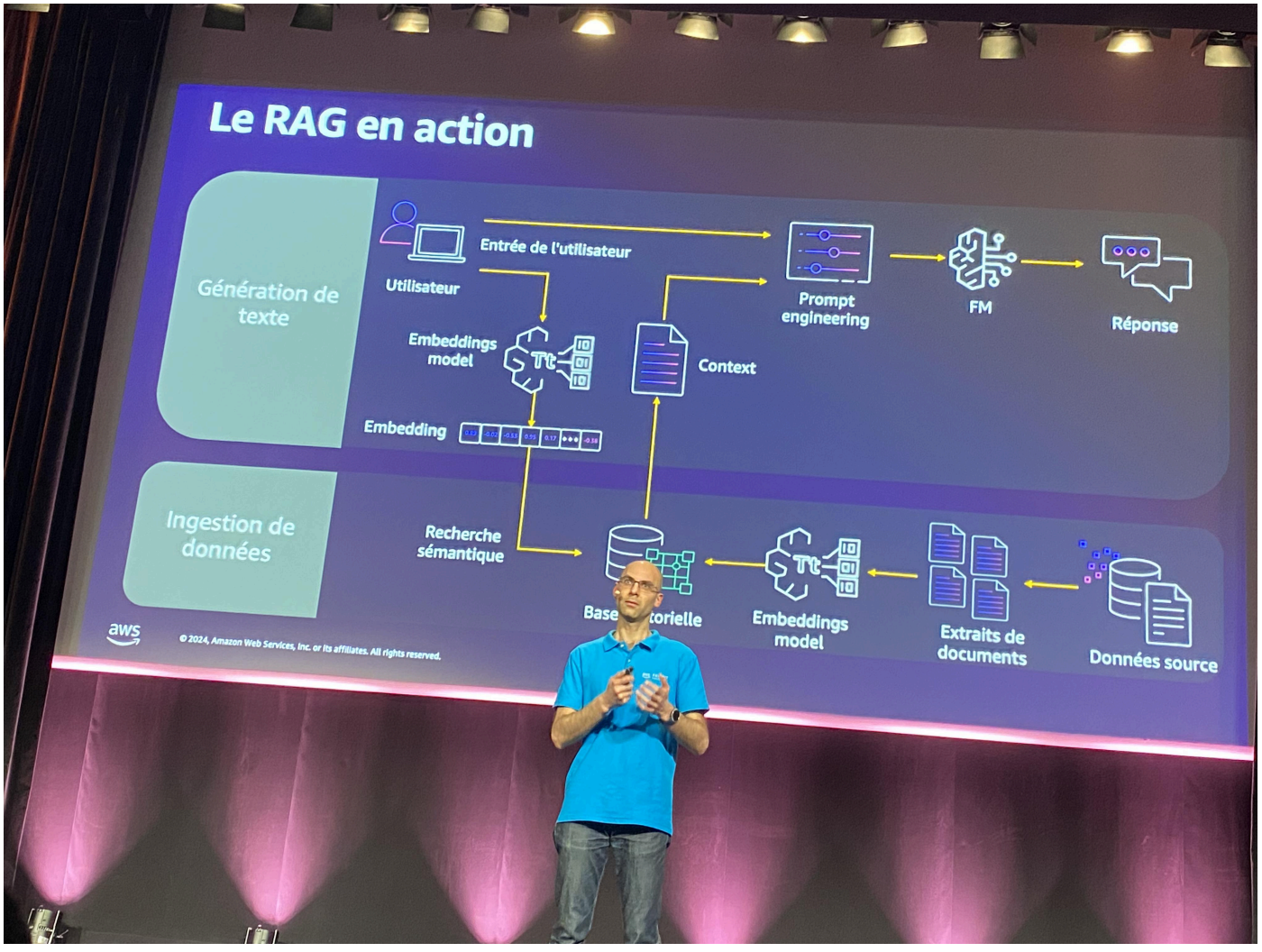

Le RAG en action
Ingestion de données
Base vectorielle
Embeddings model
Extraits de documents
Données so
aws
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

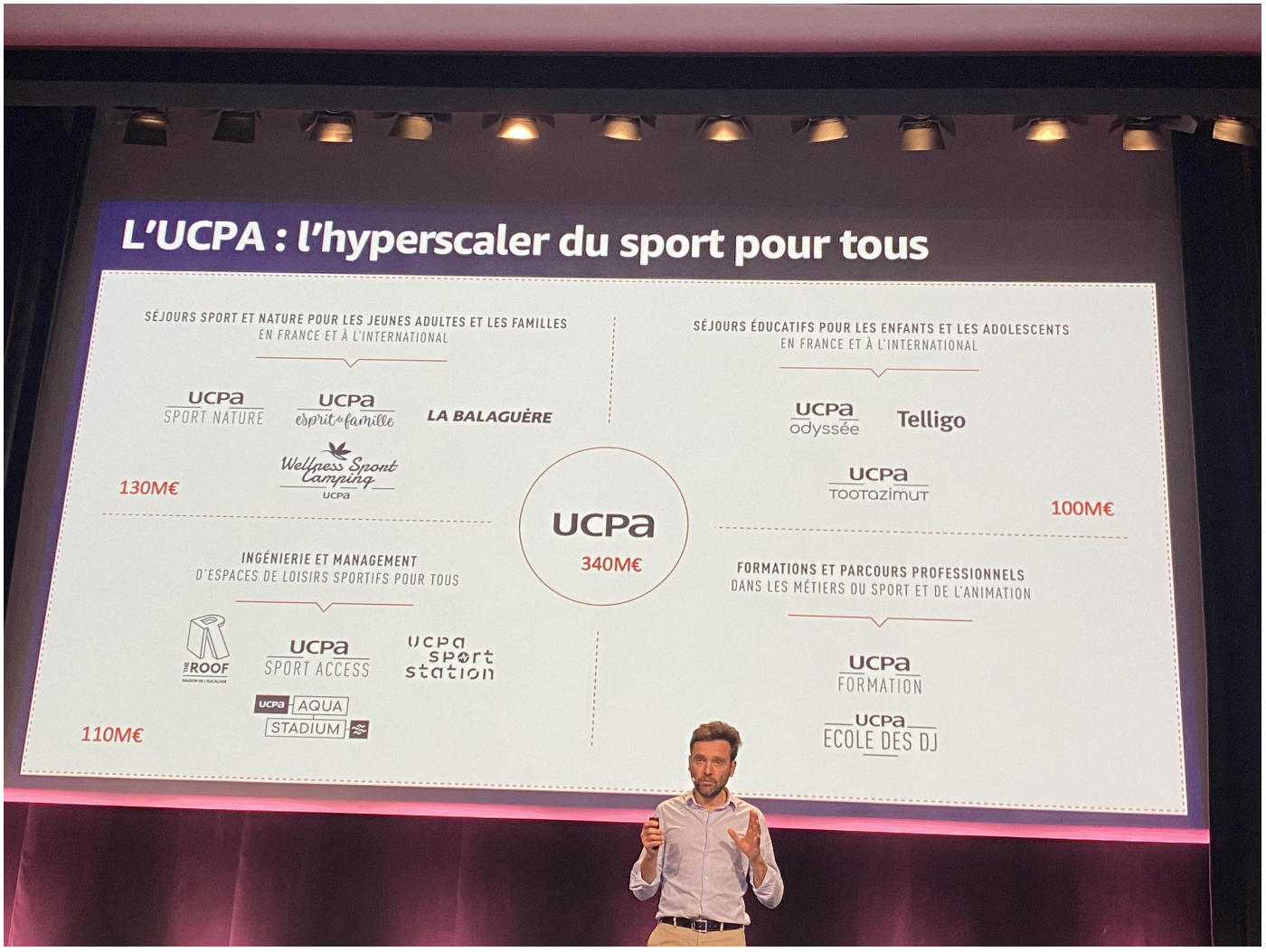

Le RAG en action
Ingestion de données
Embeddings model
Extraits de documents
Données source
aws
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

