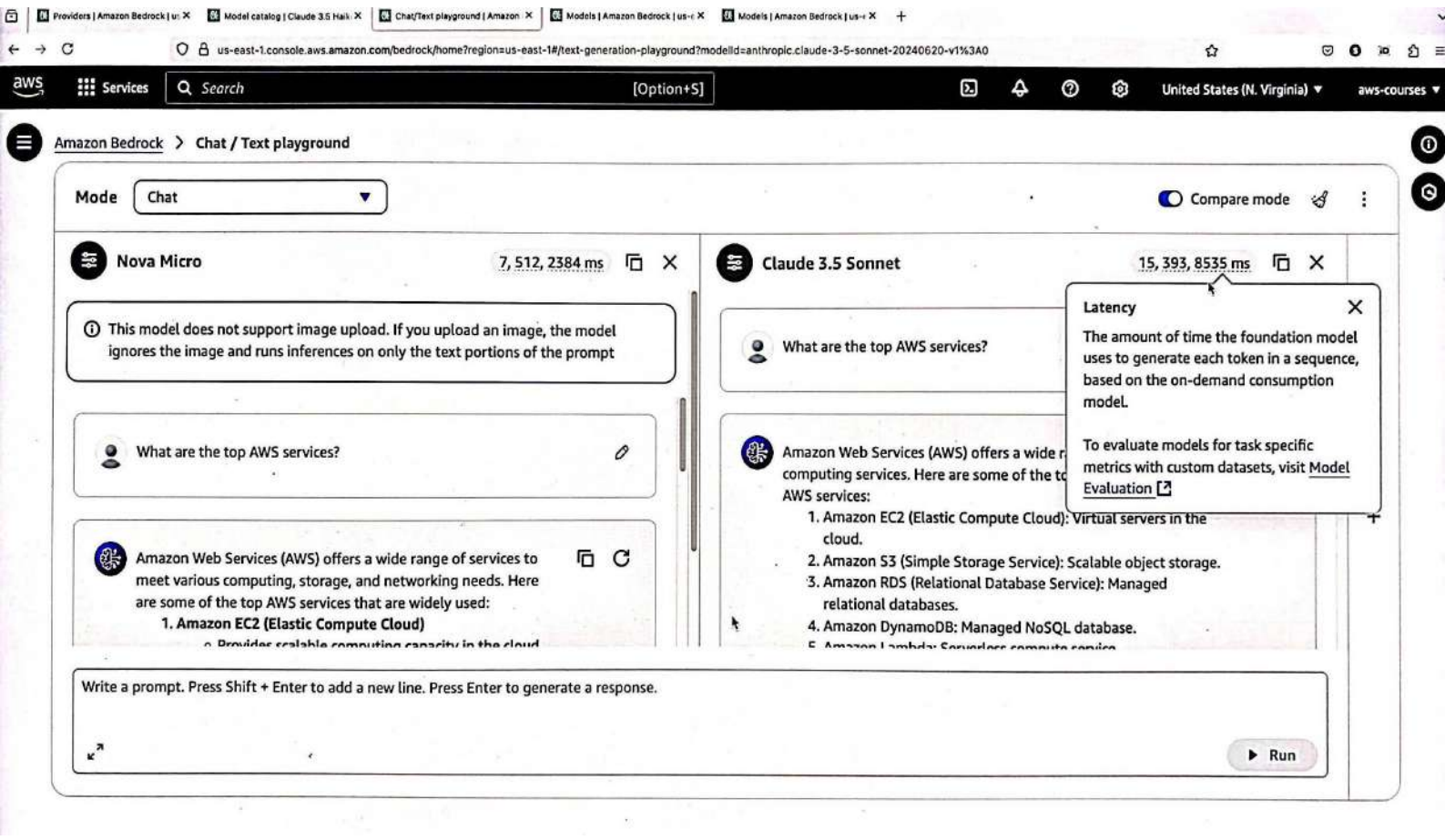
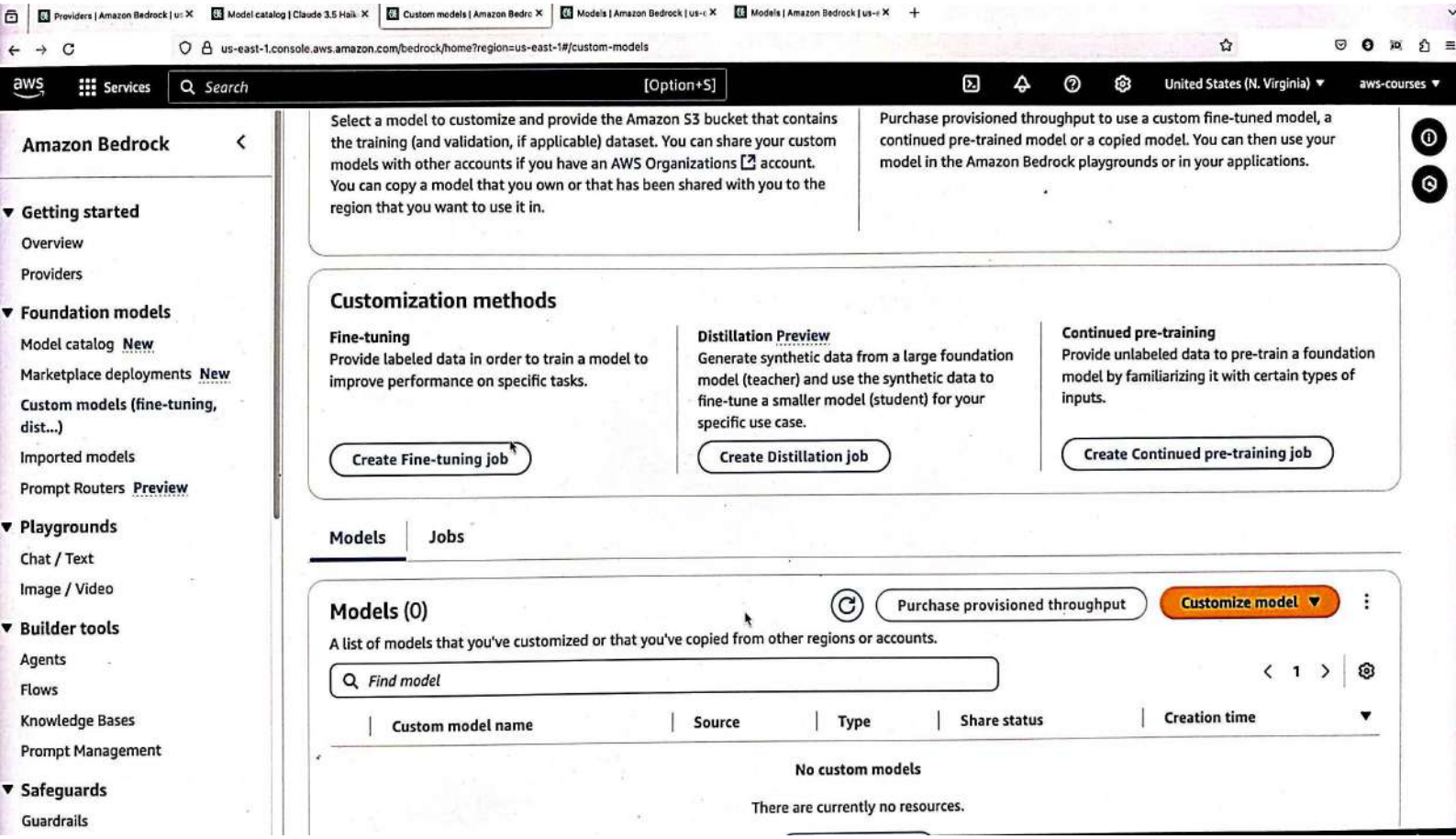
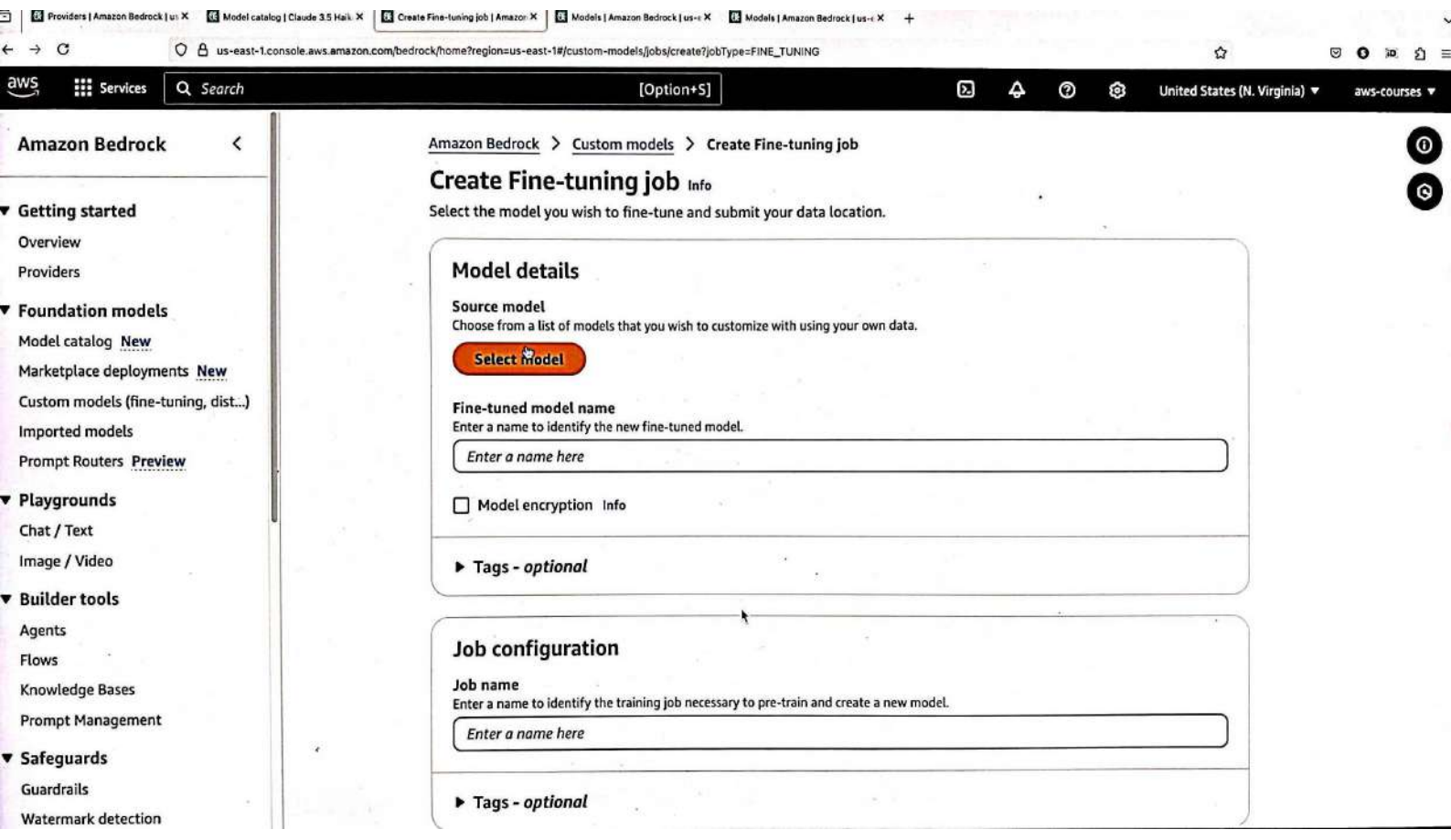
Example
Amazon Titan vs. Llama vs. Claude vs. Stable Diffusion

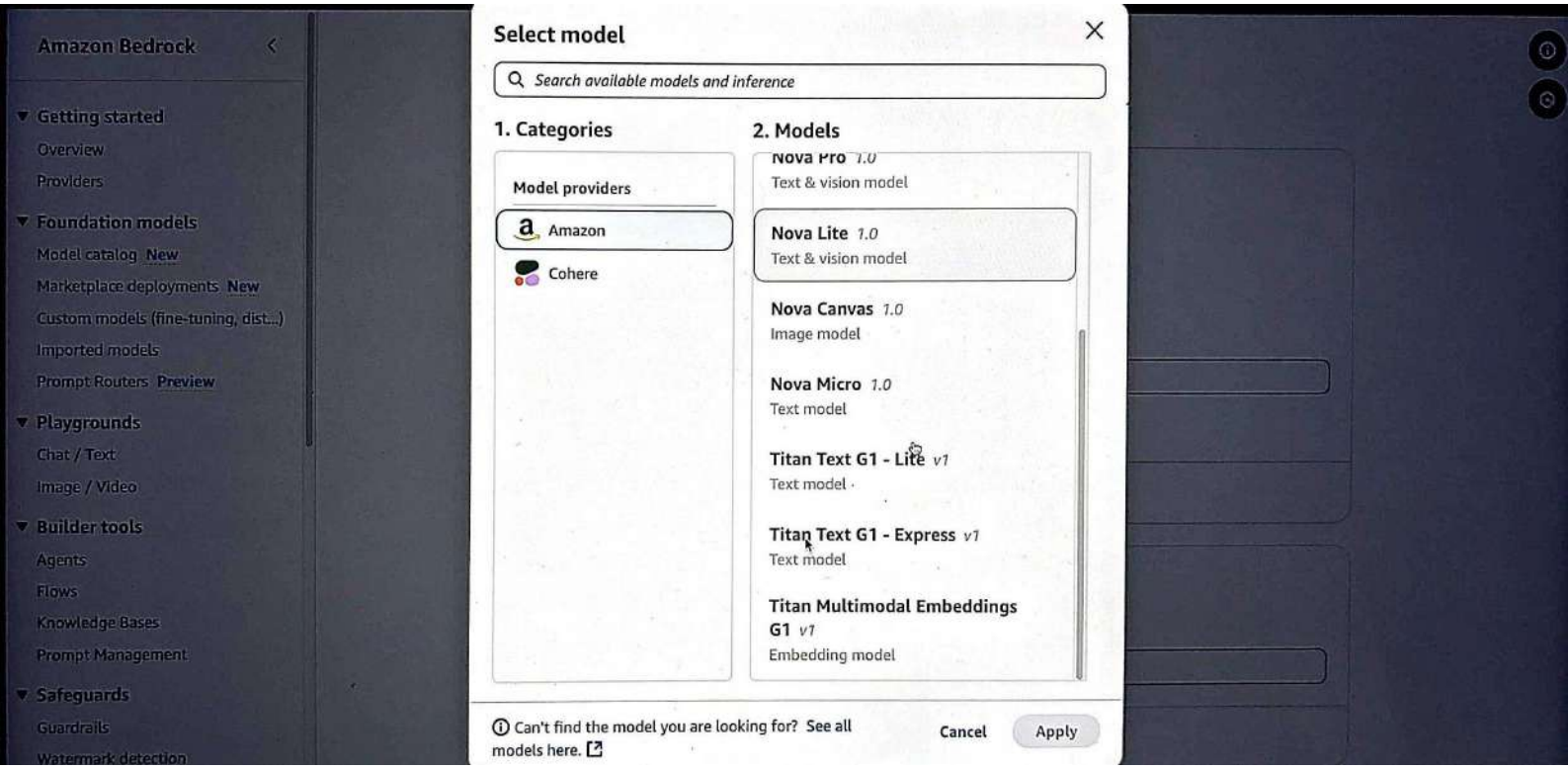
amazon LLAMA 2 Claude stability.ai

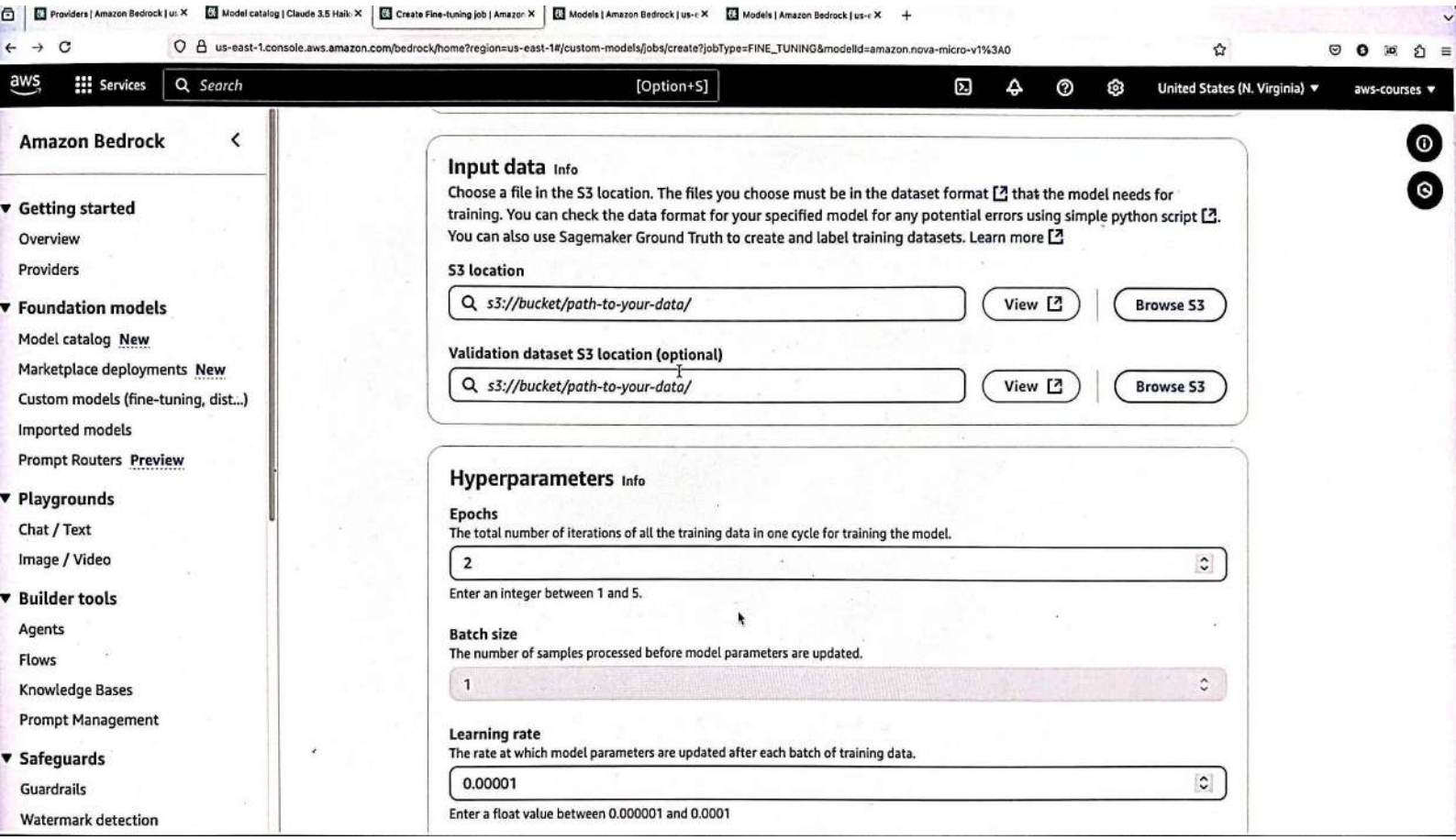
	Amazon Titan (Titan Text Express)	Llama (Llama-2 70b-chat)	Claude (Claude 2.1)	Stable Diffusion (SDXL 1.0)
Max. Tokens (=max context window)	8K Tokens	4K Tokens	200K Tokens	77-Tokens/Prompt
Features	High-performance text model, +100 languages	Large-scale tasks, dialogue, English	High-capacity text generation, multi-language	Image generation
Use cases	Content creation, classification, education...	Text generation, customer service...	Analysis, forecasting, document comparison...	Image creation for advertising, media...
Pricing (1K Tokens)	Input: $0.0008 Output: $0.0016	Input: $0.0019 Output: $0.0025	Input: $0.008 Output: $0.024	$0.04 – 0.08 / image