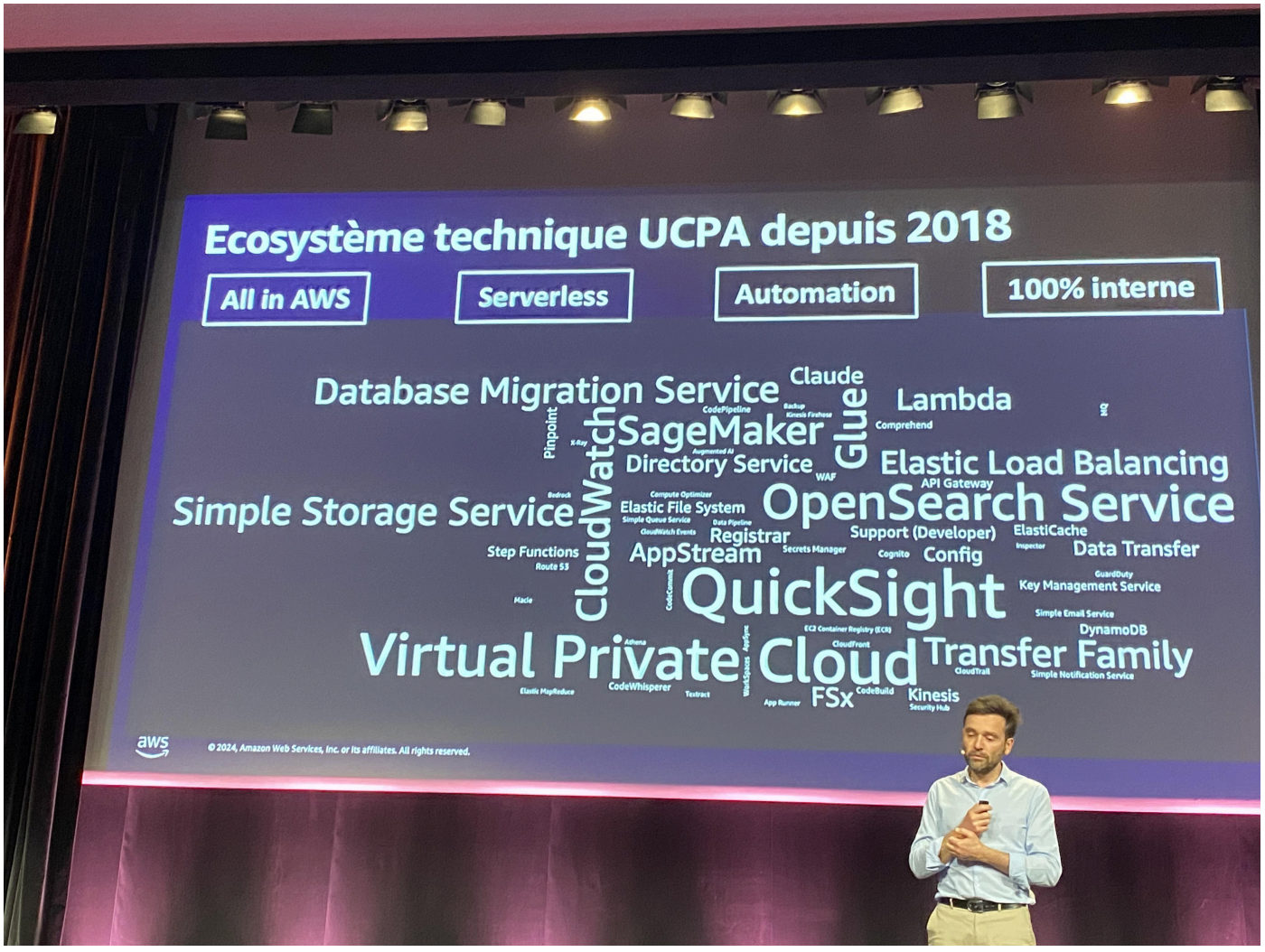

Le RAG en action
Génération de texte
Utilisateur
Embeddings model
Context
Embedding
Ingestion de données
Recherche sémantique
Embeddings model
Extraits de documents
aws
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

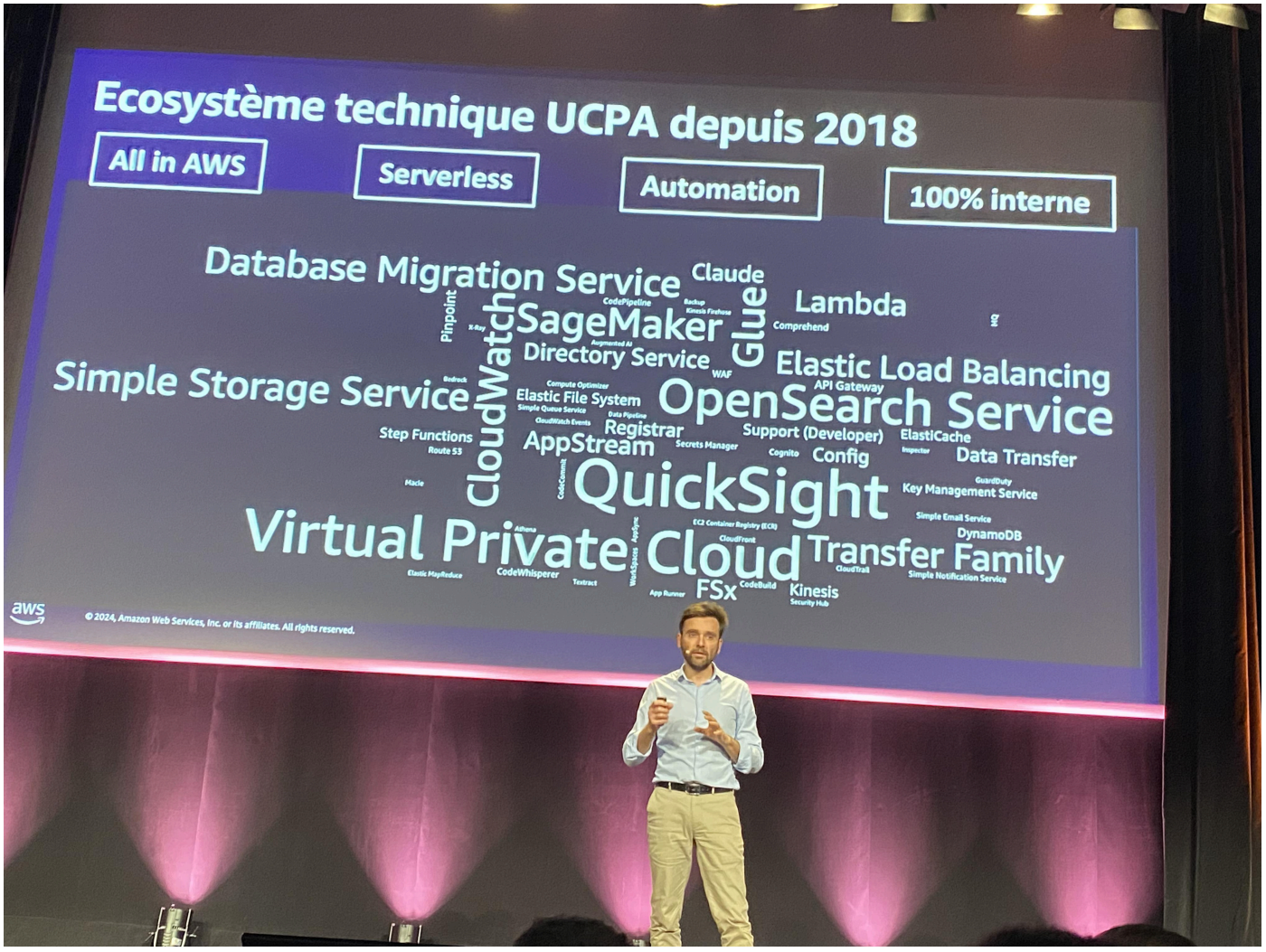

Le RAG en action
Génération de texte
Utilisateur
Entrée de l'utilisateur
Prompt engineering
FM
Réponse
Embeddings model
Context
Embedding
Ingestion de données
Recherche sémantique
Base vectorielle
Embeddings model
aws
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

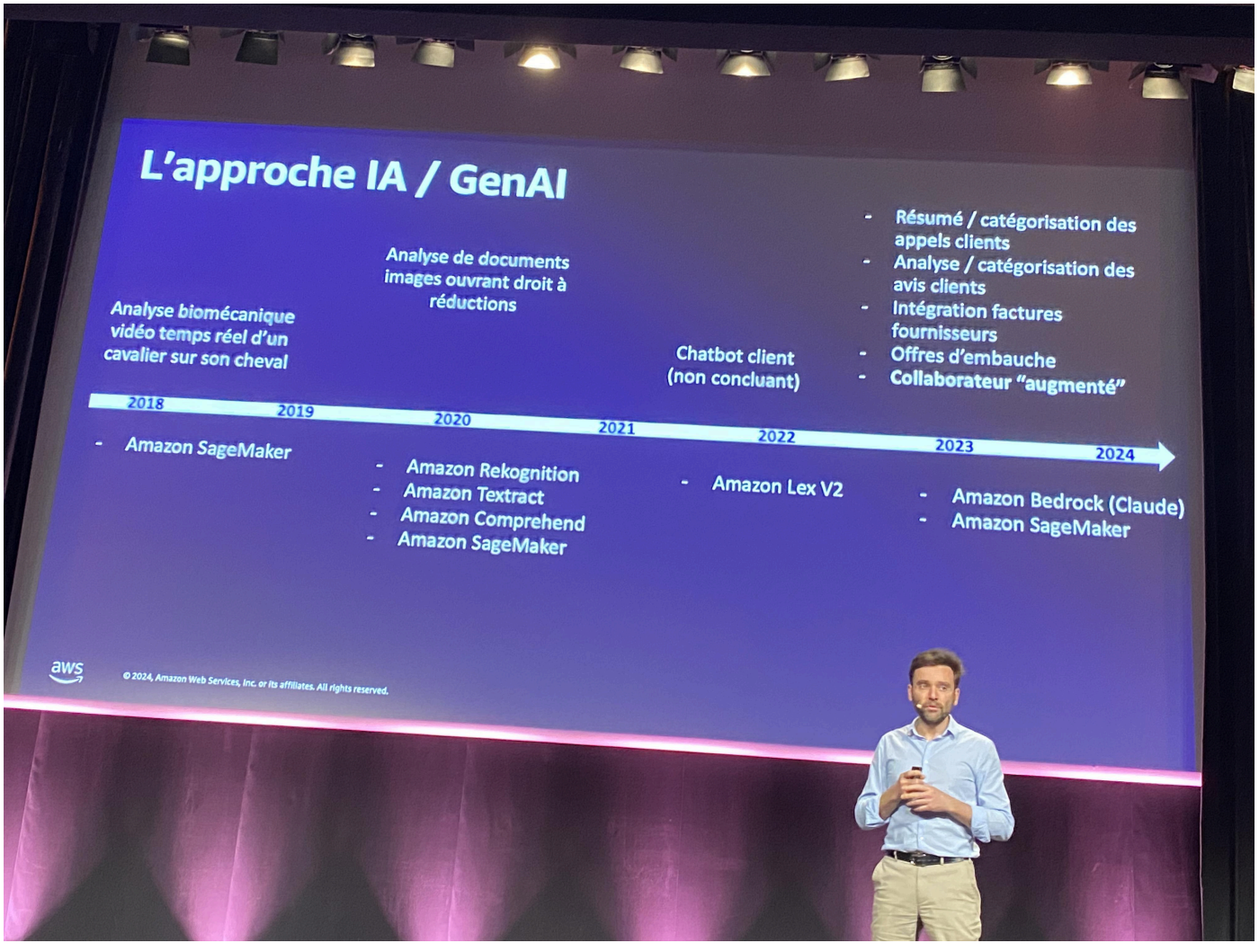

Le RAG en action
Génération de texte
Ingestion de données
Utilisateur
Entrée de l'utilisateur
Prompt engineering
FM
Réponse
Embeddings model
Context
Embedding
Recherche sémantique
Base vecto
Embeddings model
Extraits de documents
Données source
aws
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.