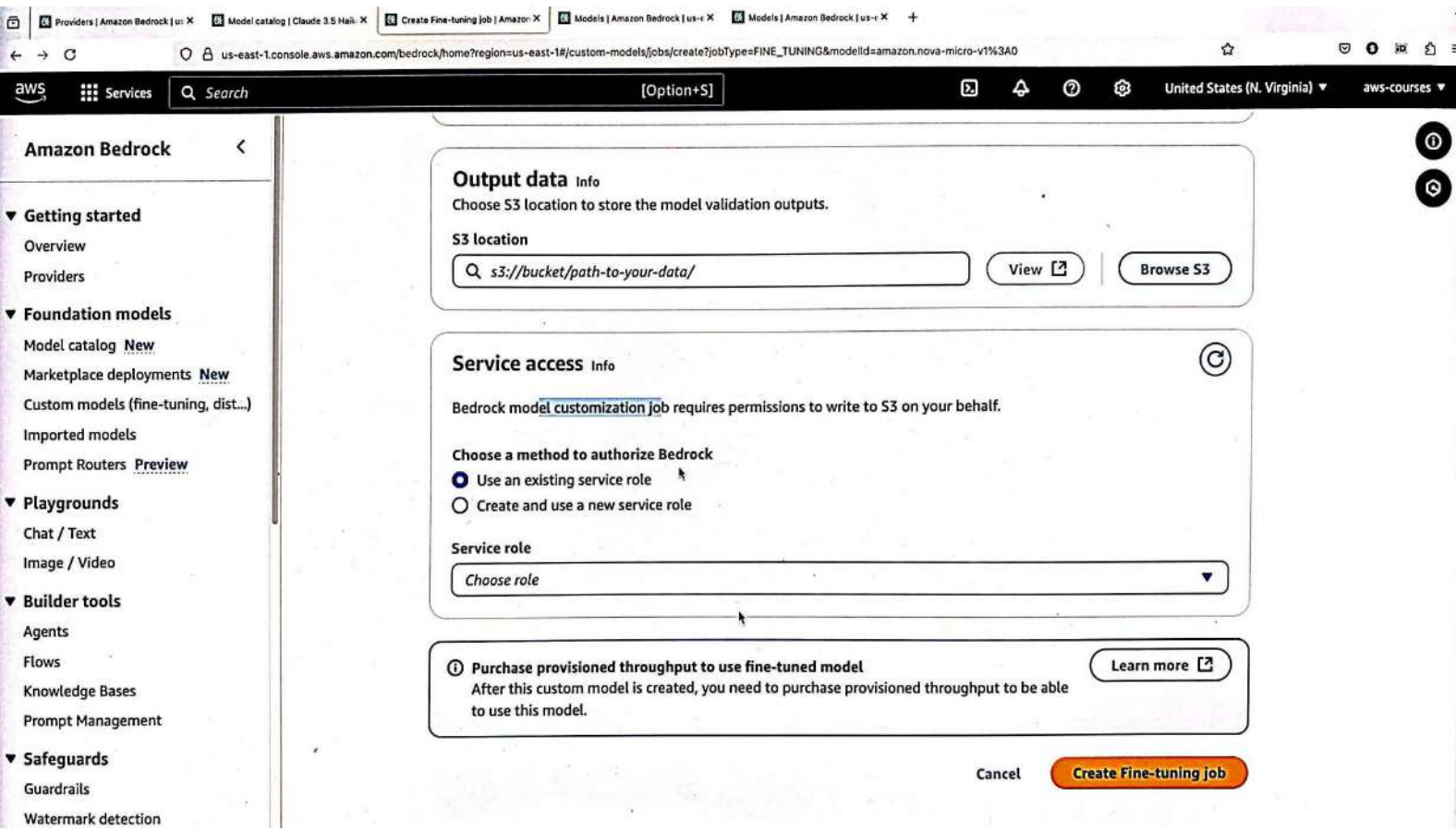

Providers | Amazon Bedrock | u
Model catalog | Claude 3.5 Hai
Chat/Text playground | Amazon
Models | Amazon Bedrock | us-e
Models | Amazon Bedrock | us-e
us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/text-generation-playground?modelId=anthropic.claude-3-5-sonnet-20240620-v1%3A0
aws
Services
Search
[Option+S]
United States (N. Virginia)
aws-courses
Amazon Bedrock > Chat / Text playground
Mode
Chat
Compare mode
Nova Micro
7, 512, 2384 ms
This model does not support image upload. If you upload an image, the model ignores the image and runs inferences on only the text portions of the prompt
What are the top AWS services?
Amazon Web Services (AWS) offers a wide range of services to meet various computing, storage, and networking needs. Here are some of the top AWS services that are widely used:
1. Amazon EC2 (Elastic Compute Cloud)
Claude 3.5 Sonnet
15, 393, 8535 ms
What are the top AWS services?
Amazon Web Services (AWS) offers a wide r
computing services. Here are some of the to
AWS services:
1. Amazon EC2 (Elastic Compute Cloud): Virtual servers in the cloud.
2. Amazon S3 (Simple Storage Service): Scalable object storage.
3. Amazon RDS (Relational Database Service): Managed relational databases.
4. Amazon DynamoDB: Managed NoSQL database.
Latency
The amount of time the foundation model uses to generate each token in a sequence, based on the on-demand consumption model.
To evaluate models for task specific metrics with custom datasets, visit Model Evaluation
Write a prompt. Press Shift + Enter to add a new line. Press Enter to generate a response.
Run


us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/custom-models
aws
Services
Search
[Option+S]
United States (N. Virginia)
aws-courses
Amazon Bedrock
Getting started
Overview
Providers
Foundation models
Model catalog New
Marketplace deployments New
Custom models (fine-tuning, dist...)
Imported models
Prompt Routers Preview
Playgrounds
Chat / Text
Image / Video
Builder tools
Agents
Flows
Knowledge Bases
Prompt Management
Safeguards
Guardrails
Select a model to customize and provide the Amazon S3 bucket that contains the training (and validation, if applicable) dataset. You can share your custom models with other accounts if you have an AWS Organizations account. You can copy a model that you own or that has been shared with you to the region that you want to use it in.
Purchase provisioned throughput to use a custom fine-tuned model, a continued pre-trained model or a copied model. You can then use your model in the Amazon Bedrock playgrounds or in your applications.
Customization methods
Fine-tuning
Provide labeled data in order to train a model to improve performance on specific tasks.
Create Fine-tuning job
Distillation Preview
Generate synthetic data from a large foundation model (teacher) and use the synthetic data to fine-tune a smaller model (student) for your specific use case.
Create Distillation job
Continued pre-training
Provide unlabeled data to pre-train a foundation model by familiarizing it with certain types of inputs.
Create Continued pre-training job
Models
Jobs
Models (0)
Purchase provisioned throughput
Customize model
A list of models that you've customized or that you've copied from other regions or accounts.
Find model
1
Custom model name
Source
Type
Share status
Creation time
No custom models
There are currently no resources.


Providers | Amazon Bedrock | us
Model catalog | Claude 3.5 Hai
Create Fine-tuning job | Amazon
Models | Amazon Bedrock | us-e
Models | Amazon Bedrock | us-e
us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/custom-models/jobs/create?jobType=FINE_TUNING
aws
Services
Search
[Option+S]
United States (N. Virginia)
aws-courses
Amazon Bedrock
Getting started
Overview
Providers
Foundation models
Model catalog New
Marketplace deployments New
Custom models (fine-tuning, dist...)
Imported models
Prompt Routers Preview
Playgrounds
Chat / Text
Image / Video
Builder tools
Agents
Flows
Knowledge Bases
Prompt Management
Safeguards
Guardrails
Watermark detection
Amazon Bedrock > Custom models > Create Fine-tuning job
Create Fine-tuning job Info
Select the model you wish to fine-tune and submit your data location.
Model details
Source model
Choose from a list of models that you wish to customize with using your own data.
Select model
Fine-tuned model name
Enter a name to identify the new fine-tuned model.
Enter a name here
Model encryption Info
Tags - optional
Job configuration
Job name
Enter a name to identify the training job necessary to pre-train and create a new model.
Enter a name here
Tags - optional

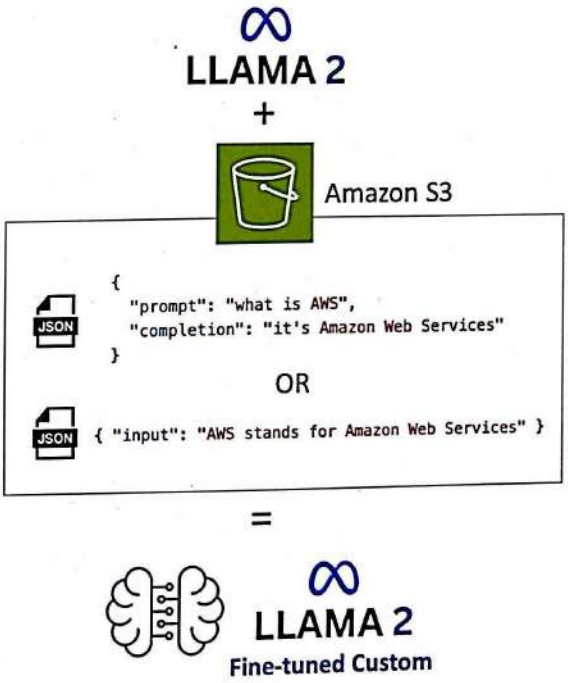

Amazon Bedrock
Getting started
Overview
Providers
Foundation models
Model catalog New
Marketplace deployments New
Custom models (fine-tuning, dist...)
Imported models
Prompt Routers Preview
Playgrounds
Chat / Text
Image / Video
Builder tools
Agents
Flows
Knowledge Bases
Prompt Management
Safeguards
Guardrails
Watermark detection
Select model
Search available models and inference
1. Categories
Model providers
Amazon
Cohere
2. Models
Nova Pro 1.0
Text & vision model
Nova Lite 1.0
Text & vision model
Nova Canvas 1.0
Image model
Nova Micro 1.0
Text model
Titan Text G1 - Lite v1
Text model
Titan Text G1 - Express v1
Text model
Titan Multimodal Embeddings G1 v1
Embedding model
Can't find the model you are looking for? See all models here.
Cancel
Apply


Providers | Amazon Bedrock | us
Model catalog | Claude 3.5 Haiku
Create Fine-tuning job | Amazon
Models | Amazon Bedrock | us-e
Models | Amazon Bedrock | us-e
us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/custom-models/jobs/create?jobType=FINE_TUNING&modelId=amazon.nova-micro-v1%3A0
aws
Services
Search
[Option+S]
United States (N. Virginia)
aws-courses
Amazon Bedrock
Getting started
Overview
Providers
Foundation models
Model catalog New
Marketplace deployments New
Custom models (fine-tuning, dist...)
Imported models
Prompt Routers Preview
Playgrounds
Chat / Text
Image / Video
Builder tools
Agents
Flows
Knowledge Bases
Prompt Management
Safeguards
Guardrails
Watermark detection
Input data Info
Choose a file in the S3 location. The files you choose must be in the dataset format that the model needs for training. You can check the data format for your specified model for any potential errors using simple python script. You can also use Sagemaker Ground Truth to create and label training datasets. Learn more
S3 location
s3://bucket/path-to-your-data/
View
Browse S3
Validation dataset S3 location (optional)
s3://bucket/path-to-your-data/
View
Browse S3
Hyperparameters Info
Epochs
The total number of iterations of all the training data in one cycle for training the model.
2
Enter an integer between 1 and 5.
Batch size
The number of samples processed before model parameters are updated.
1
Learning rate
The rate at which model parameters are updated after each batch of training data.
0.00001
Enter a float value between 0.000001 and 0.0001


Providers | Amazon Bedrock | us
Model catalog | Claude 3.5 Hai
Create Fine-tuning job | Amazon
Models | Amazon Bedrock | us-e
Models | Amazon Bedrock | us-e
us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/custom-models/jobs/create?jobType=FINE_TUNING&modelId=amazon.nova-micro-v1%3A0
aws
Services
Search
[Option+S]
United States (N. Virginia)
aws-courses
Amazon Bedrock
Getting started
Overview
Providers
Foundation models
Model catalog New
Marketplace deployments New
Custom models (fine-tuning, dist...)
Imported models
Prompt Routers Preview
Playgrounds
Chat / Text
Image / Video
Builder tools
Agents
Flows
Knowledge Bases
Prompt Management
Safeguards
Guardrails
Watermark detection
Output data Info
Choose S3 location to store the model validation outputs.
S3 location
s3://bucket/path-to-your-data/
View
Browse S3
Service access Info
Bedrock model customization job requires permissions to write to S3 on your behalf.
Choose a method to authorize Bedrock
Use an existing service role
Create and use a new service role
Service role
Choose role
Purchase provisioned throughput to use fine-tuned model
After this custom model is created, you need to purchase provisioned throughput to be able to use this model.
Learn more
Cancel
Create Fine-tuning job

Amazon Bedrock – Fine-Tuning a Model

- Adapt a copy of a foundation model with your own data
- Fine-tuning will change the weights of the base foundation model
- Training data must:
  - Adhere to a specific format
  - Be stored in Amazon S3
- You must use "Provisioned Throughput" to use a fine-tuned model
- Note: not all models can be fine-tuned



Instruction-based Fine Tuning

- Improves the performance of a pre-trained FM on domain-specific tasks
- = further trained on a particular field or area of knowledge
- Instruction-based fine-tuning uses labeled examples that are prompt-response pairs

Labeled Data

```
{
   "prompt": "Who is Stéphane Maarek?",
   "completion": "Stéphane Maarek is an
AWS instructor who dedicates his time to
make the best AWS courses so that his
students can pass all AWS certification
exams with flying color!"
}
```

Continued Pre-training

- Provide unlabeled data to continue the training of an FM
- Also called domain-adaptation fine-tuning, to make a model expert in a specific domain
- For example: feeding the entire AWS documentation to a model to make it an expert on AWS
- Good to feed industry-specific terminology into a model (acronyms, etc...)
- Can continue to train the model as more data becomes available

```
{
   "input": "Our CTA (Commodity Trading
Advisor) strategy incorporates a blend of
momentum and mean reversion algorithms,
optimized through a rolling window
backtesting methodology. The trading signals
are generated by analyzing historical price
data with a focus on Sharpe ratios and
drawdown limits. We utilize HFT (High-
Frequency Trading) systems to capitalize on
short-term price inefficiencies across
various asset classes, including
commodities, forex, and equity index
futures."
}
```

Single-Turn Messaging

- Part of instruction-based fine-tuning
- system (optional) : context for the conversation.
- messages : An array of message objects, each containing:
- role : Either user or assistant
- content : The text content of the message

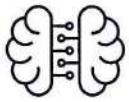
```
{
  "system": "You are an helpful assistant.",
  "messages": [
      {
         "role": "user",
         "content": "what is AWS"
      },
      {
        "role": "assistant",
    "content": "it's Amazon Web Services."
      }
   ]
}
```

Multi-Turn Messaging

- To provide instruction-based fine tuning for a conversation (vs Single-Turn Messaging)
- Chatbots = multi-turn environment
- You must alternate between "user" and "assistant" roles

```
{
"system": "You are an AI assistant specializing in AWS
services.",
  "messages": [
   {  "role": "user",      "content": "Tell me about
Amazon SageMaker." },
   {  "role": "assistant", "content": "Amazon SageMaker is
a fully managed service for building, training, and
deploying machine learning models at scale." },
   {  "role": "user",      "content": "How does it
integrate with other AWS services?" },
   {  "role": "assistant", "content": "SageMaker
integrates with AWS services like S3 for data storage,
Lambda for event-driven computing, and CloudWatch for
monitoring." }
   ]
}
```

Fine-Tuning: good to know



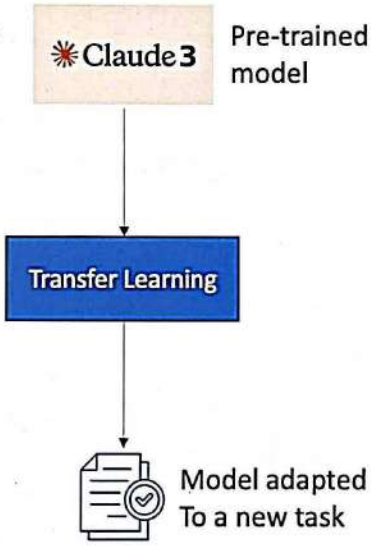
- Re-training an FM requires a higher budget
- Instruction-based fine-tuning is usually cheaper as computations are less intense and the amount of data required usually less

- It also requires experienced ML engineers to perform the task
- You must prepare the data, do the fine-tuning, evaluate the model
- Running a fine-tuned model is also more expensive (provisioned throughput)

Note: Transfer Learning

- **Transfer Learning** – the broader concept of re-using a pre-trained model to adapt it to a new related task
  - Widely used for image classification
  - And for NLP (models like BERT and GPT)

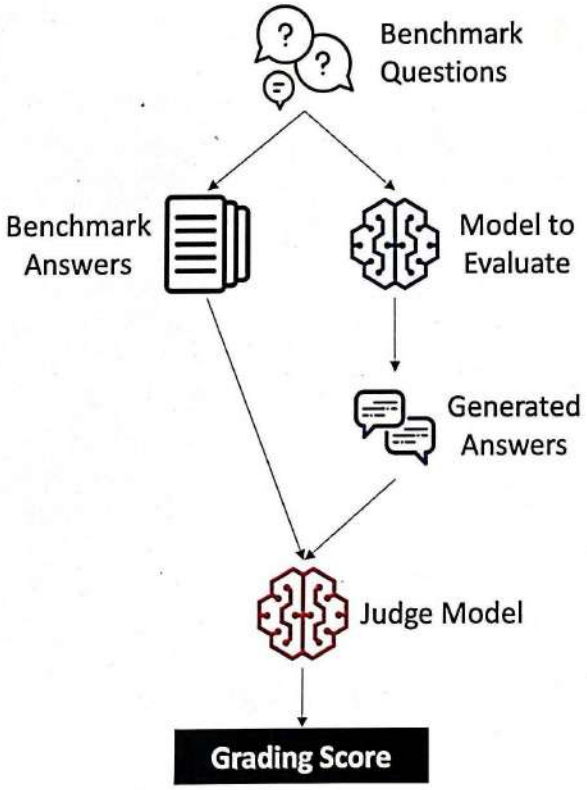
- Can appear in the exam as a general ML concept
- Fine-tuning is a specific kind of transfer learning



Amazon Bedrock – Evaluating a Model

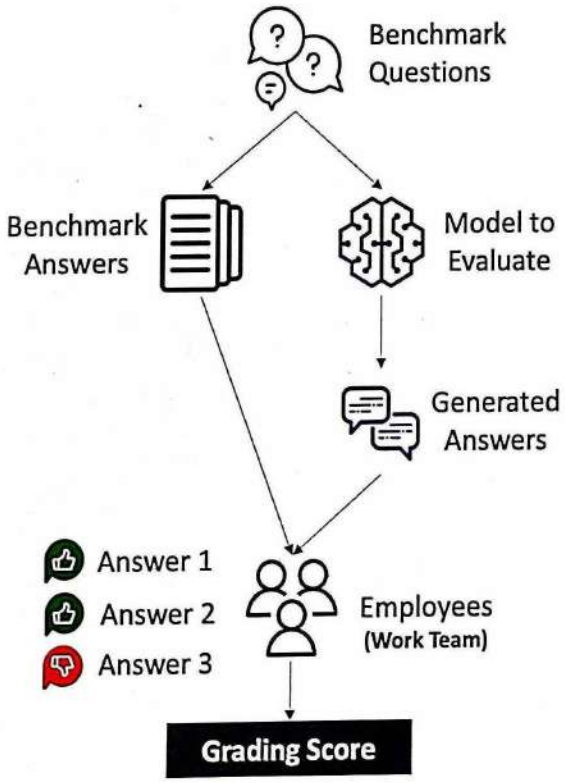
Automatic Evaluation

- Evaluate a model for quality control
- Built-in task types:
  - Text summarization
  - question and answer
  - text classification
  - open-ended text generation…
- Bring your own prompt dataset or use built-in curated prompt datasets
- Scores are calculated automatically
- Model scores are calculated using various statistical methods (e.g. BERTScore, F1…)