Le RAG en action
Génération de texte
Ingestion de données
Utilisateur
Entrée de l'utilisateur
Prompt engineering
FM
Réponse
Embeddings model
Context
Embedding
Recherche sémantique
Base torielle
Embeddings model
Extraits de documents
Données source
aws
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.


L'UCPA : l'hyperscaler du sport pour tous
SÉJOURS SPORT ET NATURE POUR LES JEUNES ADULTES ET LES FAMILLES
EN FRANCE ET À L'INTERNATIONAL
UCPA SPORT NATURE
UCPA esprit de famille
LA BALAGUÈRE
Wellness Sport Camping UCPa
130M€
SÉJOURS ÉDUCATIFS POUR LES ENFANTS ET LES ADOLESCENTS
EN FRANCE ET À L'INTERNATIONAL
UCPa odyssée
Telligo
UCPa TOOTAZIMUT
100M€
UCPa
340M€
INGÉNIERIE ET MANAGEMENT
D'ESPACES DE LOISIRS SPORTIFS POUR TOUS
THE ROOF
UCPa SPORT ACCESS
ucpa sport station
UCPa AQUA STADIUM
110M€
FORMATIONS ET PARCOURS PROFESSIONNELS
DANS LES MÉTIERS DU SPORT ET DE L'ANIMATION
UCPa FORMATION
UCPa ECOLE DES DJ

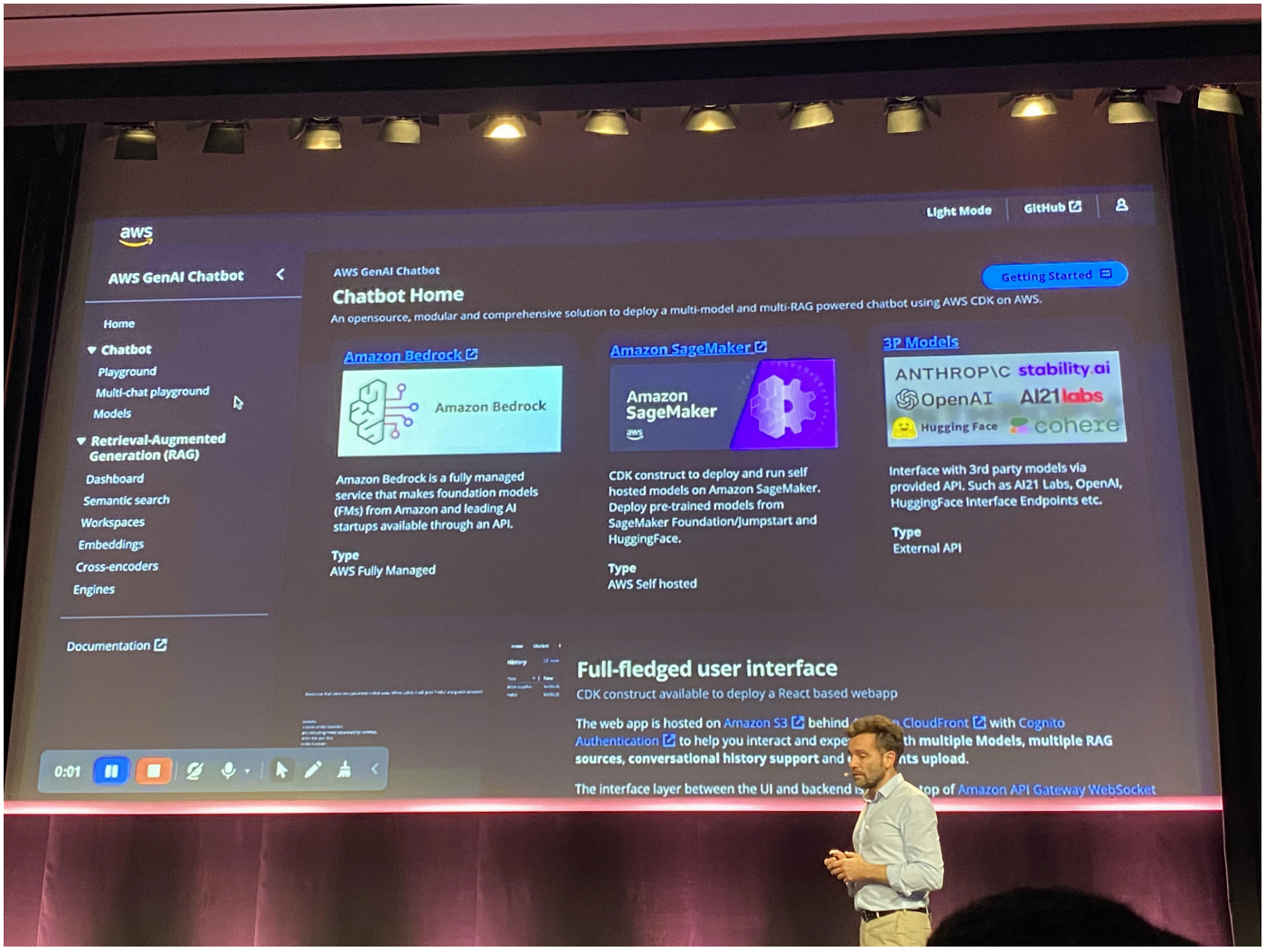

Ecosystème technique UCPA depuis 2018
All in AWS
Serverless
Automation
100% interne
Database Migration Service
Claude
Lambda
Glue
SageMaker
Comprehend
Directory Service
Elastic Load Balancing
API Gateway
Simple Storage Service
Elastic File System
OpenSearch Service
CloudWatch
Pinpoint
Registrar
Support (Developer)
ElastiCache
Step Functions
AppStream
Secrets Manager
Cognito
Config
Data Transfer
QuickSight
Key Management Service
Simple Email Service
DynamoDB
Virtual Private Cloud
Transfer Family
CodeWhisperer
FSx
CodeBuild
Kinesis
CloudTrail
Simple Notification Service
aws
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

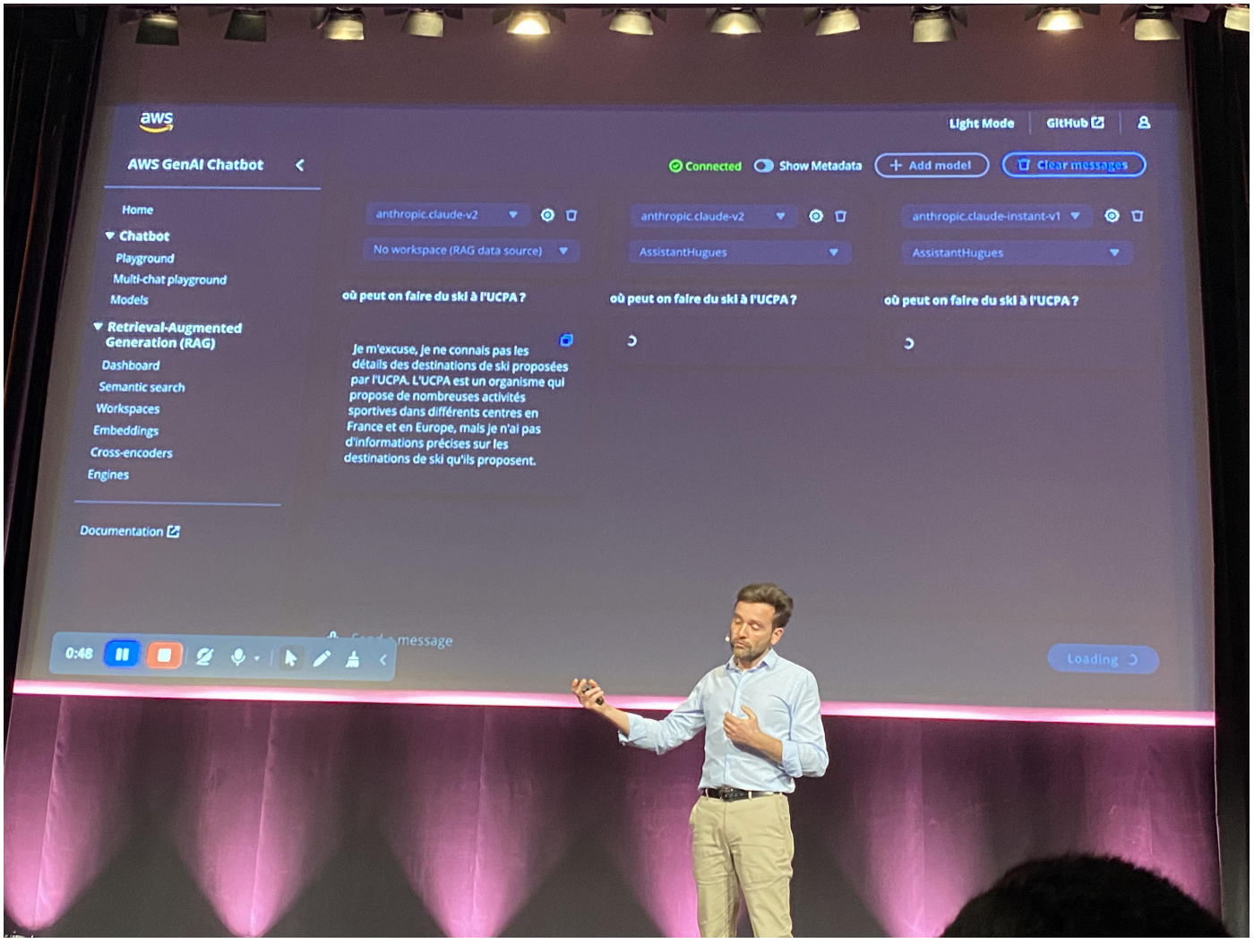

Ecosystème technique UCPA depuis 2018
All in AWS
Serverless
Automation
100% interne
Database Migration Service
Claude
Lambda
SageMaker
Glue
CloudWatch
Directory Service
Elastic Load Balancing
Simple Storage Service
OpenSearch Service
Elastic File System
Step Functions
AppStream
Registrar
Support (Developer)
ElastiCache
Data Transfer
QuickSight
Config
Key Management Service
Virtual Private Cloud
Transfer Family
DynamoDB
FSx
Kinesis
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