Amazon Bedrock – Evaluating a Model
Human Evaluation

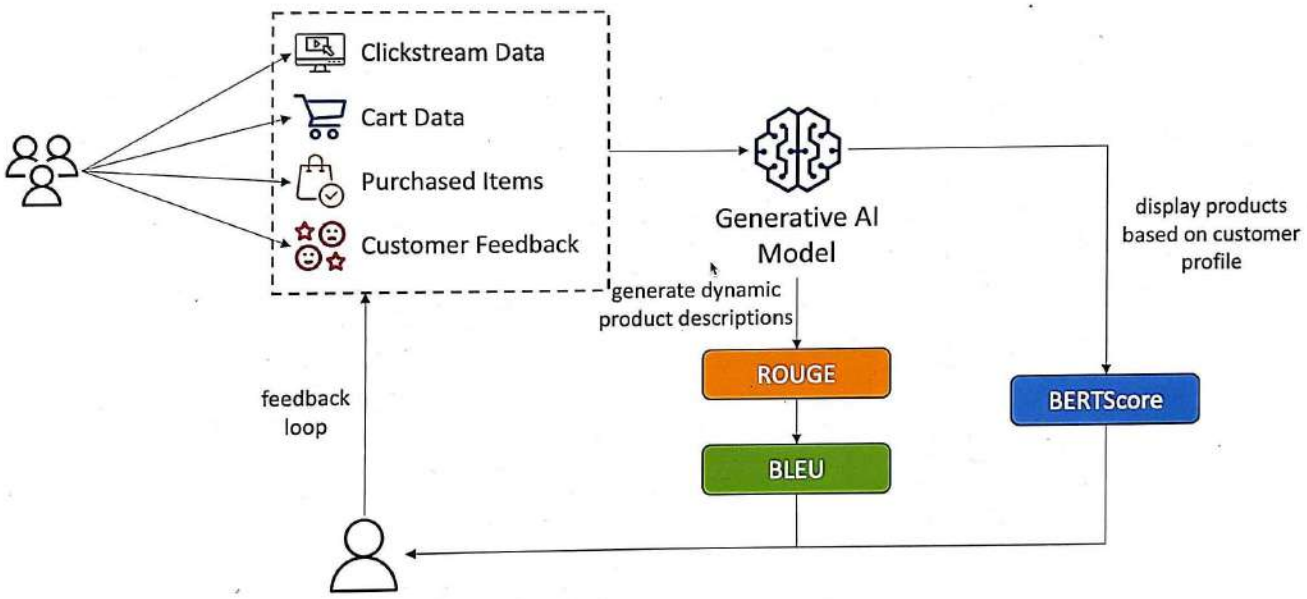
- Choose a work team to evaluate
  - Employees of your company
  - Subject-Matter Experts (SMEs)
- Define metrics and how to evaluate
  - Thumbs up/down, ranking…



Automated Metrics to Evaluate an FM

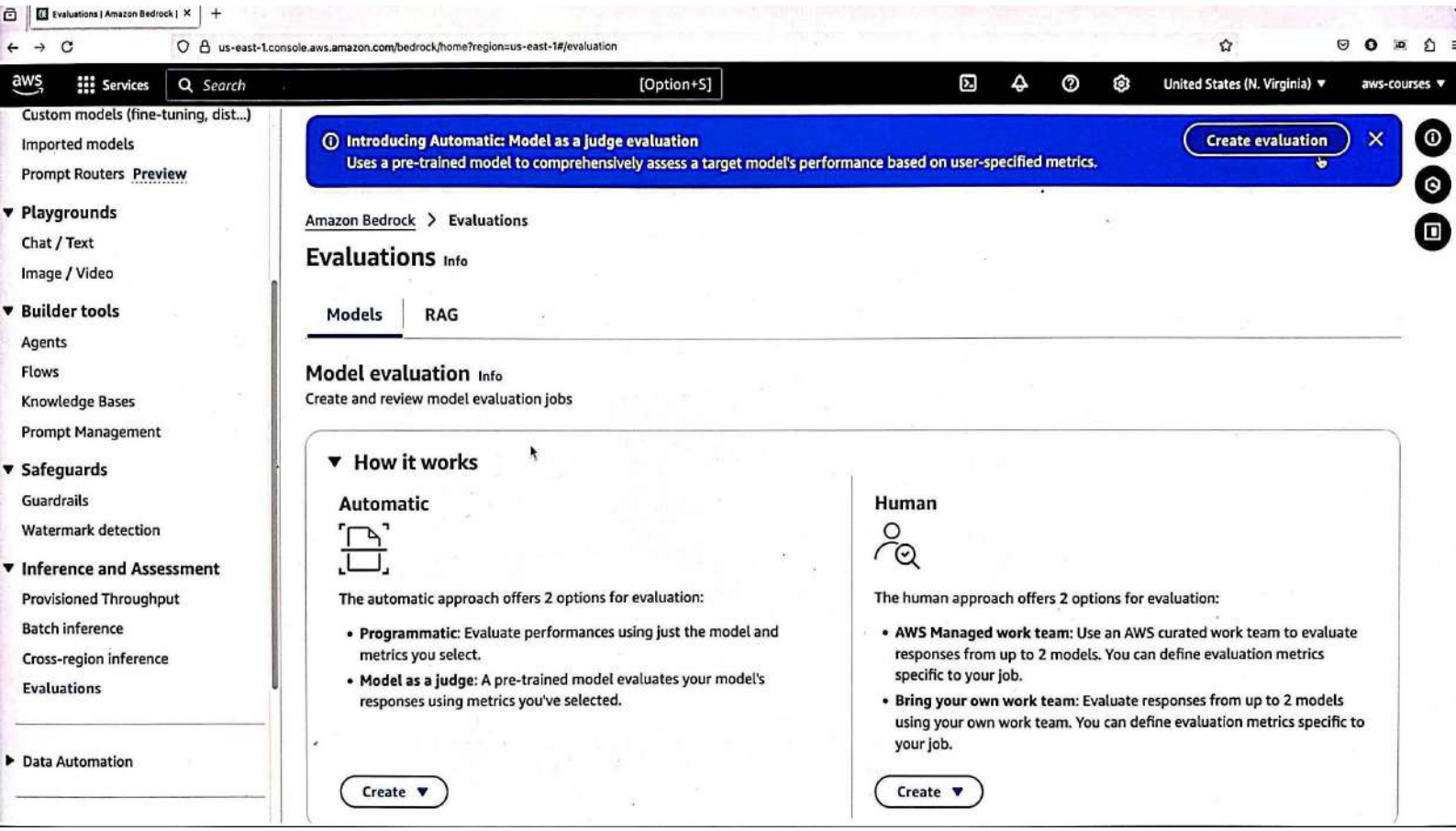
- ROUGE: Recall-Oriented Understudy for Gisting Evaluation
  - Evaluating automatic summarization and machine translation systems
  - ROUGE-N – measure the number of matching n-grams between reference and generated text
  - ROUGE-L – longest common subsequence between reference and generated text
- BLEU: Bilingual Evaluation Understudy
  - Evaluate the quality of generated text, especially for translations
  - Considers both precision and penalizes too much brevity
  - Looks at a combination of n-grams (1, 2, 3, 4)
- BERTScore
  - Semantic similarity between generated text
  - Uses pre-trained BERT models (Bidirectional Encoder Representations from Transformers) to compare the contextualized embeddings of both texts and computes the cosine similarity between them.
  - Capable of capturing more nuance between the texts
- Perplexity: how well the model predicts the next token (lower is better)