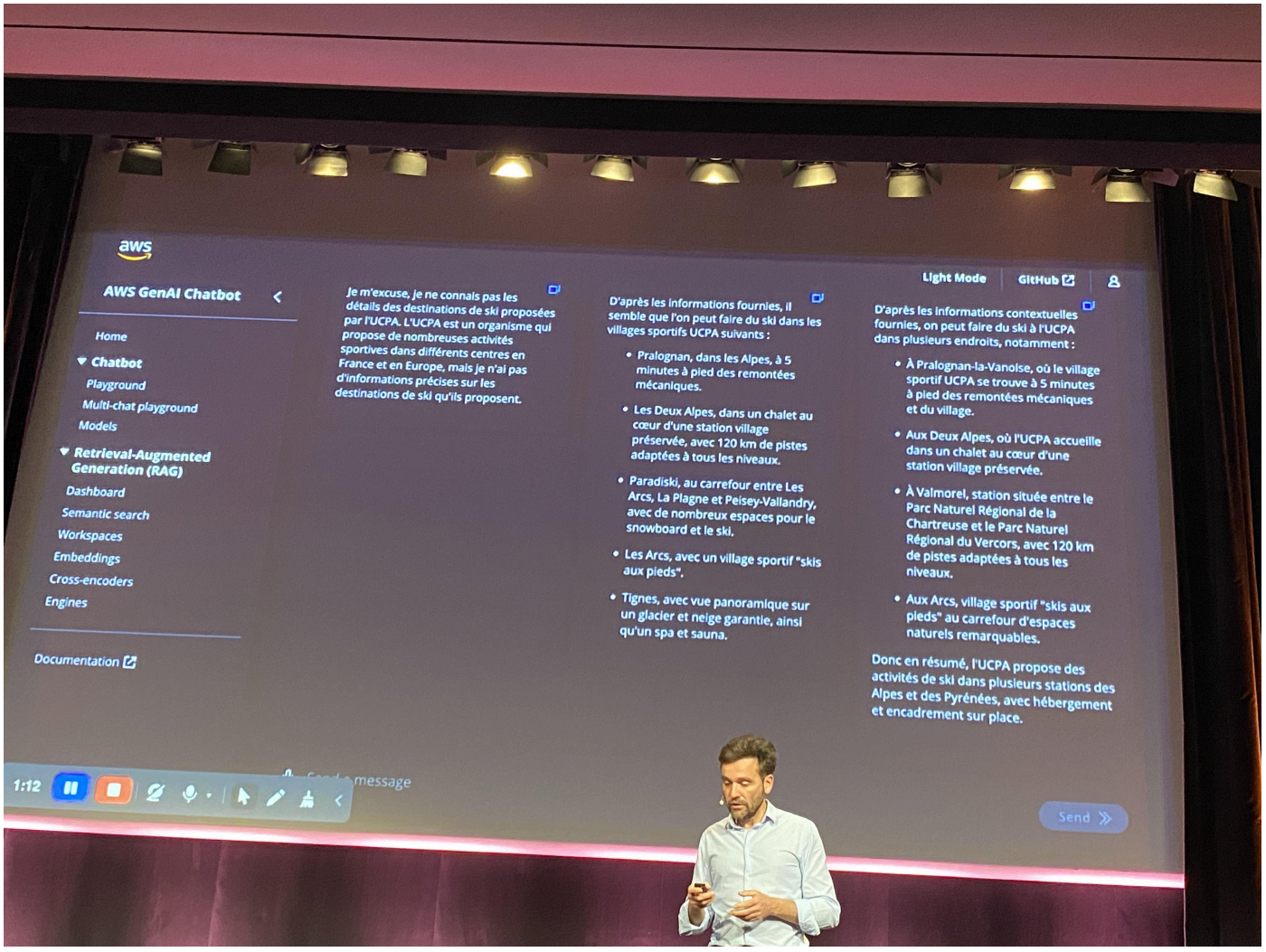

L'approche IA / GenAI
Analyse biomécanique vidéo temps réel d'un cavalier sur son cheval
Analyse de documents images ouvrant droit à réductions
Chatbot client (non concluant)
- Résumé / catégorisation des appels clients
- Analyse / catégorisation des avis clients
- Intégration factures fournisseurs
- Offres d'embauche
- Collaborateur "augmenté"
2018
2019
2020
2021
2022
2023
2024
- Amazon SageMaker
- Amazon Rekognition
- Amazon Textract
- Amazon Comprehend
- Amazon SageMaker
- Amazon Lex V2
- Amazon Bedrock (Claude)
- Amazon SageMaker
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

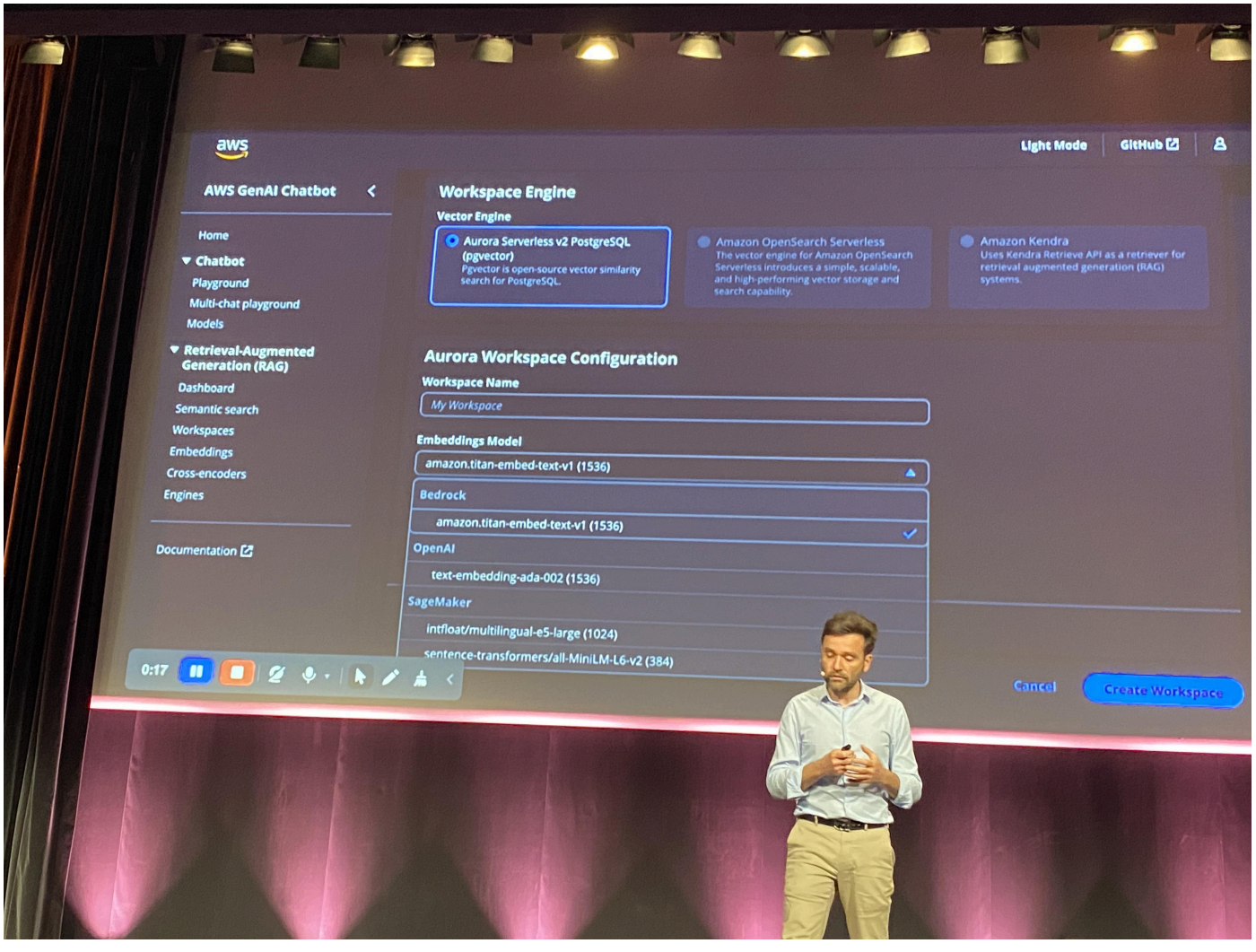

Le collaborateur “augmenté”
Le besoin : augmenter la créativité et la productivité des collaborateurs
Mettre à disposition de 200+ collaborateurs un Chatbot privé, en toute sécurité
Permettre au collaborateur de dialoguer avec ses propres données
Permettre au collaborateur de dialoguer avec les données de l’entreprise
La solution : intégrer une chaîne RAG sur Bedrock
D’abord en construisant tout nous même :
6 mois de POC itératif
Puis en intégrant un projet OpenSource : GENAI Chatbot
1 mois de développement / intégration
aws
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