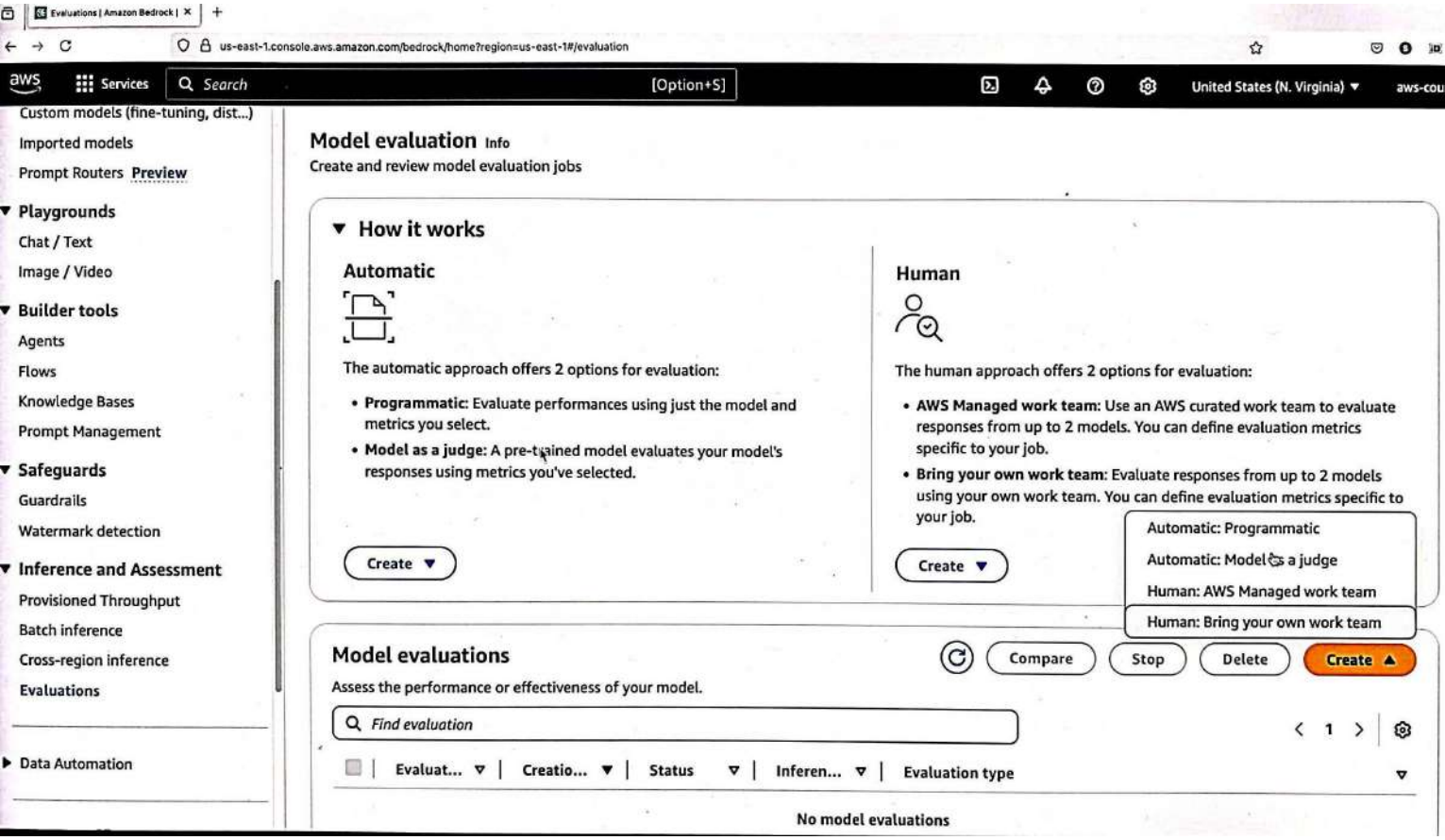
Automated Model Evaluation

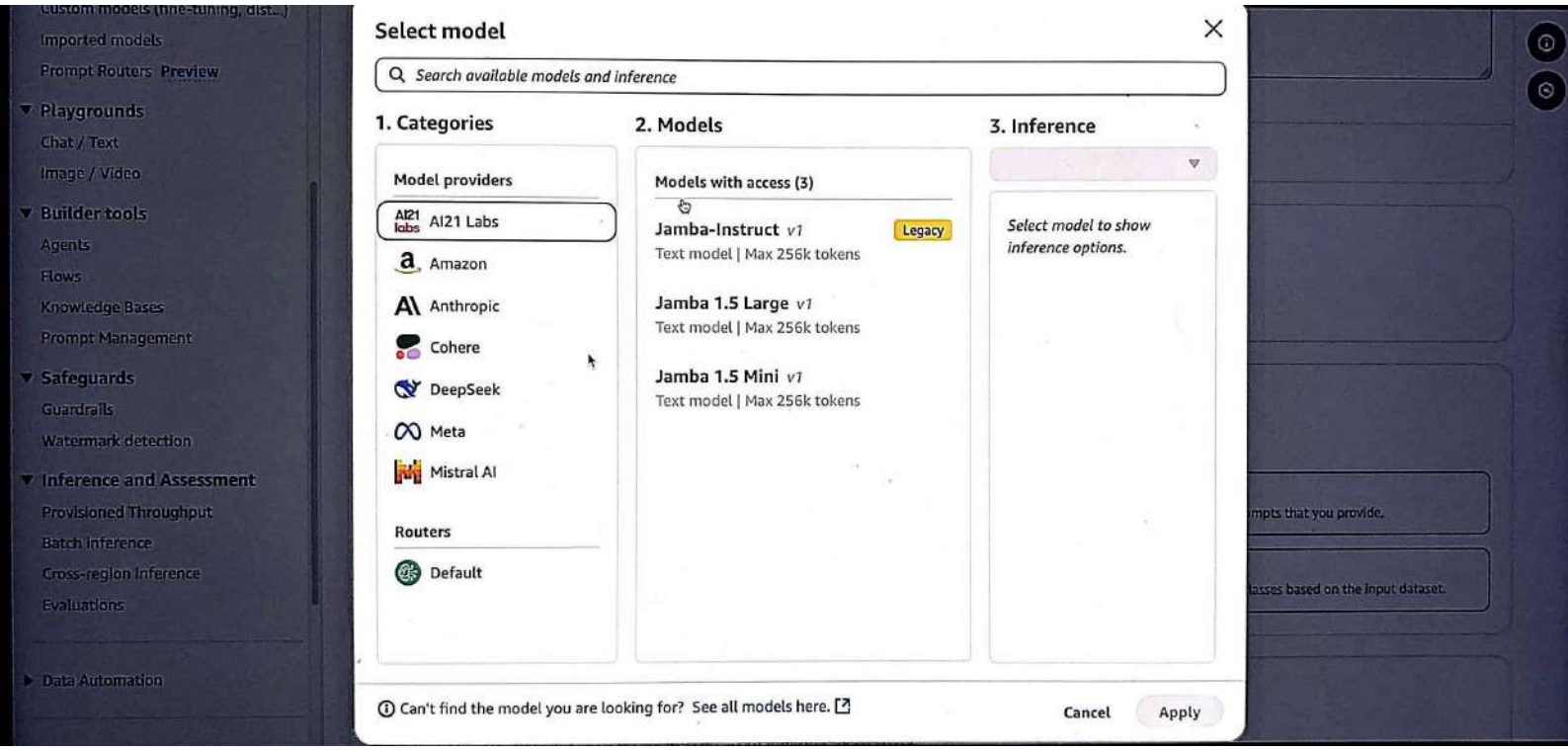


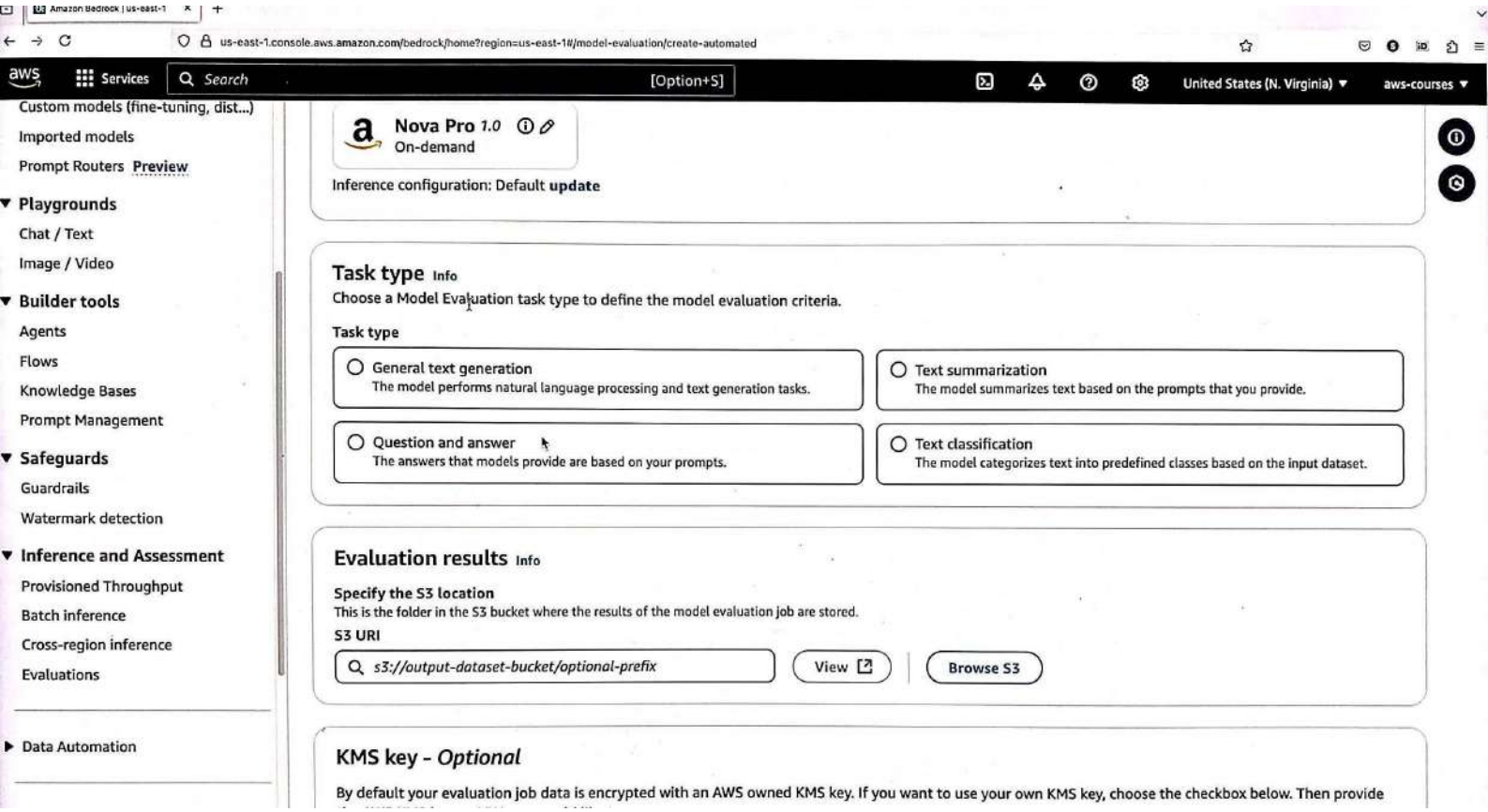
Business Metrics to Evaluate a Model On

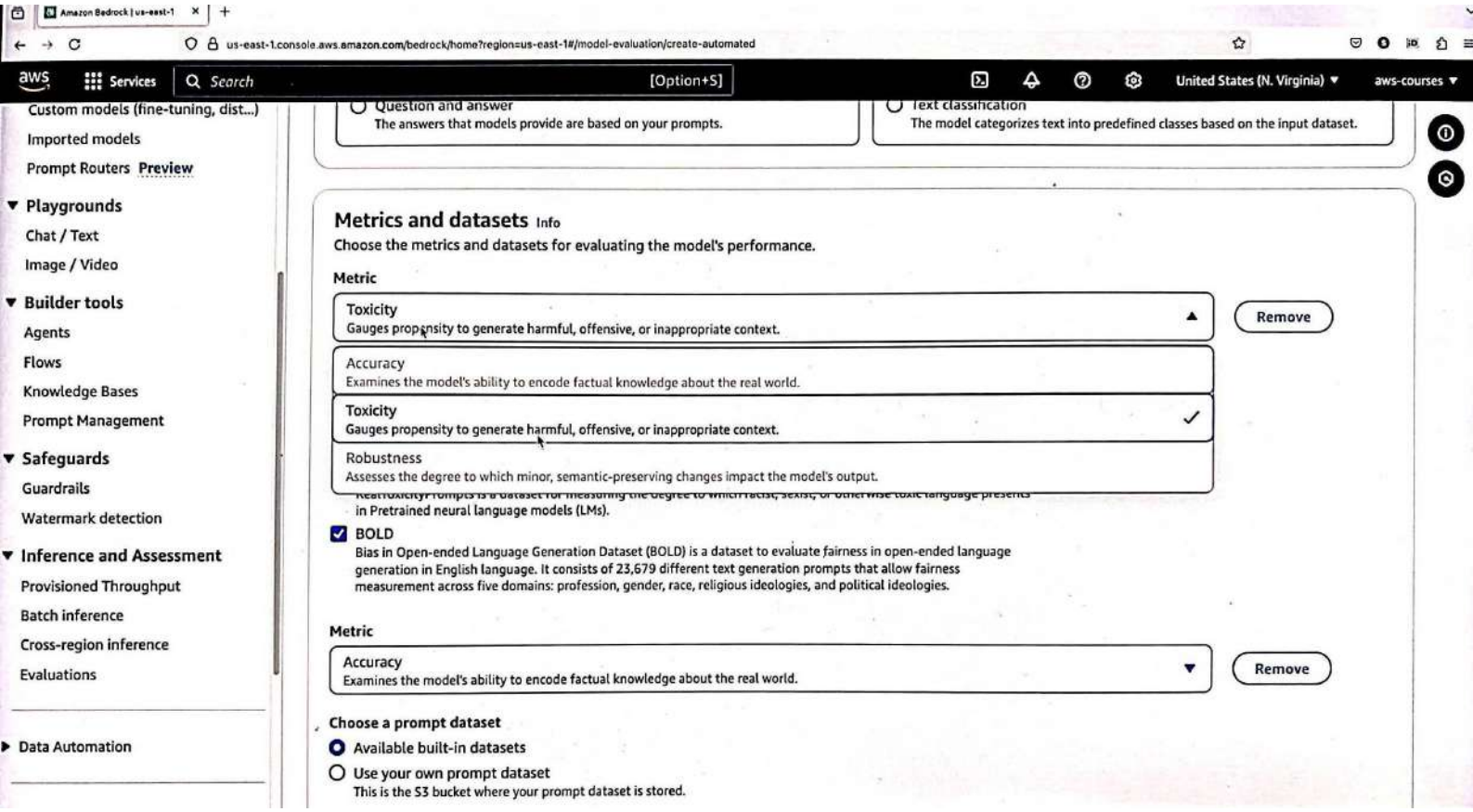
- User Satisfaction – gather users' feedbacks and assess their satisfaction with the model responses (e.g., user satisfaction for an ecommerce platform)
- Average Revenue Per User (ARPU) – average revenue per user attributed to the Gen-AI app (e.g., monitor ecommerce user base revenue)
- Cross-Domain Performance – measure the model's ability to perform cross different domains tasks (e.g., monitor multi-domain ecommerce platform)
- Conversion Rate – generate recommended desired outcomes such as purchases (e.g., optimizing ecommerce platform for higher conversion rate)
- Efficiency – evaluate the model's efficiency in computation, resource utilization… (e.g., improve production line efficiency)

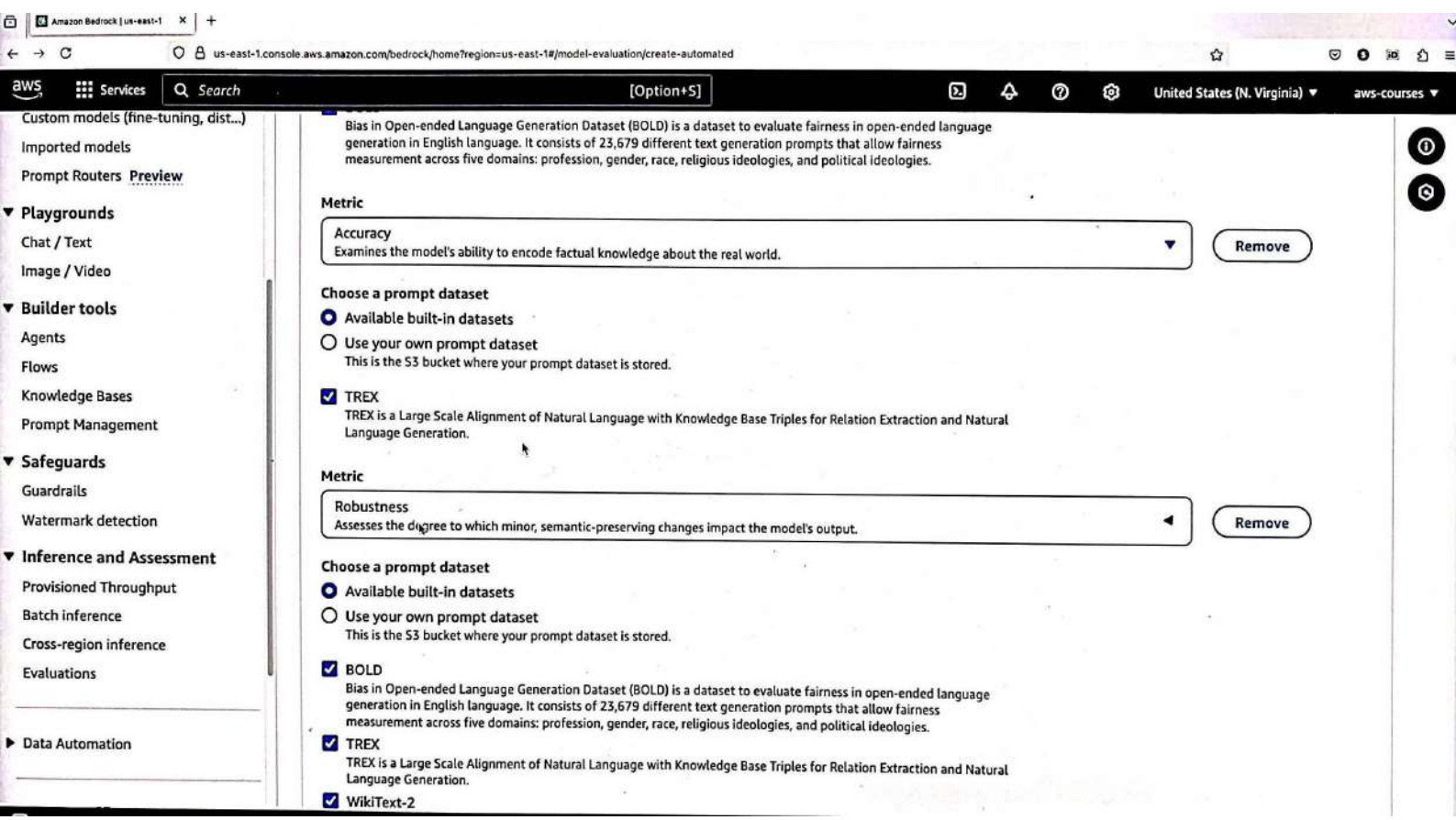

Evaluations | Amazon Bedrock |
us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/evaluation
aws
Services
Search
[Option+S]
United States (N. Virginia)
aws-courses
Custom models (fine-tuning, dist...)
Imported models
Prompt Routers Preview
Playgrounds
Chat / Text
Image / Video
Builder tools
Agents
Flows
Knowledge Bases
Prompt Management
Safeguards
Guardrails
Watermark detection
Inference and Assessment
Provisioned Throughput
Batch inference
Cross-region inference
Evaluations
Data Automation
Introducing Automatic: Model as a judge evaluation
Uses a pre-trained model to comprehensively assess a target model's performance based on user-specified metrics.
Create evaluation
Amazon Bedrock > Evaluations
Evaluations Info
Models
RAG
Model evaluation Info
Create and review model evaluation jobs
How it works
Automatic
The automatic approach offers 2 options for evaluation:
Programmatic: Evaluate performances using just the model and metrics you select.
Model as a judge: A pre-trained model evaluates your model's responses using metrics you've selected.
Create
Human
The human approach offers 2 options for evaluation:
AWS Managed work team: Use an AWS curated work team to evaluate responses from up to 2 models. You can define evaluation metrics specific to your job.
Bring your own work team: Evaluate responses from up to 2 models using your own work team. You can define evaluation metrics specific to your job.
Create


Evaluations | Amazon Bedrock |
us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/evaluation
aws
Services
Search
[Option+S]
United States (N. Virginia)
aws-courses
Custom models (fine-tuning, dist...)
Imported models
Prompt Routers Preview
Playgrounds
Chat / Text
Image / Video
Builder tools
Agents
Flows
Knowledge Bases
Prompt Management
Safeguards
Guardrails
Watermark detection
Inference and Assessment
Provisioned Throughput
Batch inference
Cross-region inference
Evaluations
Data Automation
Model evaluation Info
Create and review model evaluation jobs
How it works
Automatic
The automatic approach offers 2 options for evaluation:
Programmatic: Evaluate performances using just the model and metrics you select.
Model as a judge: A pre-trained model evaluates your model's responses using metrics you've selected.
Create
Human
The human approach offers 2 options for evaluation:
AWS Managed work team: Use an AWS curated work team to evaluate responses from up to 2 models. You can define evaluation metrics specific to your job.
Bring your own work team: Evaluate responses from up to 2 models using your own work team. You can define evaluation metrics specific to your job.
Create
Automatic: Programmatic
Automatic: Model as a judge
Human: AWS Managed work team
Human: Bring your own work team
Model evaluations
Compare
Stop
Delete
Create
Assess the performance or effectiveness of your model.
Find evaluation
1
Evaluat...
Creatio...
Status
Inferen...
Evaluation type
No model evaluations


Custom models (fine-tuning, dist...)
Imported models
Prompt Routers Preview
Playgrounds
Chat / Text
Image / Video
Builder tools
Agents
Flows
Knowledge Bases
Prompt Management
Safeguards
Guardrails
Watermark detection
Inference and Assessment
Provisioned Throughput
Batch inference
Cross-region inference
Evaluations
Data Automation
Select model
Search available models and inference
1. Categories
Model providers
AI21 Labs
Amazon
Anthropic
Cohere
DeepSeek
Meta
Mistral AI
Routers
Default
2. Models
Models with access (3)
Jamba-Instruct v1
Legacy
Text model | Max 256k tokens
Jamba 1.5 Large v1
Text model | Max 256k tokens
Jamba 1.5 Mini v1
Text model | Max 256k tokens
3. Inference
Select model to show inference options.
Can't find the model you are looking for? See all models here.
Cancel
Apply
mpts that you provide.
asses based on the input dataset.

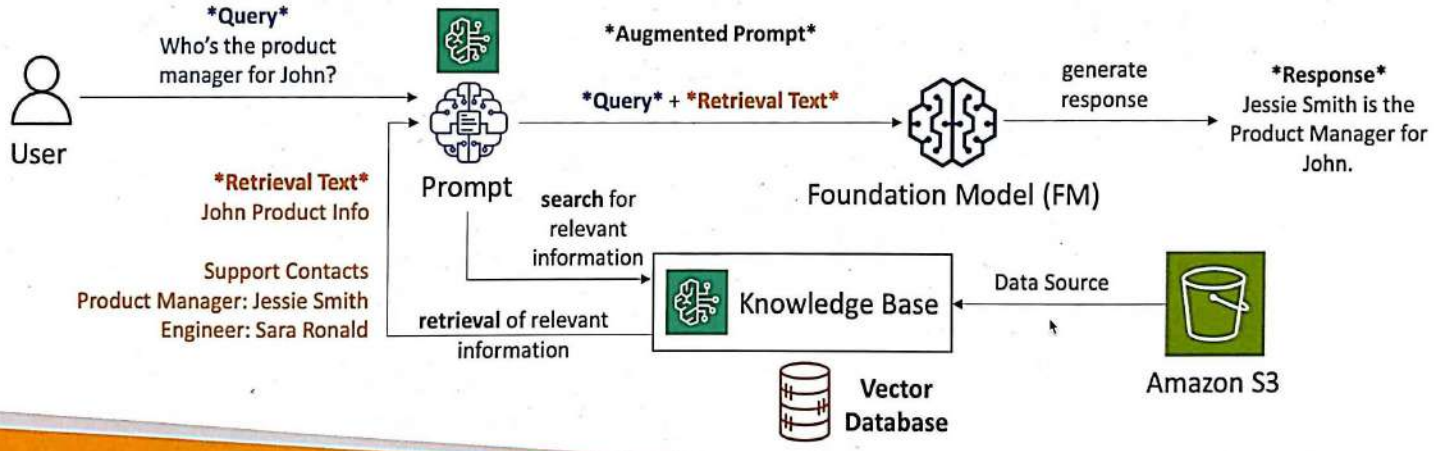

Amazon Bedrock | us-east-1
us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/model-evaluation/create-automated
aws
Services
Search
[Option+S]
United States (N. Virginia)
aws-courses
Custom models (fine-tuning, dist...)
Imported models
Prompt Routers Preview
Playgrounds
Chat / Text
Image / Video
Builder tools
Agents
Flows
Knowledge Bases
Prompt Management
Safeguards
Guardrails
Watermark detection
Inference and Assessment
Provisioned Throughput
Batch inference
Cross-region inference
Evaluations
Data Automation
Nova Pro 1.0
On-demand
Inference configuration: Default update
Task type Info
Choose a Model Evaluation task type to define the model evaluation criteria.
Task type
General text generation
The model performs natural language processing and text generation tasks.
Text summarization
The model summarizes text based on the prompts that you provide.
Question and answer
The answers that models provide are based on your prompts.
Text classification
The model categorizes text into predefined classes based on the input dataset.
Evaluation results Info
Specify the S3 location
This is the folder in the S3 bucket where the results of the model evaluation job are stored.
S3 URI
s3://output-dataset-bucket/optional-prefix
View
Browse S3
KMS key - Optional
By default your evaluation job data is encrypted with an AWS owned KMS key. If you want to use your own KMS key, choose the checkbox below. Then provide


Amazon Bedrock | us-east-1
us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/model-evaluation/create-automated
aws
Services
Search
[Option+S]
United States (N. Virginia)
aws-courses
Custom models (fine-tuning, dist...)
Imported models
Prompt Routers Preview
Playgrounds
Chat / Text
Image / Video
Builder tools
Agents
Flows
Knowledge Bases
Prompt Management
Safeguards
Guardrails
Watermark detection
Inference and Assessment
Provisioned Throughput
Batch inference
Cross-region inference
Evaluations
Data Automation
Question and answer
The answers that models provide are based on your prompts.
Text classification
The model categorizes text into predefined classes based on the input dataset.
Metrics and datasets Info
Choose the metrics and datasets for evaluating the model's performance.
Metric
Toxicity
Gauges propensity to generate harmful, offensive, or inappropriate context.
Remove
Accuracy
Examines the model's ability to encode factual knowledge about the real world.
Toxicity
Gauges propensity to generate harmful, offensive, or inappropriate context.
Robustness
Assesses the degree to which minor, semantic-preserving changes impact the model's output.
in Pretrained neural language models (LMs).
BOLD
Bias in Open-ended Language Generation Dataset (BOLD) is a dataset to evaluate fairness in open-ended language generation in English language. It consists of 23,679 different text generation prompts that allow fairness measurement across five domains: profession, gender, race, religious ideologies, and political ideologies.
Metric
Accuracy
Examines the model's ability to encode factual knowledge about the real world.
Remove
Choose a prompt dataset
Available built-in datasets
Use your own prompt dataset
This is the S3 bucket where your prompt dataset is stored.

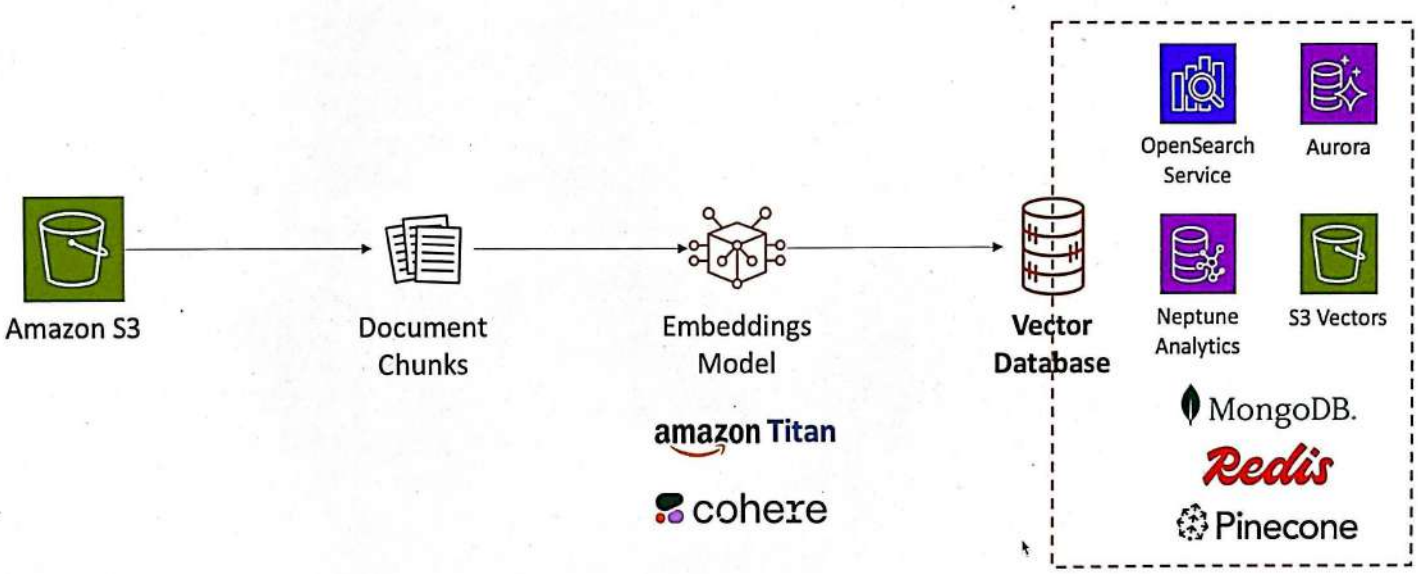

Amazon Bedrock | us-east-1
us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/model-evaluation/create-automated
aws
Services
Search
[Option+S]
United States (N. Virginia)
aws-courses
Custom models (fine-tuning, dist...)
Imported models
Prompt Routers Preview
Playgrounds
Chat / Text
Image / Video
Builder tools
Agents
Flows
Knowledge Bases
Prompt Management
Safeguards
Guardrails
Watermark detection
Inference and Assessment
Provisioned Throughput
Batch inference
Cross-region inference
Evaluations
Data Automation
Bias in Open-ended Language Generation Dataset (BOLD) is a dataset to evaluate fairness in open-ended language generation in English language. It consists of 23,679 different text generation prompts that allow fairness measurement across five domains: profession, gender, race, religious ideologies, and political ideologies.
Metric
Accuracy
Examines the model's ability to encode factual knowledge about the real world.
Remove
Choose a prompt dataset
Available built-in datasets
Use your own prompt dataset
This is the S3 bucket where your prompt dataset is stored.
TREX
TREX is a Large Scale Alignment of Natural Language with Knowledge Base Triples for Relation Extraction and Natural Language Generation.
Metric
Robustness
Assesses the degree to which minor, semantic-preserving changes impact the model's output.
Remove
Choose a prompt dataset
Available built-in datasets
Use your own prompt dataset
This is the S3 bucket where your prompt dataset is stored.
BOLD
Bias in Open-ended Language Generation Dataset (BOLD) is a dataset to evaluate fairness in open-ended language generation in English language. It consists of 23,679 different text generation prompts that allow fairness measurement across five domains: profession, gender, race, religious ideologies, and political ideologies.
TREX
TREX is a Large Scale Alignment of Natural Language with Knowledge Base Triples for Relation Extraction and Natural Language Generation.
WikiText-2

Amazon Bedrock – RAG & Knowledge Base

- RAG = Retrieval-Augmented Generation
- Allows a Foundation Model to reference a data source outside of its training data
- Bedrock takes care of creating Vector Embeddings in the database of your choice based on your data



Amazon Bedrock – RAG Vector Databases



RAG Vector Databases by AWS

- Amazon OpenSearch Service (Serverless & Managed Cluster) search & analytics database real time similarity queries, store millions of vector embeddings scalable index management, and fast nearest-neighbor (kNN) search capability

- Amazon Aurora PostgreSQL – relational database, proprietary on AWS

- Amazon Neptune Analytics – graph database that enables high performance graph analytics and graph-based RAG (GraphRAG) solutions

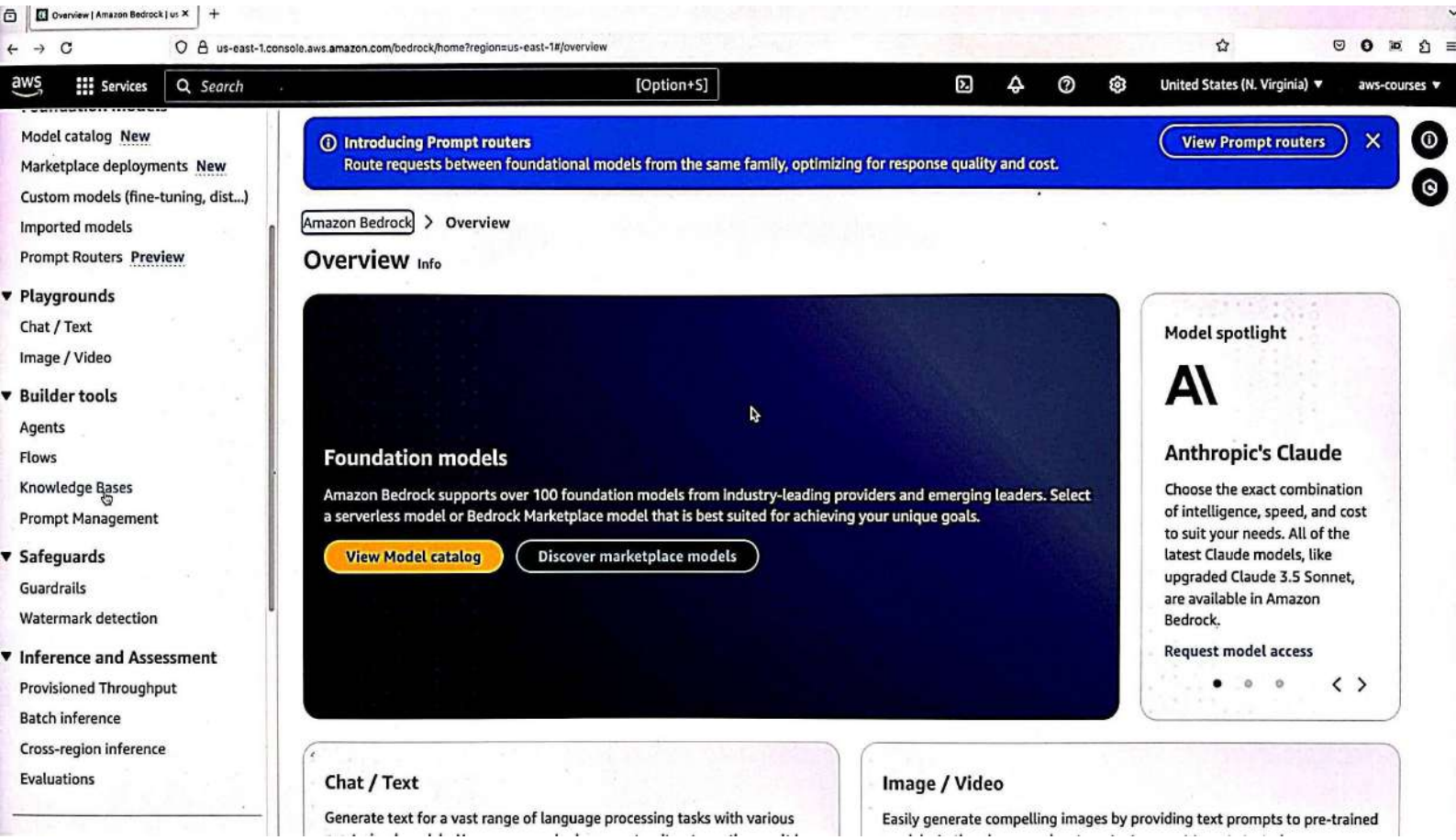
- Amazon S3 Vectors – cost-effective and durable storage with sub-second query performance

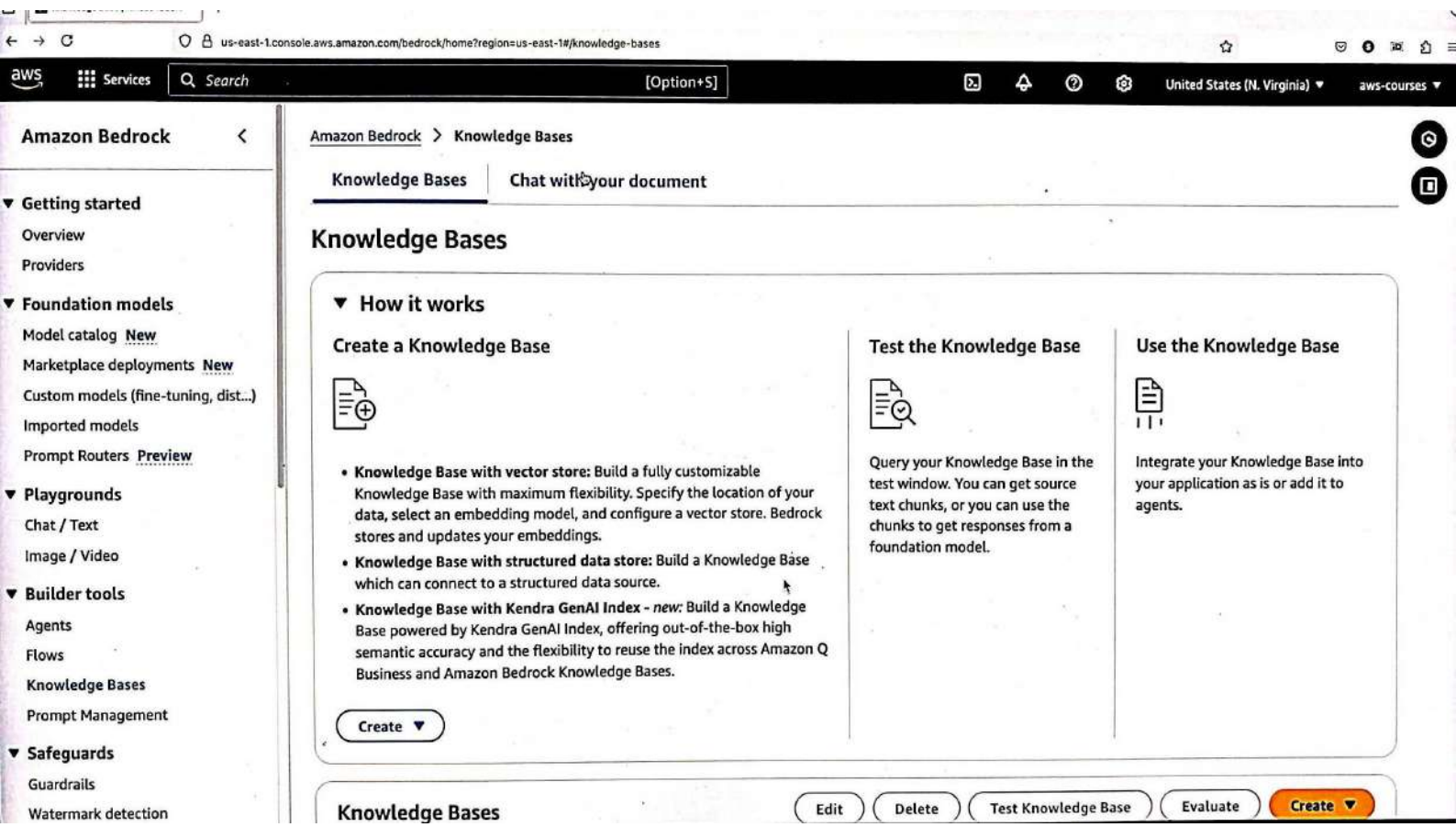
Amazon Bedrock – RAG Data Sources

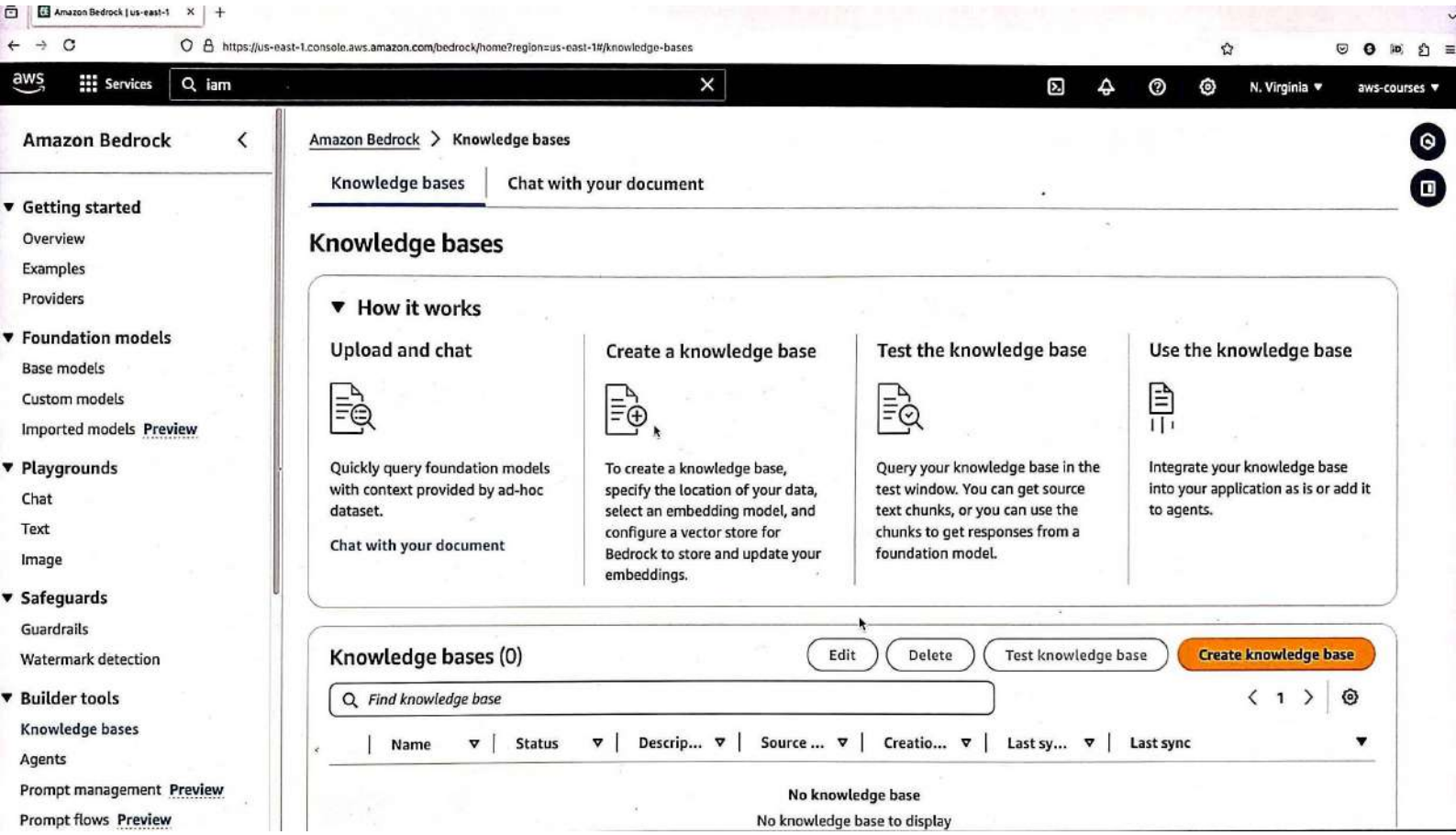
- Amazon S3
- Confluence
- Microsoft SharePoint
- Salesforce
- Web pages (your website, your social media feed, etc…)
- *More added over time...*

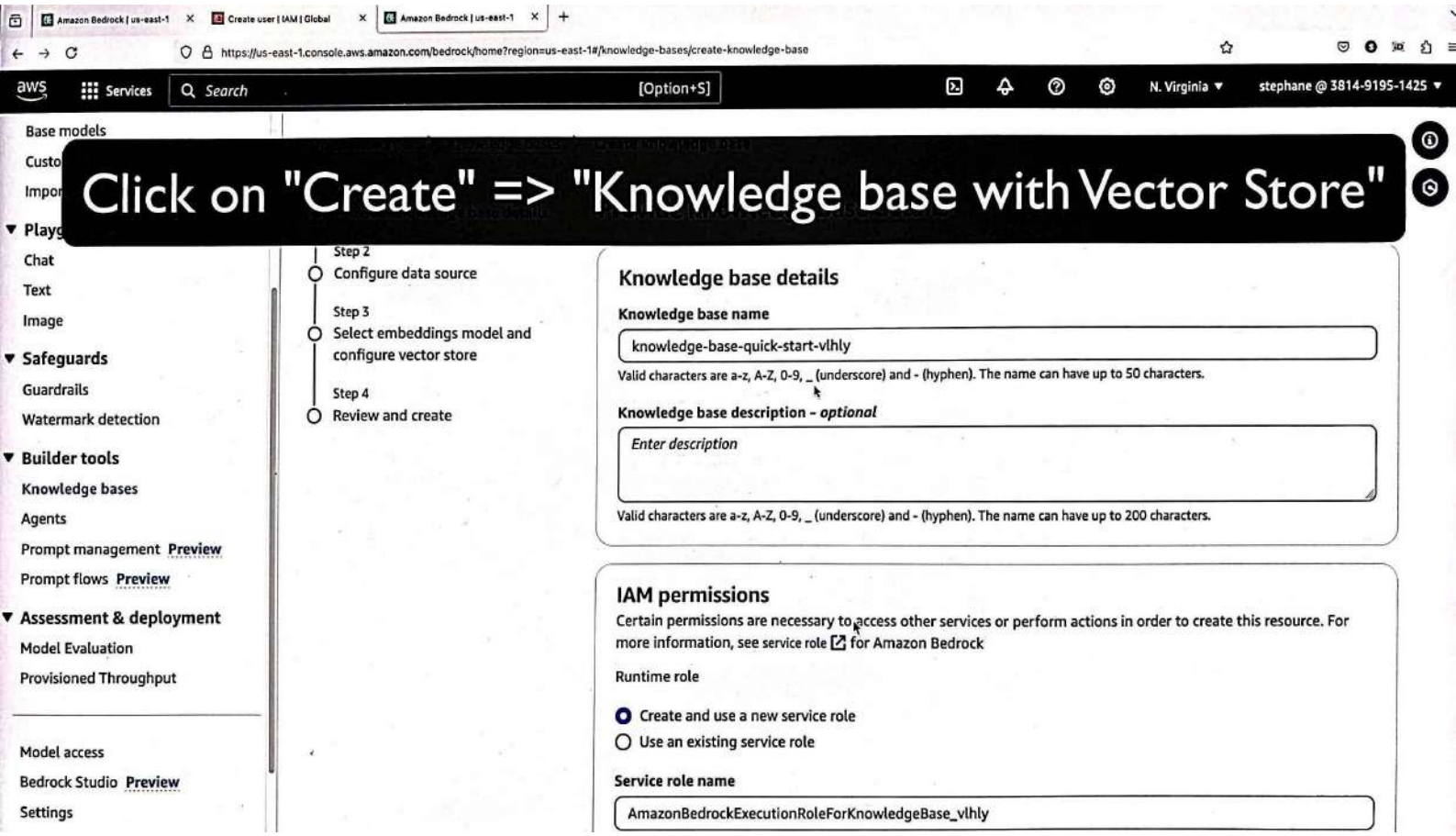
Amazon Bedrock – RAG – Use Cases

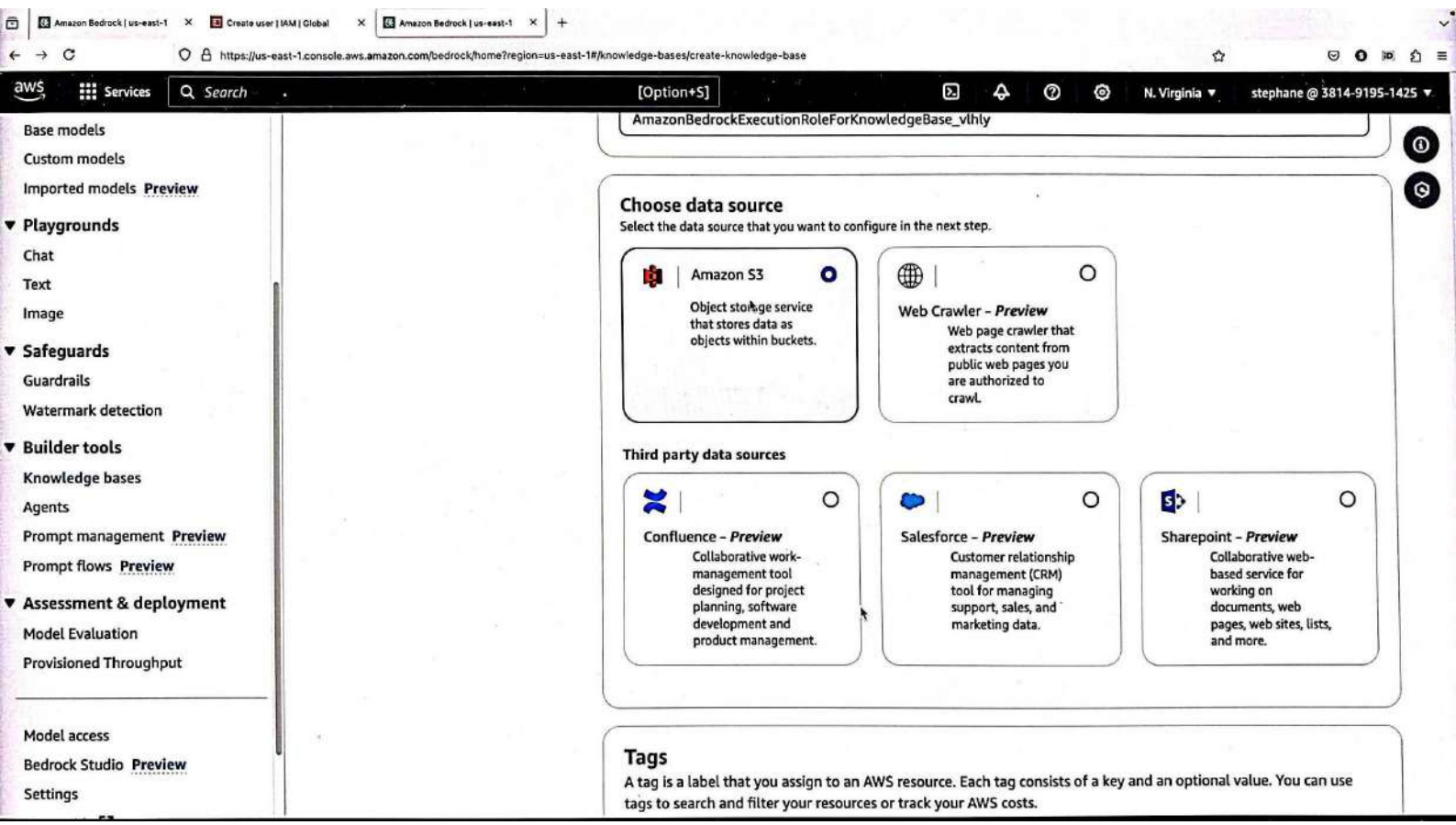
- Customer Service Chatbot
  - Knowledge Base – products, features, specifications, troubleshooting guides, and FAQs
  - RAG application – chatbot that can answer customer queries
- Legal Research and Analysis
  - Knowledge Base – laws, regulations, case precedents, legal opinions, and expert analysis
  - RAG Application – chatbot that can provide relevant information for specific legal queries
- Healthcare Question-Answering
  - Knowledge base – diseases, treatments, clinical guidelines, research papers, patients…
  - RAG application – chatbot that can answer complex medical queries