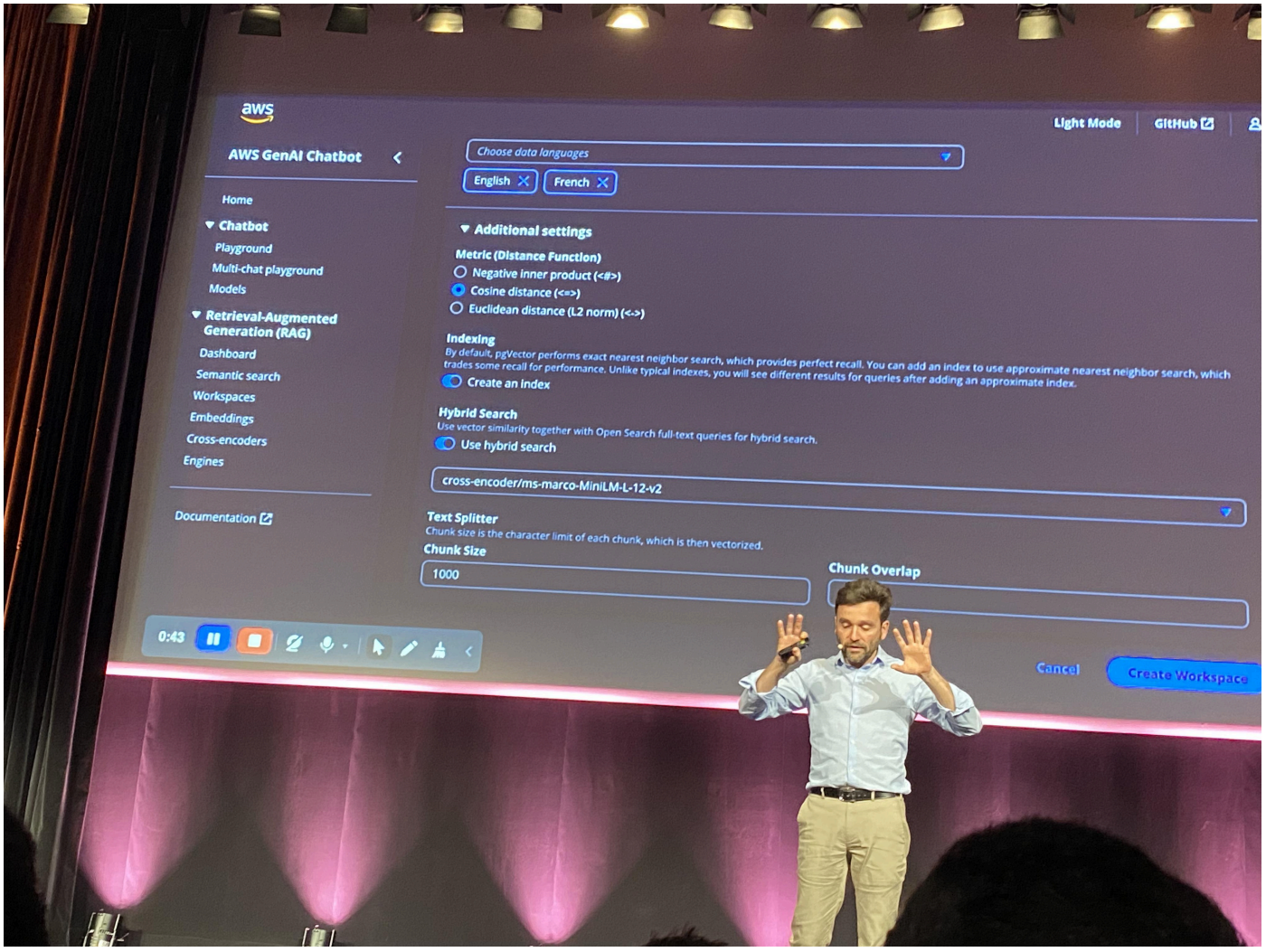

Le collaborateur “augmenté”
Le besoin : augmenter la créativité et la productivité des collaborateurs
Mettre à disposition de 200+ collaborateurs un Chatbot privé, en toute sécurité
Permettre au collaborateur de dialoguer avec ses propres données
Permettre au collaborateur de dialoguer avec les données de l’entreprise
La solution : intégrer une chaîne RAG sur Bedrock
D’abord en construisant tout nous même :
6 mois de POC itératif
Puis en intégrant un projet OpenSource : GENAI Chatbot
1 mois de développement / intégration
aws
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

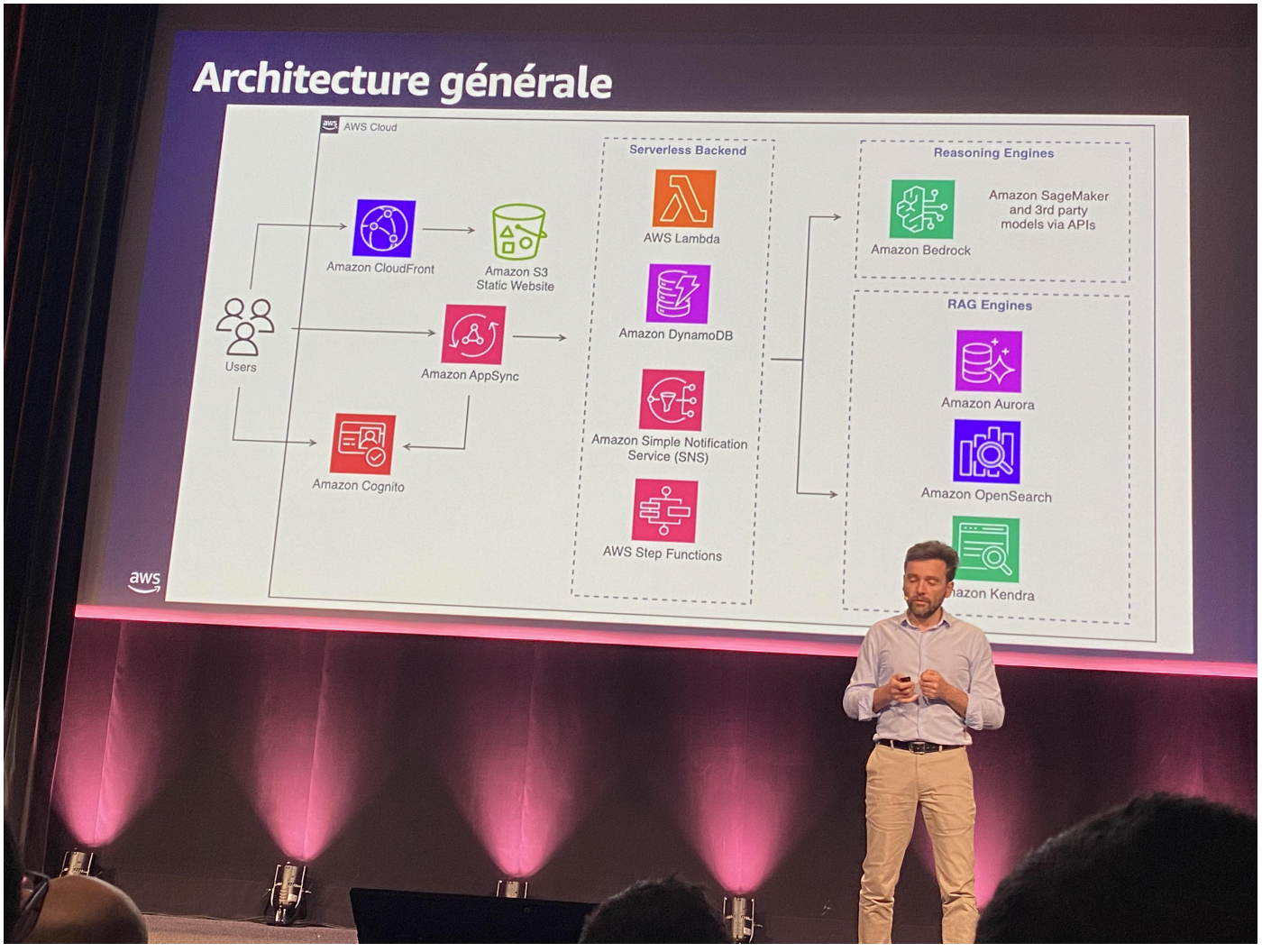

Light Mode
GitHub
AWS GenAI Chatbot
Home
Chatbot
Playground
Multi-chat playground
Models
Retrieval-Augmented Generation (RAG)
Dashboard
Semantic search
Workspaces
Embeddings
Cross-encoders
Engines
Documentation
AWS GenAI Chatbot
Chatbot Home
Getting Started
An opensource, modular and comprehensive solution to deploy a multi-model and multi-RAG powered chatbot using AWS CDK on AWS.
Amazon Bedrock
Amazon Bedrock is a fully managed service that makes foundation models (FMs) from Amazon and leading AI startups available through an API.
Type
AWS Fully Managed
Amazon SageMaker
Amazon SageMaker
CDK construct to deploy and run self hosted models on Amazon SageMaker. Deploy pre-trained models from SageMaker Foundation/Jumpstart and HuggingFace.
Type
AWS Self hosted
3P Models
ANTHROPIC stability.ai
OpenAI AI21labs
Hugging Face cohere
Interface with 3rd party models via provided API. Such as AI21 Labs, OpenAI, HuggingFace Interface Endpoints etc.
Type
External API
Full-fledged user interface
CDK construct available to deploy a React based webapp
The web app is hosted on Amazon S3 behind
CloudFront with Cognito Authentication to help you interact and expe
th multiple Models, multiple RAG sources, conversational history support and
nts upload.
The interface layer between the UI and backend
op of Amazon API Gateway WebSocket
0:01