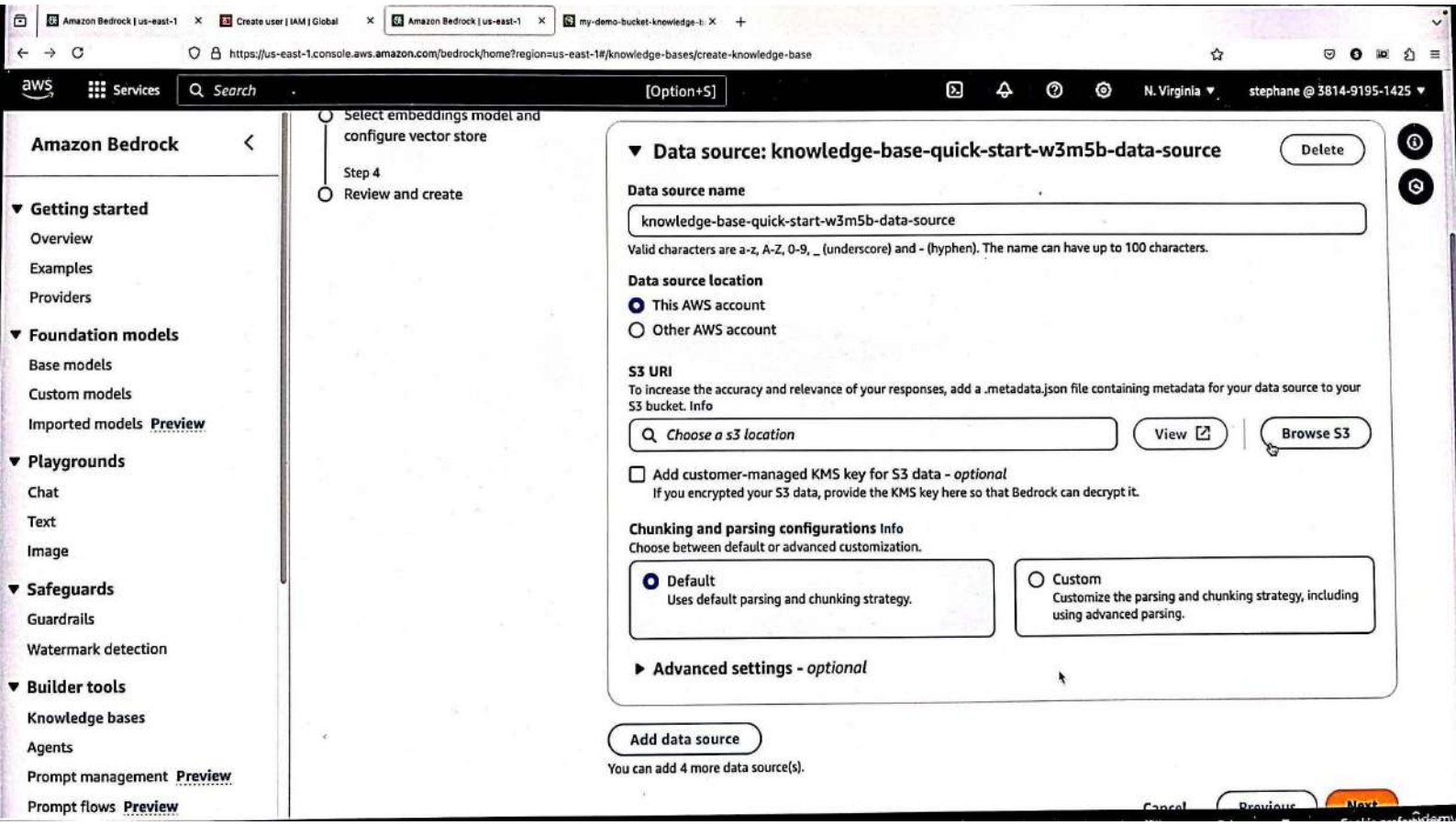

Overview | Amazon Bedrock | us
us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/overview
aws
Services
Search
[Option+S]
United States (N. Virginia)
aws-courses
Model catalog New
Marketplace deployments New
Custom models (fine-tuning, dist...)
Imported models
Prompt Routers Preview
Playgrounds
Chat / Text
Image / Video
Builder tools
Agents
Flows
Knowledge Bases
Prompt Management
Safeguards
Guardrails
Watermark detection
Inference and Assessment
Provisioned Throughput
Batch inference
Cross-region inference
Evaluations
Introducing Prompt routers
Route requests between foundational models from the same family, optimizing for response quality and cost.
View Prompt routers
Amazon Bedrock
Overview
Overview Info
Foundation models
Amazon Bedrock supports over 100 foundation models from industry-leading providers and emerging leaders. Select a serverless model or Bedrock Marketplace model that is best suited for achieving your unique goals.
View Model catalog
Discover marketplace models
Model spotlight
Anthropic's Claude
Choose the exact combination of intelligence, speed, and cost to suit your needs. All of the latest Claude models, like upgraded Claude 3.5 Sonnet, are available in Amazon Bedrock.
Request model access
Chat / Text
Generate text for a vast range of language processing tasks with various
Image / Video
Easily generate compelling images by providing text prompts to pre-trained

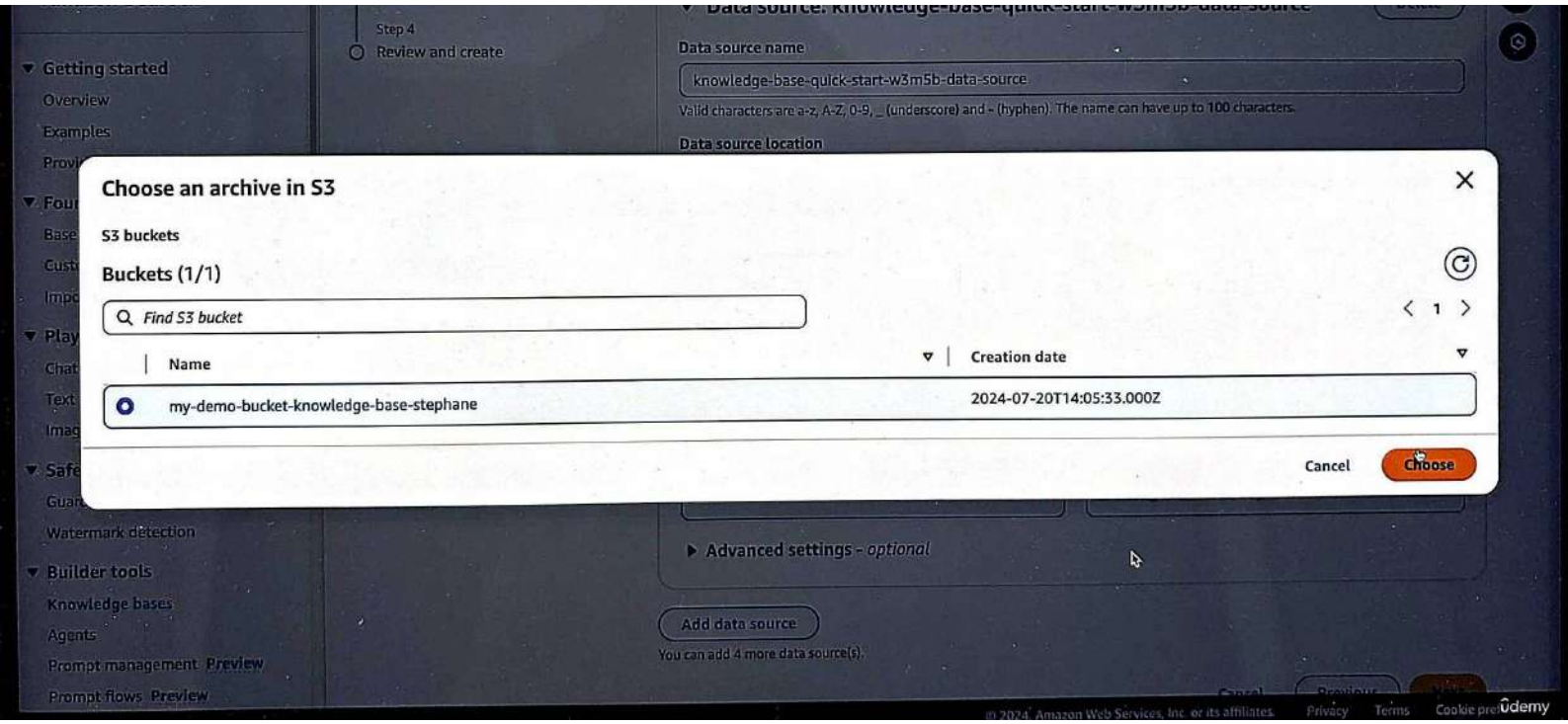

us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/knowledge-bases
aws
Services
Search
[Option+S]
United States (N. Virginia)
aws-courses
Amazon Bedrock
Getting started
Overview
Providers
Foundation models
Model catalog New
Marketplace deployments New
Custom models (fine-tuning, dist...)
Imported models
Prompt Routers Preview
Playgrounds
Chat / Text
Image / Video
Builder tools
Agents
Flows
Knowledge Bases
Prompt Management
Safeguards
Guardrails
Watermark detection
Amazon Bedrock > Knowledge Bases
Knowledge Bases
Chat with your document
Knowledge Bases
How it works
Create a Knowledge Base
Knowledge Base with vector store: Build a fully customizable Knowledge Base with maximum flexibility. Specify the location of your data, select an embedding model, and configure a vector store. Bedrock stores and updates your embeddings.
Knowledge Base with structured data store: Build a Knowledge Base which can connect to a structured data source.
Knowledge Base with Kendra GenAI Index - new: Build a Knowledge Base powered by Kendra GenAI Index, offering out-of-the-box high semantic accuracy and the flexibility to reuse the index across Amazon Q Business and Amazon Bedrock Knowledge Bases.
Create
Test the Knowledge Base
Query your Knowledge Base in the test window. You can get source text chunks, or you can use the chunks to get responses from a foundation model.
Use the Knowledge Base
Integrate your Knowledge Base into your application as is or add it to agents.
Knowledge Bases
Edit
Delete
Test Knowledge Base
Evaluate
Create

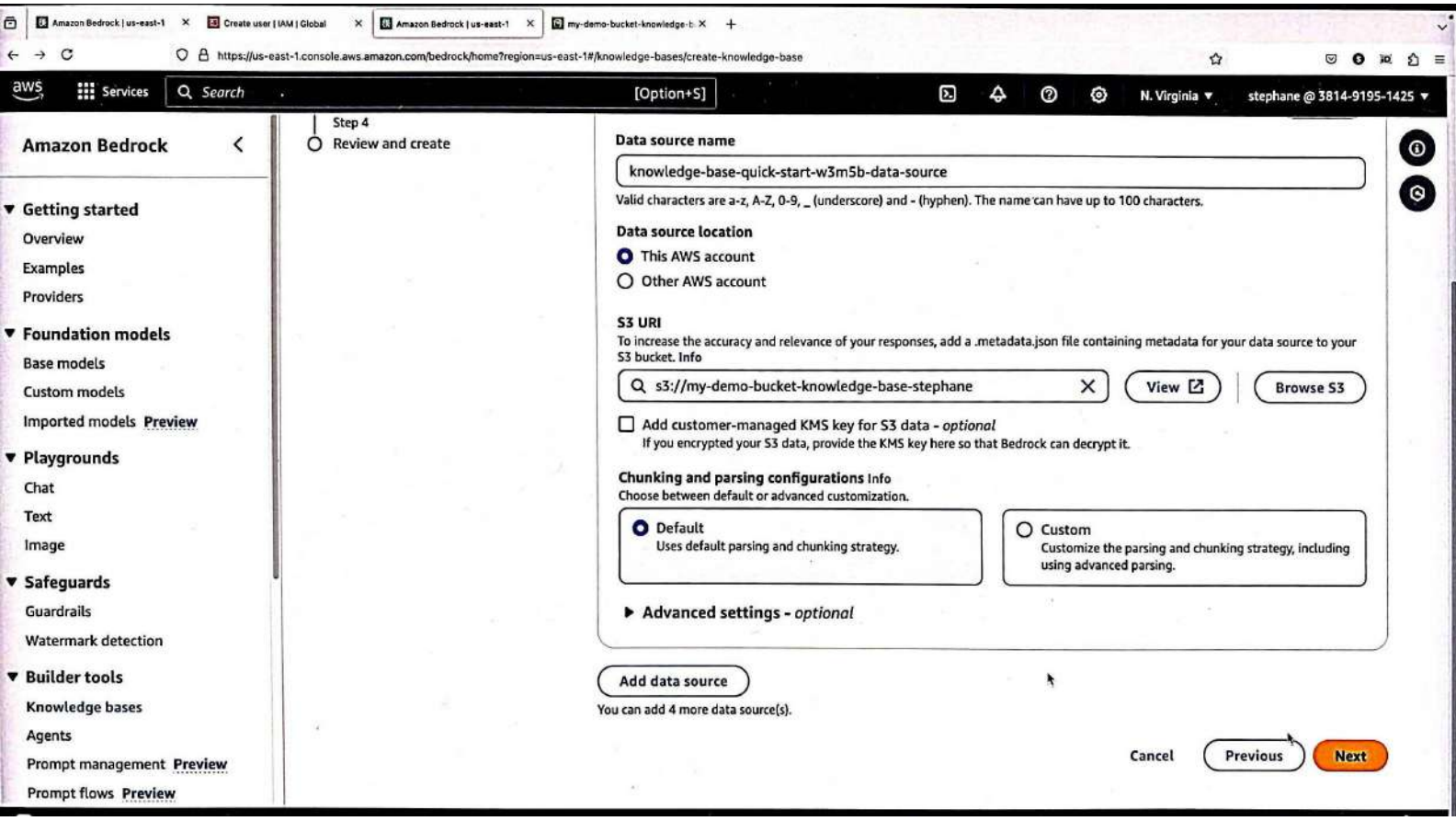

Amazon Bedrock | us-east-1
https://us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/knowledge-bases
aws
Services
iam
N. Virginia
aws-courses
Amazon Bedrock
Getting started
Overview
Examples
Providers
Foundation models
Base models
Custom models
Imported models Preview
Playgrounds
Chat
Text
Image
Safeguards
Guardrails
Watermark detection
Builder tools
Knowledge bases
Agents
Prompt management Preview
Prompt flows Preview
Amazon Bedrock > Knowledge bases
Knowledge bases
Chat with your document
Knowledge bases
How it works
Upload and chat
Quickly query foundation models with context provided by ad-hoc dataset.
Chat with your document
Create a knowledge base
To create a knowledge base, specify the location of your data, select an embedding model, and configure a vector store for Bedrock to store and update your embeddings.
Test the knowledge base
Query your knowledge base in the test window. You can get source text chunks, or you can use the chunks to get responses from a foundation model.
Use the knowledge base
Integrate your knowledge base into your application as is or add it to agents.
Knowledge bases (0)
Edit
Delete
Test knowledge base
Create knowledge base
Find knowledge base
1
Name
Status
Descrip...
Source ...
Creatio...
Last sy...
Last sync
No knowledge base
No knowledge base to display

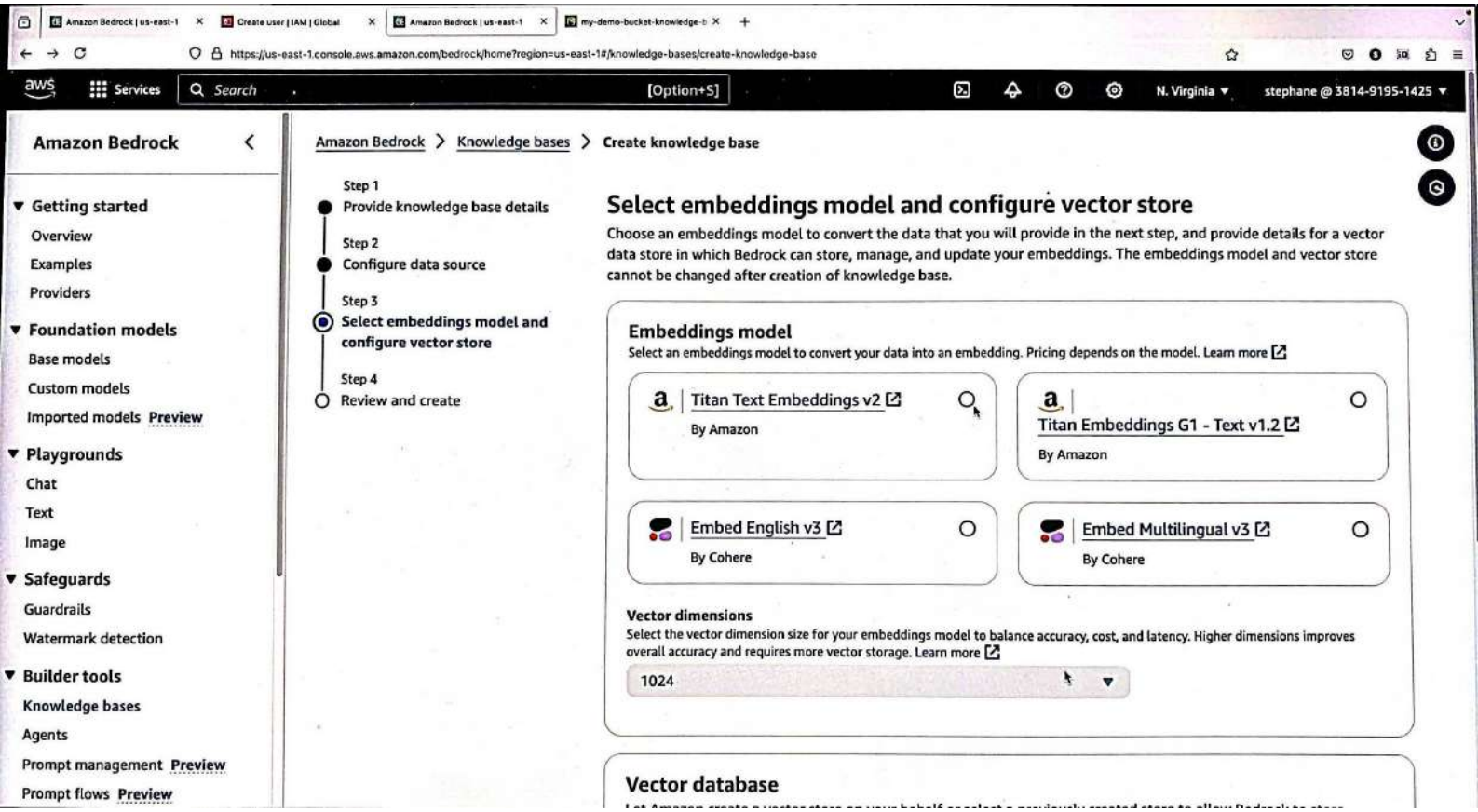

Amazon Bedrock | us-east-1
Create user | IAM | Global
Amazon Bedrock | us-east-1
https://us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/knowledge-bases/create-knowledge-base
aws
Services
Search
[Option+S]
N. Virginia
stephane @ 3814-9195-1425
Base models
Custo
Impor
Playg
Chat
Text
Image
Safeguards
Guardrails
Watermark detection
Builder tools
Knowledge bases
Agents
Prompt management Preview
Prompt flows Preview
Assessment & deployment
Model Evaluation
Provisioned Throughput
Model access
Bedrock Studio Preview
Settings
Click on "Create" => "Knowledge base with Vector Store"
Step 2
Configure data source
Step 3
Select embeddings model and configure vector store
Step 4
Review and create
Knowledge base details
Knowledge base name
knowledge-base-quick-start-vlhly
Valid characters are a-z, A-Z, 0-9, _ (underscore) and - (hyphen). The name can have up to 50 characters.
Knowledge base description - optional
Enter description
Valid characters are a-z, A-Z, 0-9, _ (underscore) and - (hyphen). The name can have up to 200 characters.
IAM permissions
Certain permissions are necessary to access other services or perform actions in order to create this resource. For more information, see service role for Amazon Bedrock
Runtime role
Create and use a new service role
Use an existing service role
Service role name
AmazonBedrockExecutionRoleForKnowledgeBase_vlhly

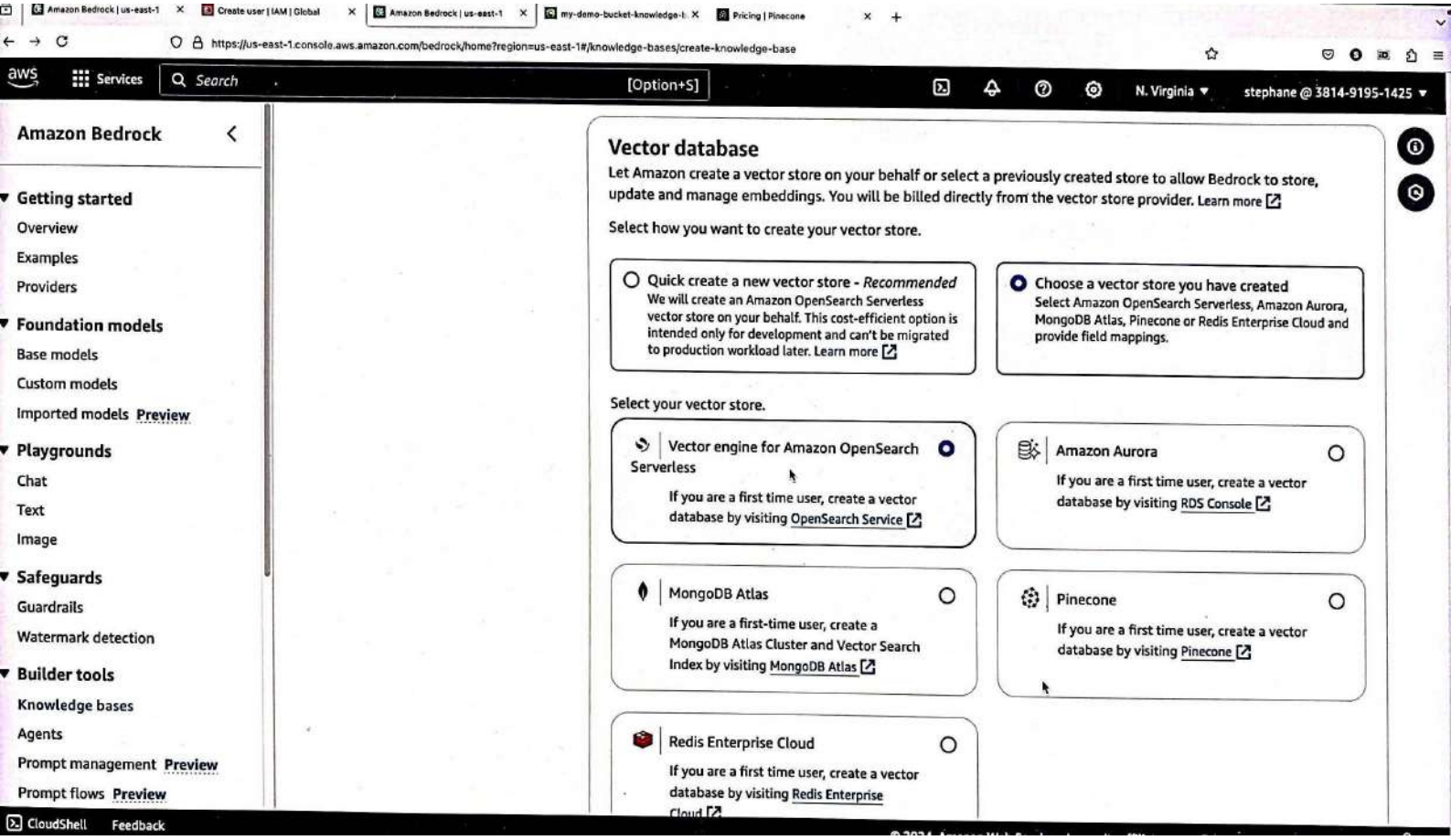

Amazon Bedrock | us-east-1
Create user | IAM | Global
Amazon Bedrock | us-east-1
https://us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/knowledge-bases/create-knowledge-base
aws
Services
Search
[Option+S]
N. Virginia
stephane @ 3814-9195-1425
Base models
Custom models
Imported models Preview
Playgrounds
Chat
Text
Image
Safeguards
Guardrails
Watermark detection
Builder tools
Knowledge bases
Agents
Prompt management Preview
Prompt flows Preview
Assessment & deployment
Model Evaluation
Provisioned Throughput
Model access
Bedrock Studio Preview
Settings
AmazonBedrockExecutionRoleForKnowledgeBase_vlhly
Choose data source
Select the data source that you want to configure in the next step.
Amazon S3
Object storage service that stores data as objects within buckets.
Web Crawler – Preview
Web page crawler that extracts content from public web pages you are authorized to crawl.
Third party data sources
Confluence – Preview
Collaborative work-management tool designed for project planning, software development and product management.
Salesforce – Preview
Customer relationship management (CRM) tool for managing support, sales, and marketing data.
Sharepoint – Preview
Collaborative web-based service for working on documents, web pages, web sites, lists, and more.
Tags
A tag is a label that you assign to an AWS resource. Each tag consists of a key and an optional value. You can use tags to search and filter your resources or track your AWS costs.

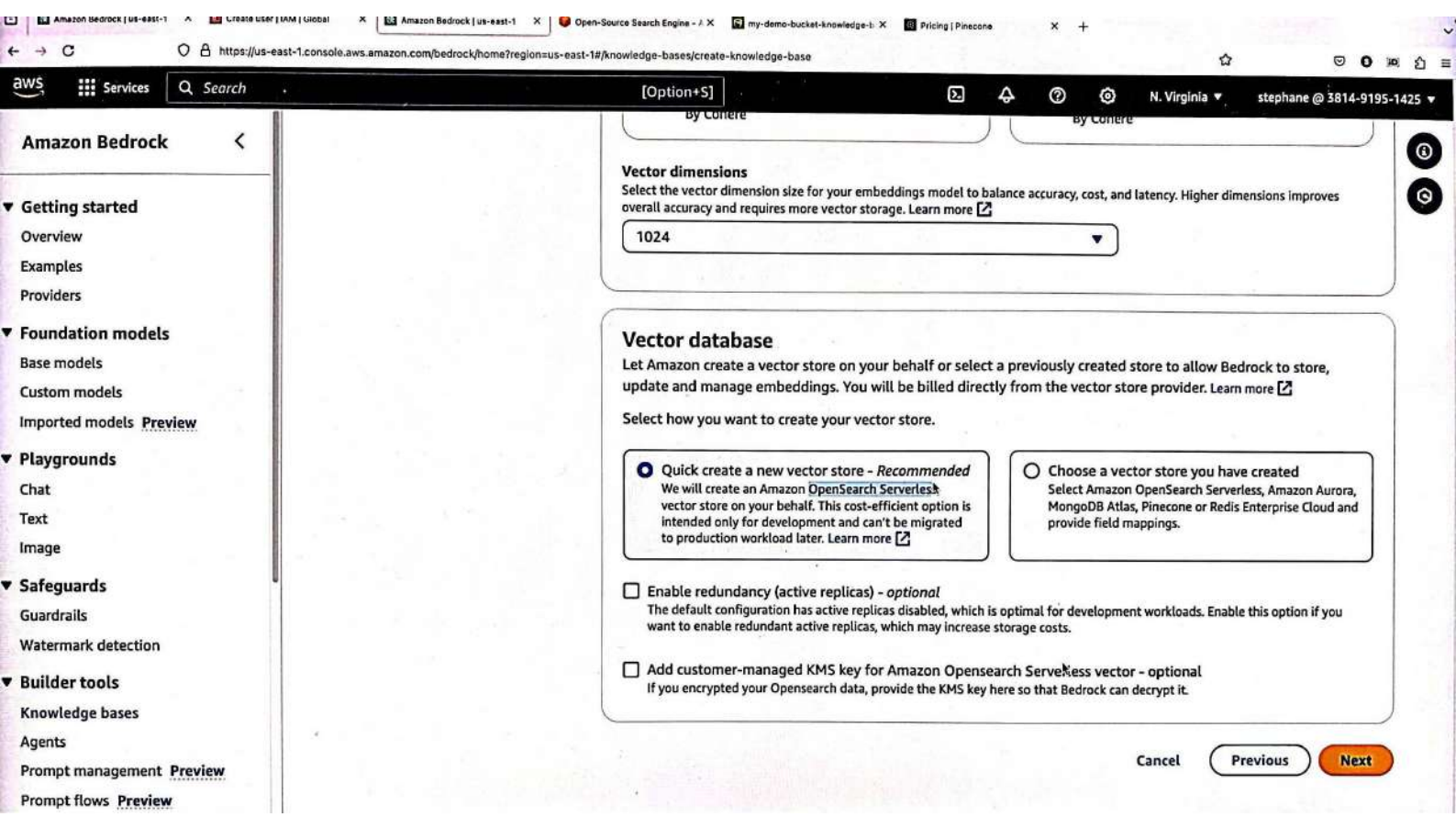

Amazon Bedrock | us-east-1
Create user | IAM | Global
Amazon Bedrock | us-east-1
my-demo-bucket-knowledge-b
https://us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/knowledge-bases/create-knowledge-base
aws
Services
Search
[Option+S]
N. Virginia
stephane @ 3814-9195-1425
Amazon Bedrock
Getting started
Overview
Examples
Providers
Foundation models
Base models
Custom models
Imported models Preview
Playgrounds
Chat
Text
Image
Safeguards
Guardrails
Watermark detection
Builder tools
Knowledge bases
Agents
Prompt management Preview
Prompt flows Preview
Select embeddings model and configure vector store
Step 4
Review and create
Data source: knowledge-base-quick-start-w3m5b-data-source
Delete
Data source name
knowledge-base-quick-start-w3m5b-data-source
Valid characters are a-z, A-Z, 0-9, _ (underscore) and - (hyphen). The name can have up to 100 characters.
Data source location
This AWS account
Other AWS account
S3 URI
To increase the accuracy and relevance of your responses, add a .metadata.json file containing metadata for your data source to your S3 bucket. Info
Choose a s3 location
View
Browse S3
Add customer-managed KMS key for S3 data - optional
If you encrypted your S3 data, provide the KMS key here so that Bedrock can decrypt it.
Chunking and parsing configurations Info
Choose between default or advanced customization.
Default
Uses default parsing and chunking strategy.
Custom
Customize the parsing and chunking strategy, including using advanced parsing.
Advanced settings - optional
Add data source
You can add 4 more data source(s).
Cancel
Previous
Next

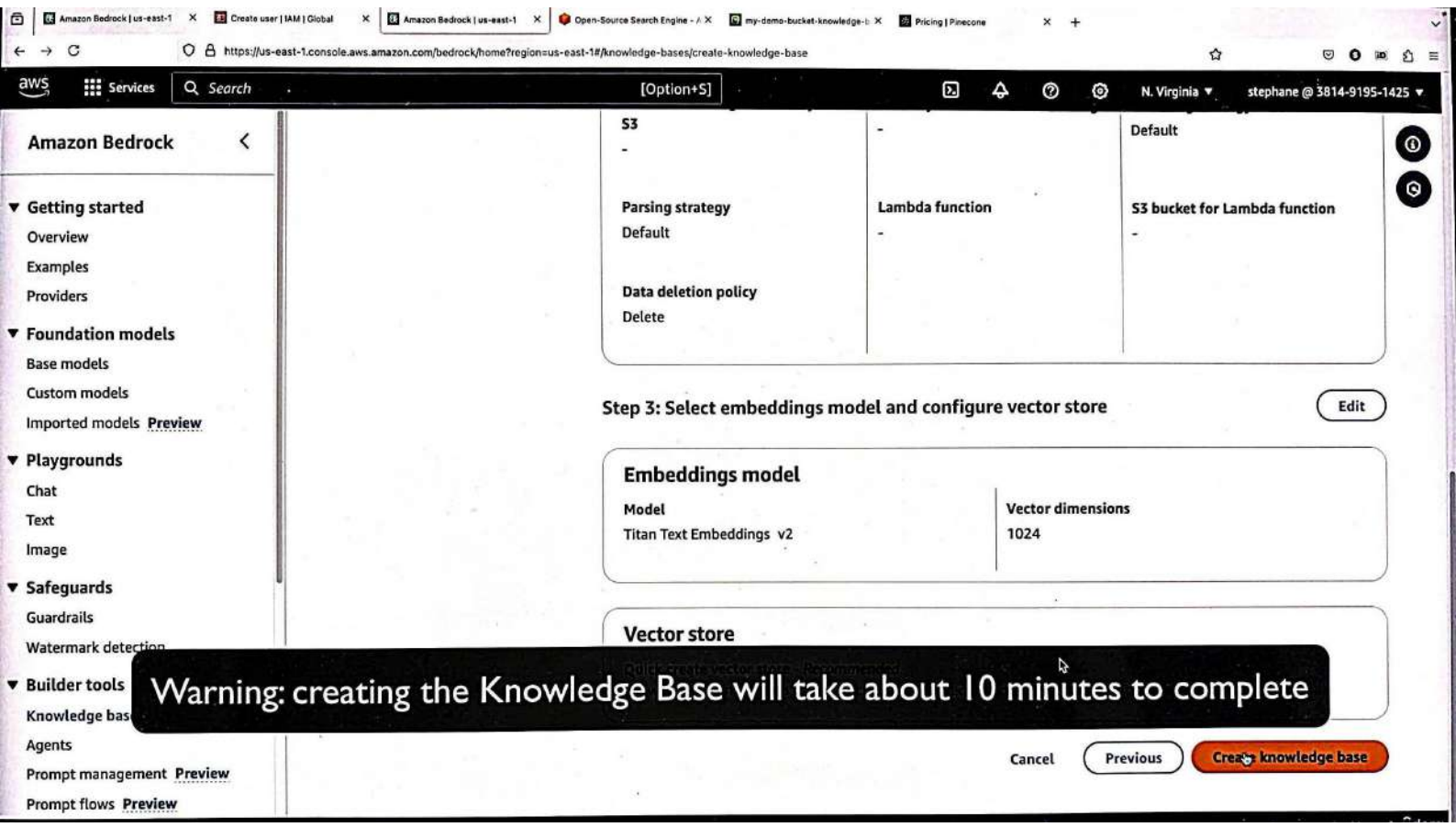

Step 4
Review and create
Data source name
knowledge-base-quick-start-w3m5b-data-source
Valid characters are a-z, A-Z, 0-9, _ (underscore) and - (hyphen). The name can have up to 100 characters.
Data source location
Choose an archive in S3
S3 buckets
Buckets (1/1)
Find S3 bucket
1
Name
Creation date
my-demo-bucket-knowledge-base-stephane
2024-07-20T14:05:33.000Z
Cancel
Choose
Advanced settings - optional
Add data source
You can add 4 more data source(s).
Getting started
Overview
Examples
Builder tools
Knowledge bases
Agents
Prompt management Preview
Prompt flows Preview
Watermark detection
Privacy
Terms
Cookie pref