Light Mode
GitHub
AWS GenAI Chatbot
Connected
Show Metadata
Add model
Clear messages
Home
Chatbot
Playground
Multi-chat playground
Models
Retrieval-Augmented Generation (RAG)
Dashboard
Semantic search
Workspaces
Embeddings
Cross-encoders
Engines
Documentation
anthropic.claude-v2
No workspace (RAG data source)
où peut on faire du ski à l'UCPA ?
Je m'excuse, je ne connais pas les détails des destinations de ski proposées par l'UCPA. L'UCPA est un organisme qui propose de nombreuses activités sportives dans différents centres en France et en Europe, mais je n'ai pas d'informations précises sur les destinations de ski qu'ils proposent.
anthropic.claude-v2
AssistantHugues
où peut on faire du ski à l'UCPA ?
anthropic.claude-instant-v1
AssistantHugues
où peut on faire du ski à l'UCPA ?
Loading
0:48


aws
AWS GenAI Chatbot
Home
Chatbot
Playground
Multi-chat playground
Models
Retrieval-Augmented Generation (RAG)
Dashboard
Semantic search
Workspaces
Embeddings
Cross-encoders
Engines
Documentation
Light Mode
GitHub
Je m'excuse, je ne connais pas les détails des destinations de ski proposées par l'UCPA. L'UCPA est un organisme qui propose de nombreuses activités sportives dans différents centres en France et en Europe, mais je n'ai pas d'informations précises sur les destinations de ski qu'ils proposent.
D'après les informations fournies, il semble que l'on peut faire du ski dans les villages sportifs UCPA suivants :
Pralognan, dans les Alpes, à 5 minutes à pied des remontées mécaniques.
Les Deux Alpes, dans un chalet au cœur d'une station village préservée, avec 120 km de pistes adaptées à tous les niveaux.
Paradiski, au carrefour entre Les Arcs, La Plagne et Peisey-Vallandry, avec de nombreux espaces pour le snowboard et le ski.
Les Arcs, avec un village sportif "skis aux pieds".
Tignes, avec vue panoramique sur un glacier et neige garantie, ainsi qu'un spa et sauna.
D'après les informations contextuelles fournies, on peut faire du ski à l'UCPA dans plusieurs endroits, notamment :
À Pralognan-la-Vanoise, où le village sportif UCPA se trouve à 5 minutes à pied des remontées mécaniques et du village.
Aux Deux Alpes, où l'UCPA accueille dans un chalet au cœur d'une station village préservée.
À Valmorel, station située entre le Parc Naturel Régional de la Chartreuse et le Parc Naturel Régional du Vercors, avec 120 km de pistes adaptées à tous les niveaux.
Aux Arcs, village sportif "skis aux pieds" au carrefour d'espaces naturels remarquables.
Donc en résumé, l'UCPA propose des activités de ski dans plusieurs stations des Alpes et des Pyrénées, avec hébergement et encadrement sur place.
1:12
Send


aws
AWS GenAI Chatbot
Home
Chatbot
Playground
Multi-chat playground
Models
Retrieval-Augmented Generation (RAG)
Dashboard
Semantic search
Workspaces
Embeddings
Cross-encoders
Engines
Documentation
Light Mode
GitHub
Workspace Engine
Vector Engine
Aurora Serverless v2 PostgreSQL (pgvector)
Pgvector is open-source vector similarity search for PostgreSQL.
Amazon OpenSearch Serverless
The vector engine for Amazon OpenSearch Serverless introduces a simple, scalable, and high-performing vector storage and search capability.
Amazon Kendra
Uses Kendra Retrieve API as a retriever for retrieval augmented generation (RAG) systems.
Aurora Workspace Configuration
Workspace Name
My Workspace
Embeddings Model
amazon.titan-embed-text-v1 (1536)
Bedrock
amazon.titan-embed-text-v1 (1536)
OpenAI
text-embedding-ada-002 (1536)
SageMaker
intfloat/multilingual-e5-large (1024)
sentence-transformers/all-MiniLM-L6-v2 (384)
0:17
Cancel
Create Workspace


aws
AWS GenAI Chatbot
Light Mode
GitHub
Choose data languages
English
French
Home
Chatbot
Playground
Multi-chat playground
Models
Retrieval-Augmented Generation (RAG)
Dashboard
Semantic search
Workspaces
Embeddings
Cross-encoders
Engines
Documentation
Additional settings
Metric (Distance Function)
Negative inner product (<#>)
Cosine distance (<=>)
Euclidean distance (L2 norm) (<->)
Indexing
By default, pgVector performs exact nearest neighbor search, which provides perfect recall. You can add an index to use approximate nearest neighbor search, which trades some recall for performance. Unlike typical indexes, you will see different results for queries after adding an approximate index.
Create an index
Hybrid Search
Use vector similarity together with Open Search full-text queries for hybrid search.
Use hybrid search
cross-encoder/ms-marco-MiniLM-L-12-v2
Text Splitter
Chunk size is the character limit of each chunk, which is then vectorized.
Chunk Size
1000
Chunk Overlap
0:43
Cancel
Create Workspace


Architecture générale
AWS Cloud
Users
Amazon CloudFront
Amazon S3
Static Website
Amazon AppSync
Amazon Cognito
Serverless Backend
AWS Lambda
Amazon DynamoDB
Amazon Simple Notification Service (SNS)
AWS Step Functions
Reasoning Engines
Amazon Bedrock
Amazon SageMaker and 3rd party models via APIs
RAG Engines
Amazon Aurora
Amazon OpenSearch
aws

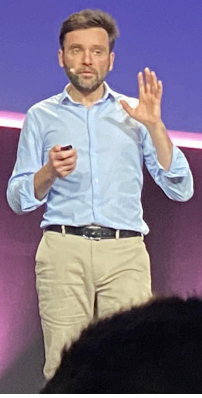
Les points clés à retenir

La confidentialité est indispensable

- Amazon Bedrock obligatoire (Francfort)

Le RAG n'est pas magique

- choisir la meilleure stratégie de "chunk" (taille, paragraphe, phrase, thèmes)
- adapter la stratégie de "chunk" aux types de contenus
- intégrer une brique de reranking pour améliorer la qualité des réponses
- essayer plusieurs LLM et choisir celui qui fonctionne le mieux pour l'usage



Les points clés à retenir

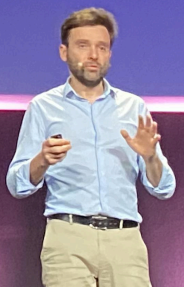
La confidentialité est indispensable

- Amazon Bedrock obligatoire (Francfort)

Le RAG n'est pas magique

- choisir la meilleure stratégie de "chunk" (taille, paragraphe, phrase, thèmes)
- adapter la stratégie de "chunk" aux types de contenus
- intégrer une brique de reranking pour améliorer la qualité des réponses
- essayer plusieurs LLM et choisir celui qui fonctionne le mieux pour l'usage

aws © 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Les points clés à retenir

Gérer le cycle de vie de la “connaissance”
- éviter les mauvaises données

Hallucinations
- à gérer impérativement

Peu de recul à ce jour sur les coûts au token
- Attention à la facture !

Peu de LLM “parlent” français
- Claude privilégié, en attente de Mistral (sur Francfort)

Knowledge Bases pour Amazon Bedrock

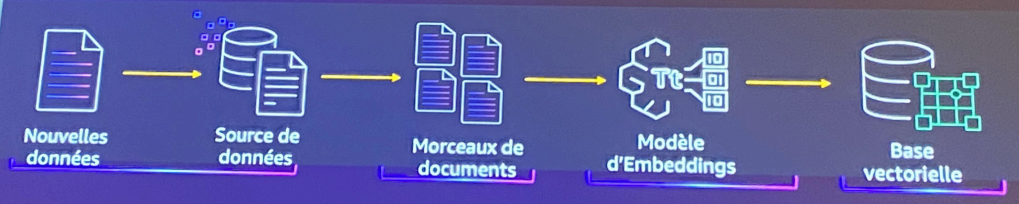
Fournit aux FM et aux agents des informations contextuelles provenant de vos sources de données privées pour RAG afin de fournir des réponses plus pertinentes, précises et personnalisées.

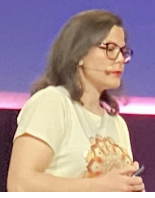
- Support entièrement géré pour le workflow RAG de bout en bout
- Connectez en toute sécurité les FM et les agents aux sources de données
- Récupérez facilement les données pertinentes et améliorez les prompts
- Indiquez l'attribution de la source

Workflow d'ingestion de données

KNOWLEDGE BASES SUR AMAZON BEDROCK

Workflow d'ingestion de données entièrement géré





Workflow d'ingestion de données

KNOWLEDGE BASES SUR AMAZON BEDROCK

Workflow d'ingestion de données entièrement géré

Nouvelles données
Source de données
Morceaux de documents
Modèle d'Embeddings
Base vectorielle

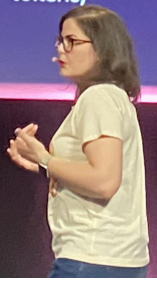
- Choisissez votre source de données (Amazon S3)
- Prise en charge des mises à jour incrémentales
- Plusieurs formats de fichiers de données pris en charge

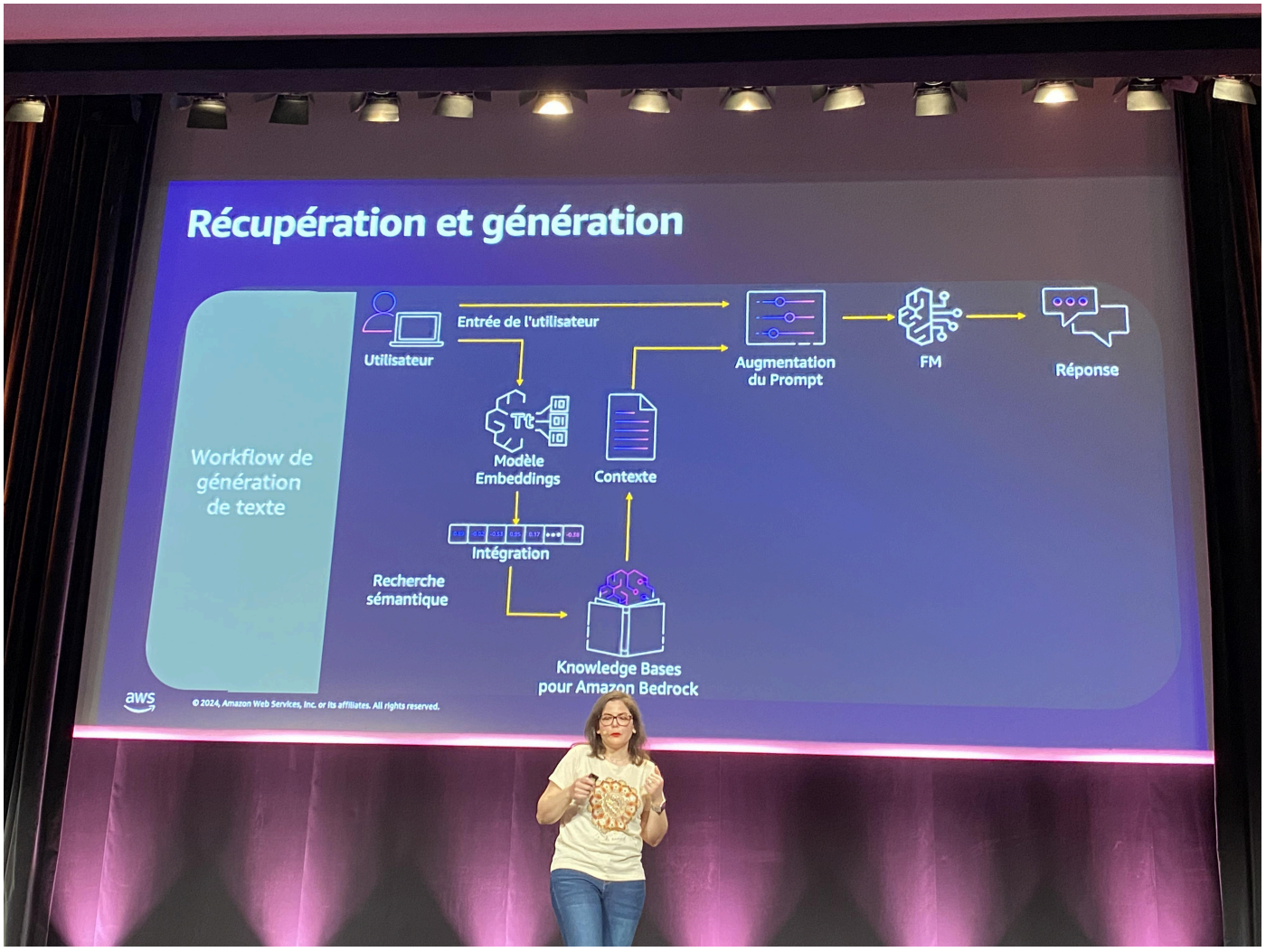
- Choisissez votre stratégie de découpage
  - Morceaux fixes
  - Pas de découpage
  - Par défaut (200 tokens)

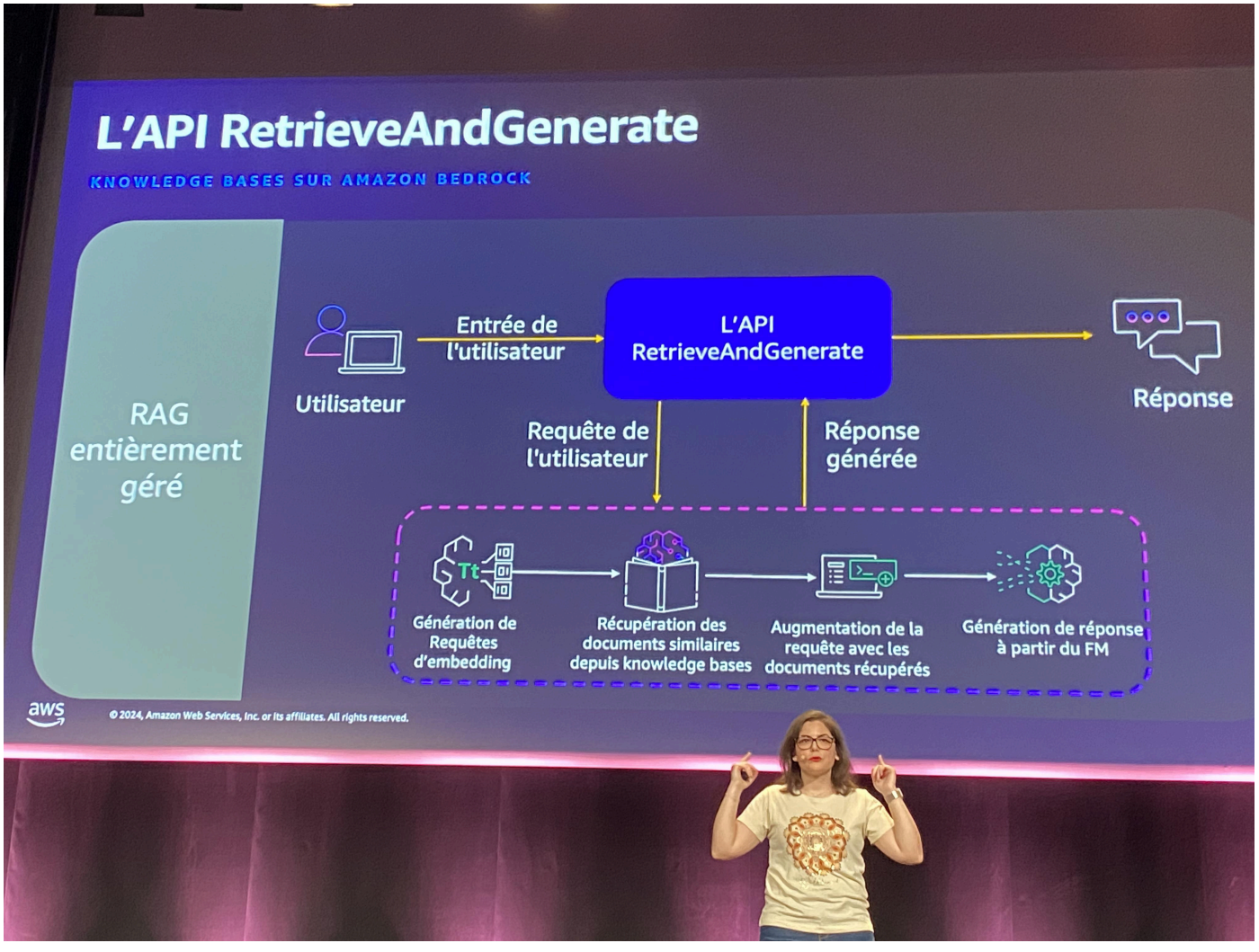
- Choisissez votre modèle d'intégration
  - Amazon Titan
  - Cohere Embed Multilingual
  - Cohere Embed English

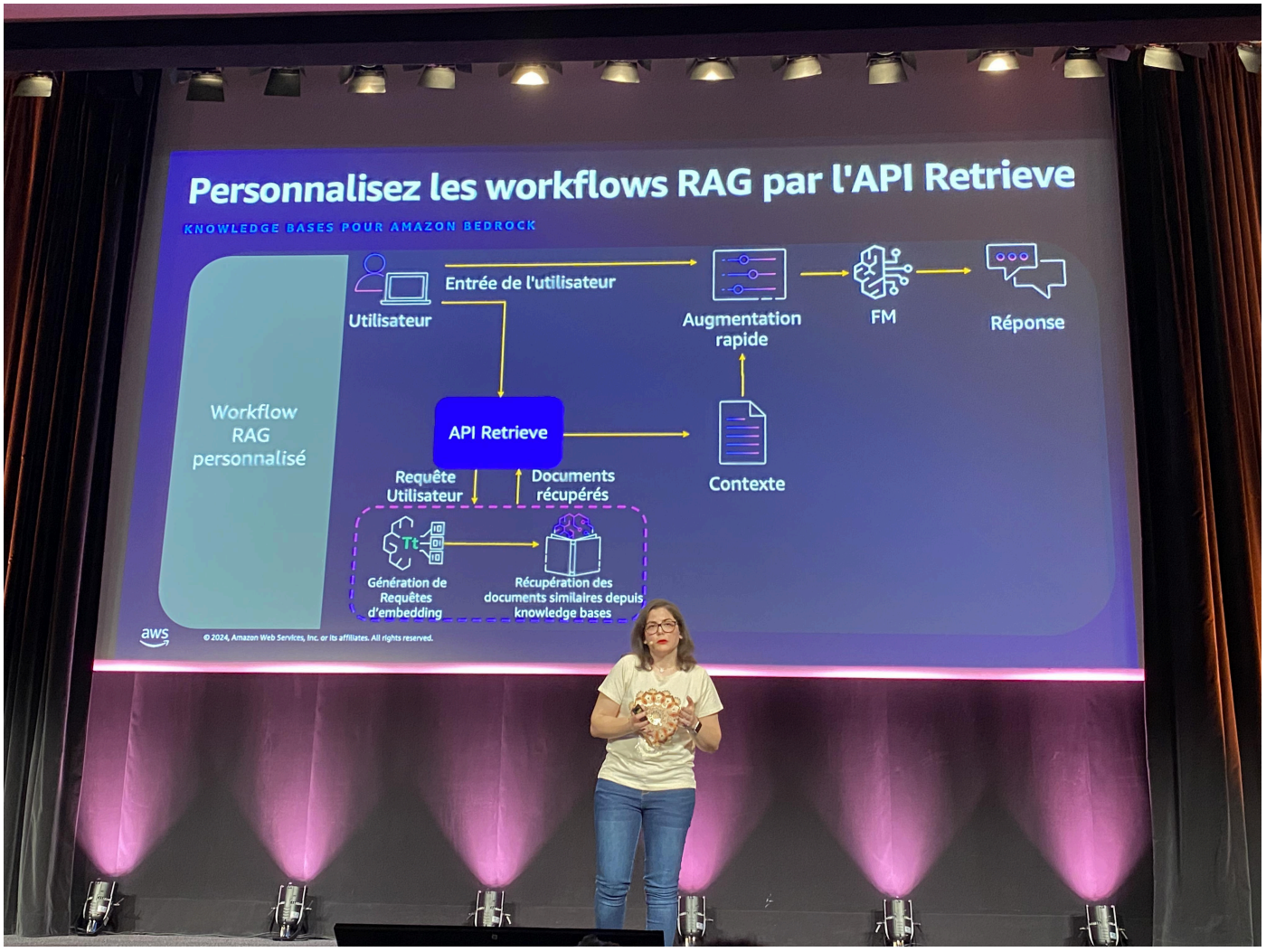
- Choisissez votre base vectorielle
  - OpenSearch serverless
  - Aurora Postgres
  - Redis
  - Pinecone

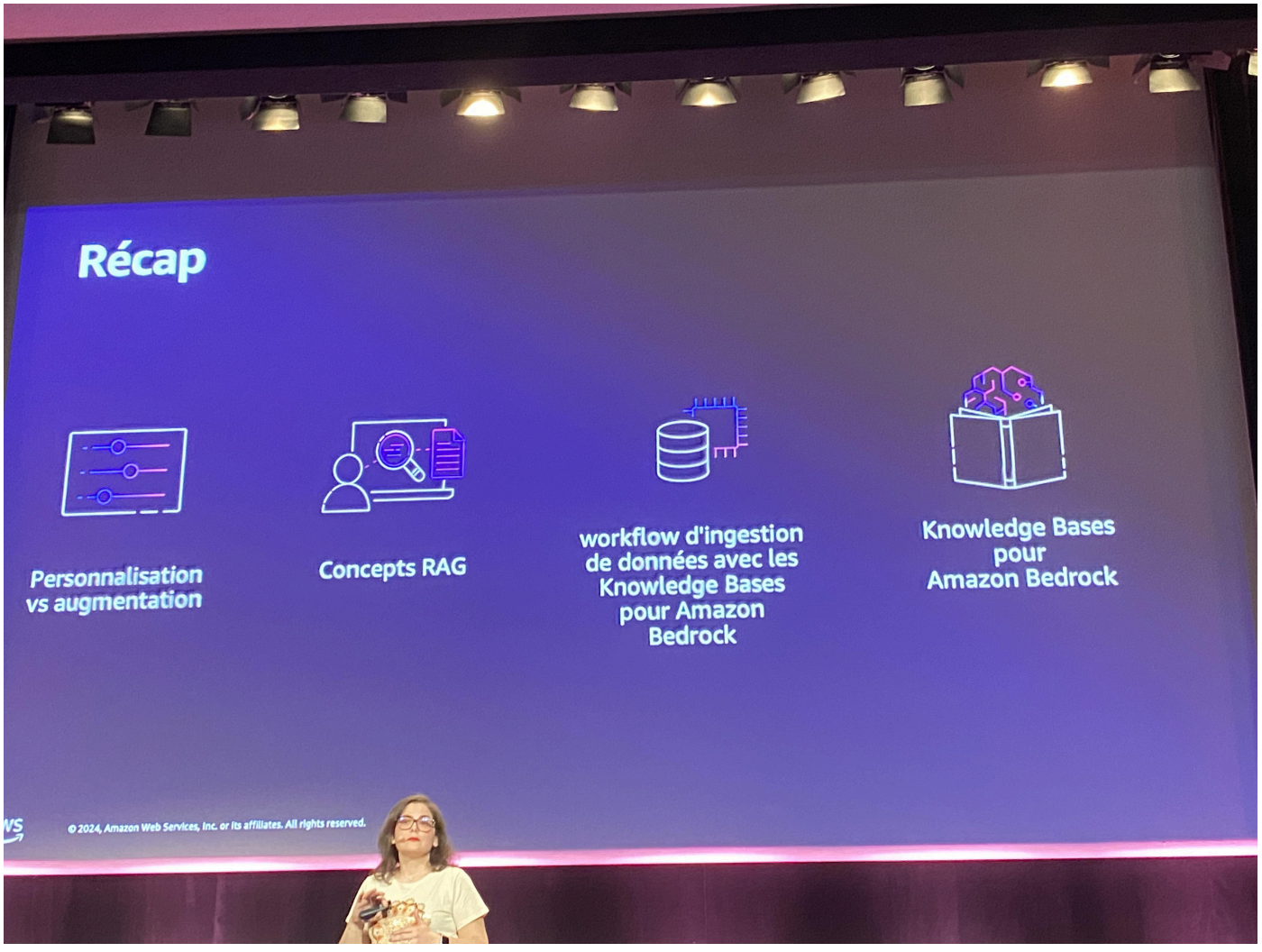
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

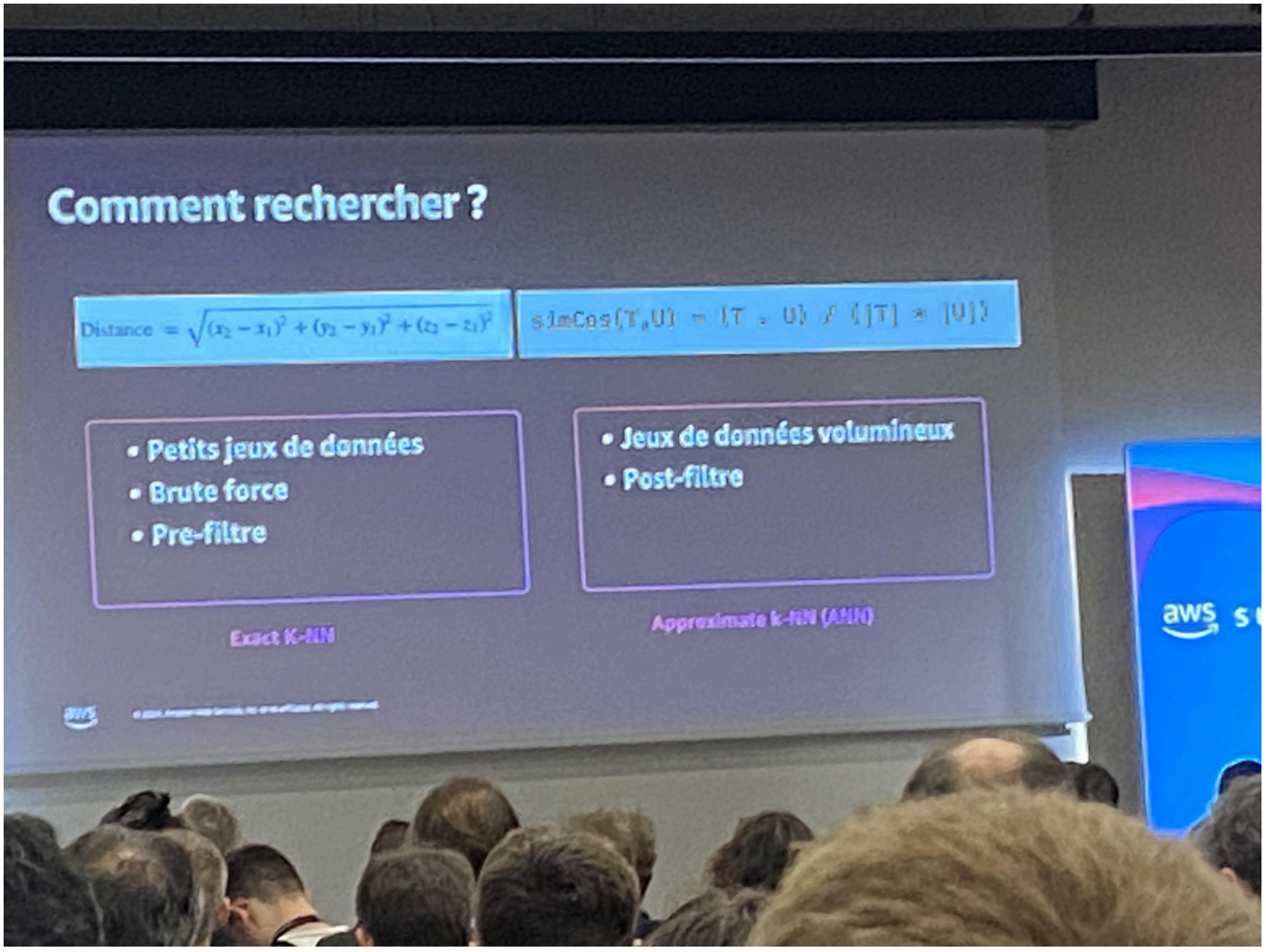



Récupération et génération
Workflow de génération de texte
Utilisateur
Entrée de l'utilisateur
Augmentation du Prompt
FM
Réponse
Modèle Embeddings
Contexte
Intégration
Recherche sémantique
Knowledge Bases pour Amazon Bedrock
aws
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.


L'API RetrieveAndGenerate
KNOWLEDGE BASES SUR AMAZON BEDROCK
RAG entièrement géré
Utilisateur
Entrée de l'utilisateur
L'API RetrieveAndGenerate
Réponse
Requête de l'utilisateur
Réponse générée
Génération de Requêtes d'embedding
Récupération des documents similaires depuis knowledge bases
Augmentation de la requête avec les documents récupérés
Génération de réponse à partir du FM
aws
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.


Personnalisez les workflows RAG par l'API Retrieve
KNOWLEDGE BASES POUR AMAZON BEDROCK
Workflow RAG personnalisé
Utilisateur
Entrée de l'utilisateur
Augmentation rapide
FM
Réponse
API Retrieve
Requête Utilisateur
Documents récupérés
Contexte
Tt
Génération de Requêtes d'embedding
Récupération des documents similaires depuis knowledge bases
aws
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.


Récap
Personnalisation vs augmentation
Concepts RAG
workflow d'ingestion de données avec les Knowledge Bases pour Amazon Bedrock
Knowledge Bases pour Amazon Bedrock
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.


Comment rechercher ?
Distance = $\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$
simCos(T,U) = (T . U) / (|T| * |U|)
• Petits jeux de données
• Brute force
• Pre-filtre
Exact K-NN
• Jeux de données volumineux
• Post-filtre
Approximate k-NN (ANN)
aws