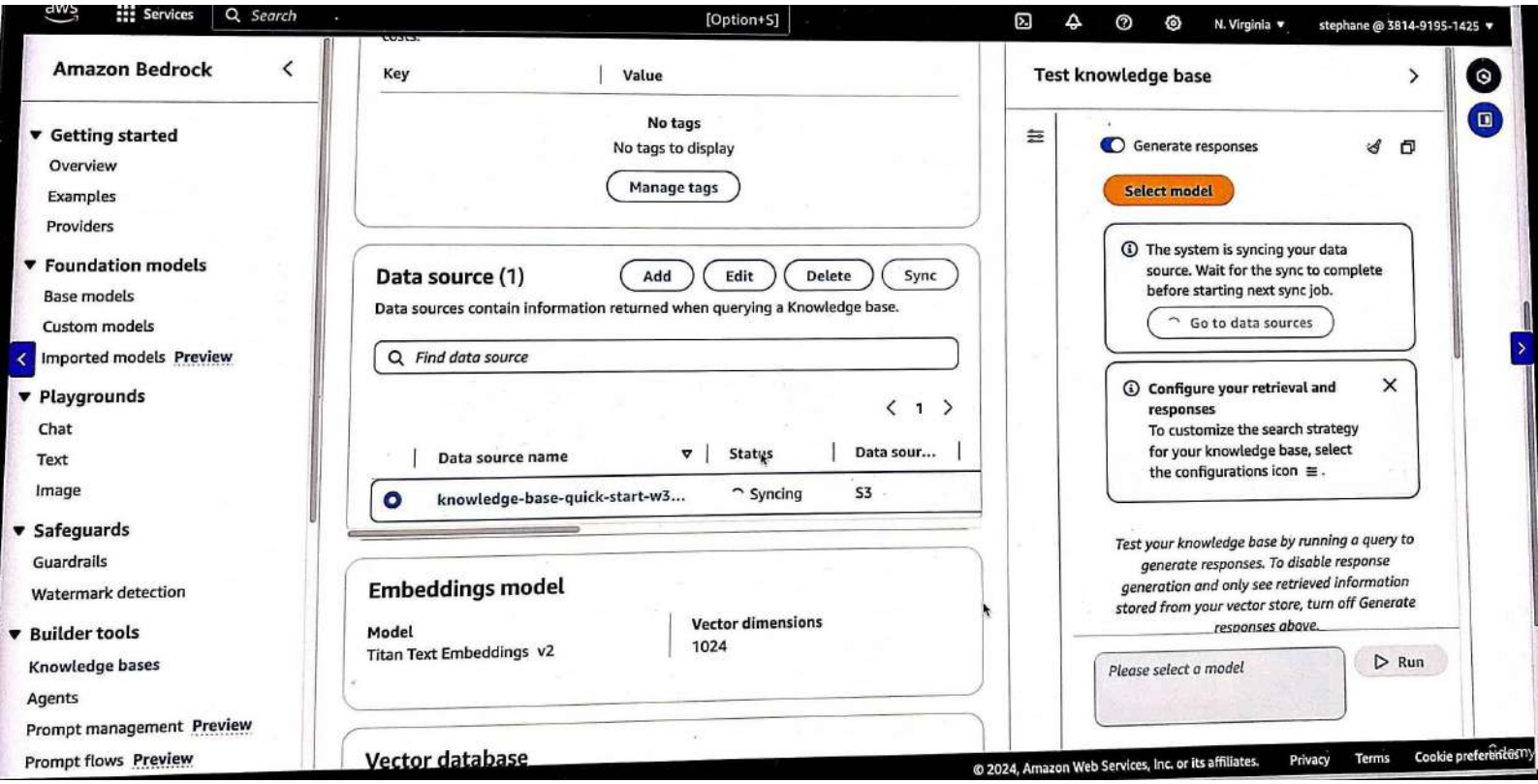

Amazon Bedrock | us-east-1
Create user | IAM | Global
Amazon Bedrock | us-east-1
my-demo-bucket-knowledge-b
https://us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/knowledge-bases/create-knowledge-base
aws
Services
Search
[Option+S]
N. Virginia
stephane @ 3814-9195-1425
Amazon Bedrock
Getting started
Overview
Examples
Providers
Foundation models
Base models
Custom models
Imported models Preview
Playgrounds
Chat
Text
Image
Safeguards
Guardrails
Watermark detection
Builder tools
Knowledge bases
Agents
Prompt management Preview
Prompt flows Preview
Step 4
Review and create
Data source name
knowledge-base-quick-start-w3m5b-data-source
Valid characters are a-z, A-Z, 0-9, _ (underscore) and - (hyphen). The name can have up to 100 characters.
Data source location
This AWS account
Other AWS account
S3 URI
To increase the accuracy and relevance of your responses, add a .metadata.json file containing metadata for your data source to your S3 bucket. Info
s3://my-demo-bucket-knowledge-base-stephane
View
Browse S3
Add customer-managed KMS key for S3 data - optional
If you encrypted your S3 data, provide the KMS key here so that Bedrock can decrypt it.
Chunking and parsing configurations Info
Choose between default or advanced customization.
Default
Uses default parsing and chunking strategy.
Custom
Customize the parsing and chunking strategy, including using advanced parsing.
Advanced settings - optional
Add data source
You can add 4 more data source(s).
Cancel
Previous
Next

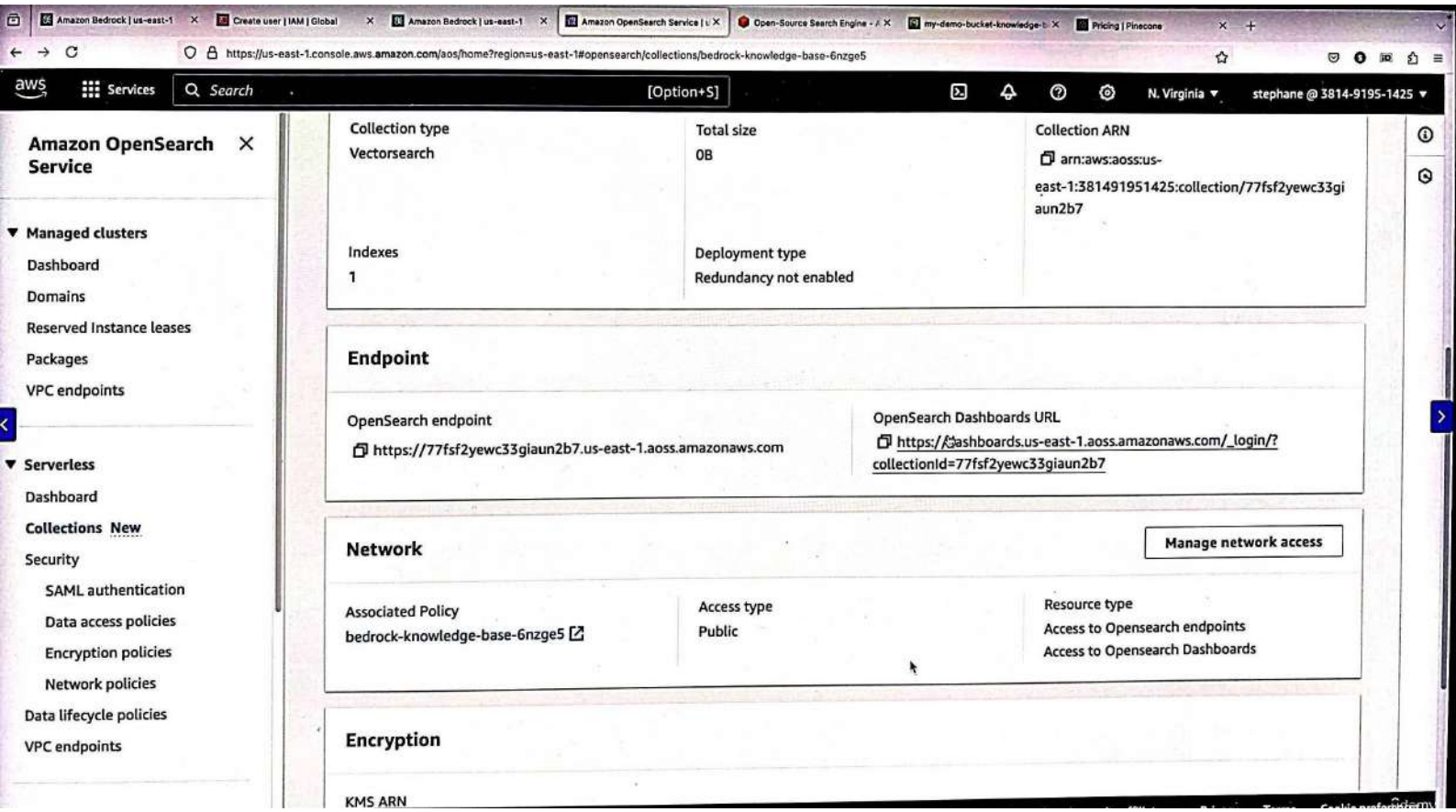

Amazon Bedrock | us-east-1
Create user | IAM | Global
Amazon Bedrock | us-east-1
my-demo-bucket-knowledge-b
https://us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/knowledge-bases/create-knowledge-base
aws
Services
Search
[Option+S]
N. Virginia
stephane @ 3814-9195-1425
Amazon Bedrock
Getting started
Overview
Examples
Providers
Foundation models
Base models
Custom models
Imported models Preview
Playgrounds
Chat
Text
Image
Safeguards
Guardrails
Watermark detection
Builder tools
Knowledge bases
Agents
Prompt management Preview
Prompt flows Preview
Amazon Bedrock > Knowledge bases > Create knowledge base
Step 1
Provide knowledge base details
Step 2
Configure data source
Step 3
Select embeddings model and configure vector store
Step 4
Review and create
Select embeddings model and configure vector store
Choose an embeddings model to convert the data that you will provide in the next step, and provide details for a vector data store in which Bedrock can store, manage, and update your embeddings. The embeddings model and vector store cannot be changed after creation of knowledge base.
Embeddings model
Select an embeddings model to convert your data into an embedding. Pricing depends on the model. Learn more
Titan Text Embeddings v2
By Amazon
Titan Embeddings G1 - Text v1.2
By Amazon
Embed English v3
By Cohere
Embed Multilingual v3
By Cohere
Vector dimensions
Select the vector dimension size for your embeddings model to balance accuracy, cost, and latency. Higher dimensions improves overall accuracy and requires more vector storage. Learn more
1024
Vector database

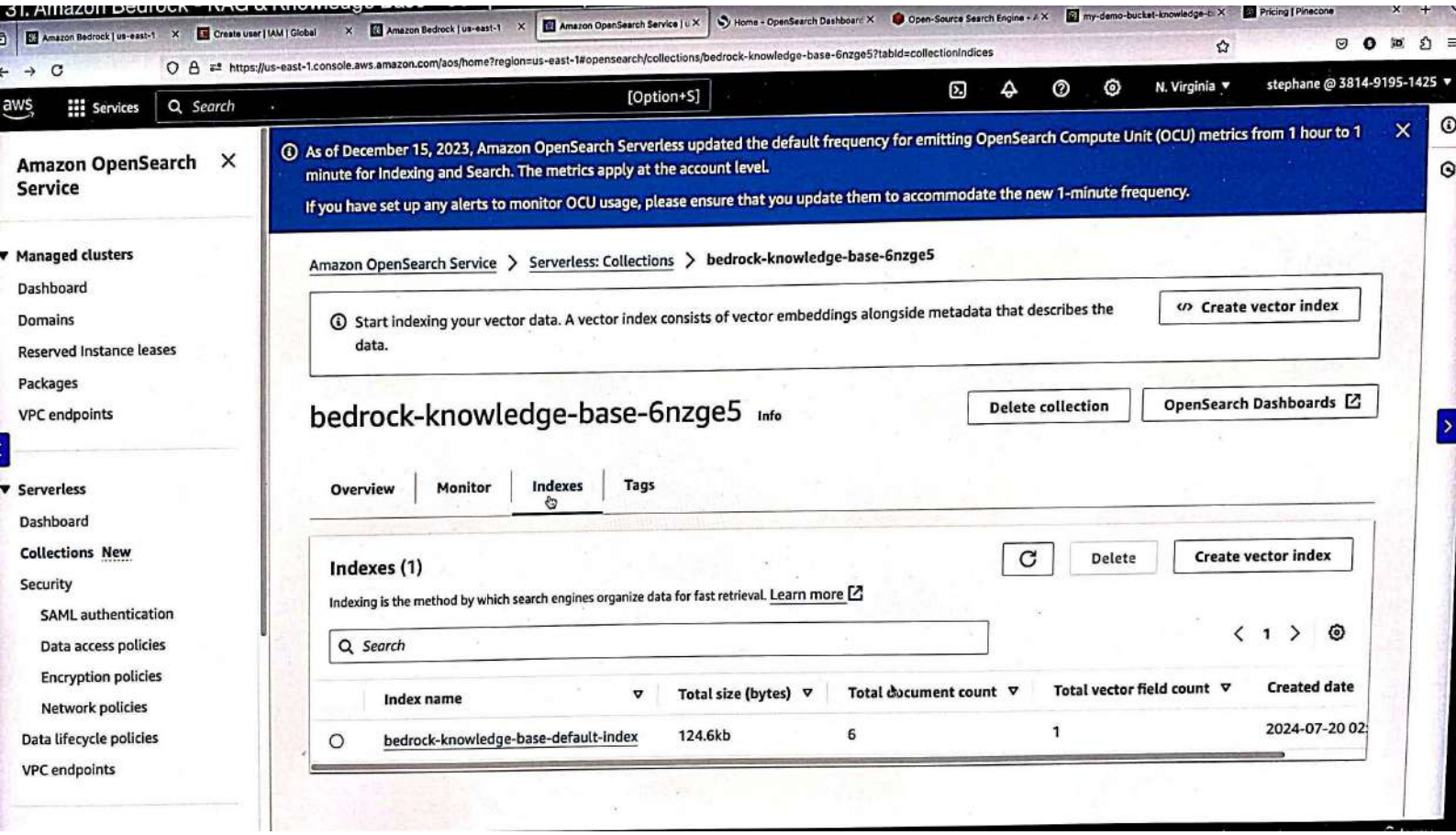

Amazon Bedrock | us-east-1
Create user | IAM | Global
Amazon Bedrock | us-east-1
my-demo-bucket-knowledge-b
Pricing | Pinecone
https://us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/knowledge-bases/create-knowledge-base
aws
Services
Search
[Option+S]
N. Virginia
stephane @ 3814-9195-1425
Amazon Bedrock
Getting started
Overview
Examples
Providers
Foundation models
Base models
Custom models
Imported models Preview
Playgrounds
Chat
Text
Image
Safeguards
Guardrails
Watermark detection
Builder tools
Knowledge bases
Agents
Prompt management Preview
Prompt flows Preview
Vector database
Let Amazon create a vector store on your behalf or select a previously created store to allow Bedrock to store, update and manage embeddings. You will be billed directly from the vector store provider. Learn more
Select how you want to create your vector store.
Quick create a new vector store - Recommended
We will create an Amazon OpenSearch Serverless vector store on your behalf. This cost-efficient option is intended only for development and can't be migrated to production workload later. Learn more
Choose a vector store you have created
Select Amazon OpenSearch Serverless, Amazon Aurora, MongoDB Atlas, Pinecone or Redis Enterprise Cloud and provide field mappings.
Select your vector store.
Vector engine for Amazon OpenSearch Serverless
If you are a first time user, create a vector database by visiting OpenSearch Service
Amazon Aurora
If you are a first time user, create a vector database by visiting RDS Console
MongoDB Atlas
If you are a first-time user, create a MongoDB Atlas Cluster and Vector Search Index by visiting MongoDB Atlas
Pinecone
If you are a first time user, create a vector database by visiting Pinecone
Redis Enterprise Cloud
If you are a first time user, create a vector database by visiting Redis Enterprise Cloud
CloudShell
Feedback

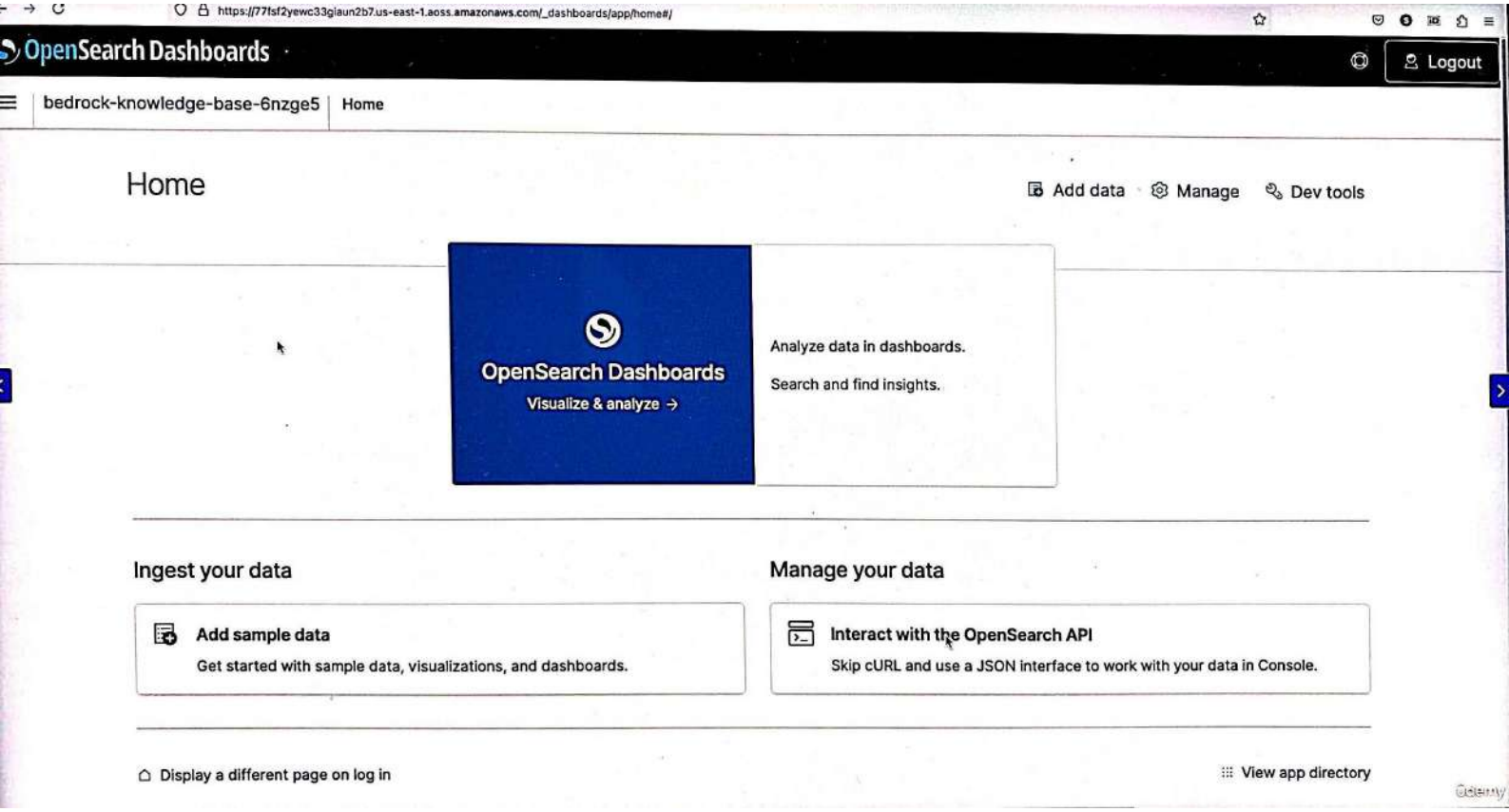

Amazon Bedrock | us-east-1
Create user | IAM | Global
Amazon Bedrock | us-east-1
Open-Source Search Engine - A
my-demo-bucket-knowledge-b
Pricing | Pinecone
https://us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/knowledge-bases/create-knowledge-base
aws
Services
Search
[Option+S]
N. Virginia
stephane @ 3814-9195-1425
Amazon Bedrock
Getting started
Overview
Examples
Providers
Foundation models
Base models
Custom models
Imported models Preview
Playgrounds
Chat
Text
Image
Safeguards
Guardrails
Watermark detection
Builder tools
Knowledge bases
Agents
Prompt management Preview
Prompt flows Preview
By Cohere
By Cohere
Vector dimensions
Select the vector dimension size for your embeddings model to balance accuracy, cost, and latency. Higher dimensions improves overall accuracy and requires more vector storage. Learn more
1024
Vector database
Let Amazon create a vector store on your behalf or select a previously created store to allow Bedrock to store, update and manage embeddings. You will be billed directly from the vector store provider. Learn more
Select how you want to create your vector store.
Quick create a new vector store - Recommended
We will create an Amazon OpenSearch Serverless vector store on your behalf. This cost-efficient option is intended only for development and can't be migrated to production workload later. Learn more
Choose a vector store you have created
Select Amazon OpenSearch Serverless, Amazon Aurora, MongoDB Atlas, Pinecone or Redis Enterprise Cloud and provide field mappings.
Enable redundancy (active replicas) - optional
The default configuration has active replicas disabled, which is optimal for development workloads. Enable this option if you want to enable redundant active replicas, which may increase storage costs.
Add customer-managed KMS key for Amazon Opensearch Serverless vector - optional
If you encrypted your Opensearch data, provide the KMS key here so that Bedrock can decrypt it.
Cancel
Previous
Next

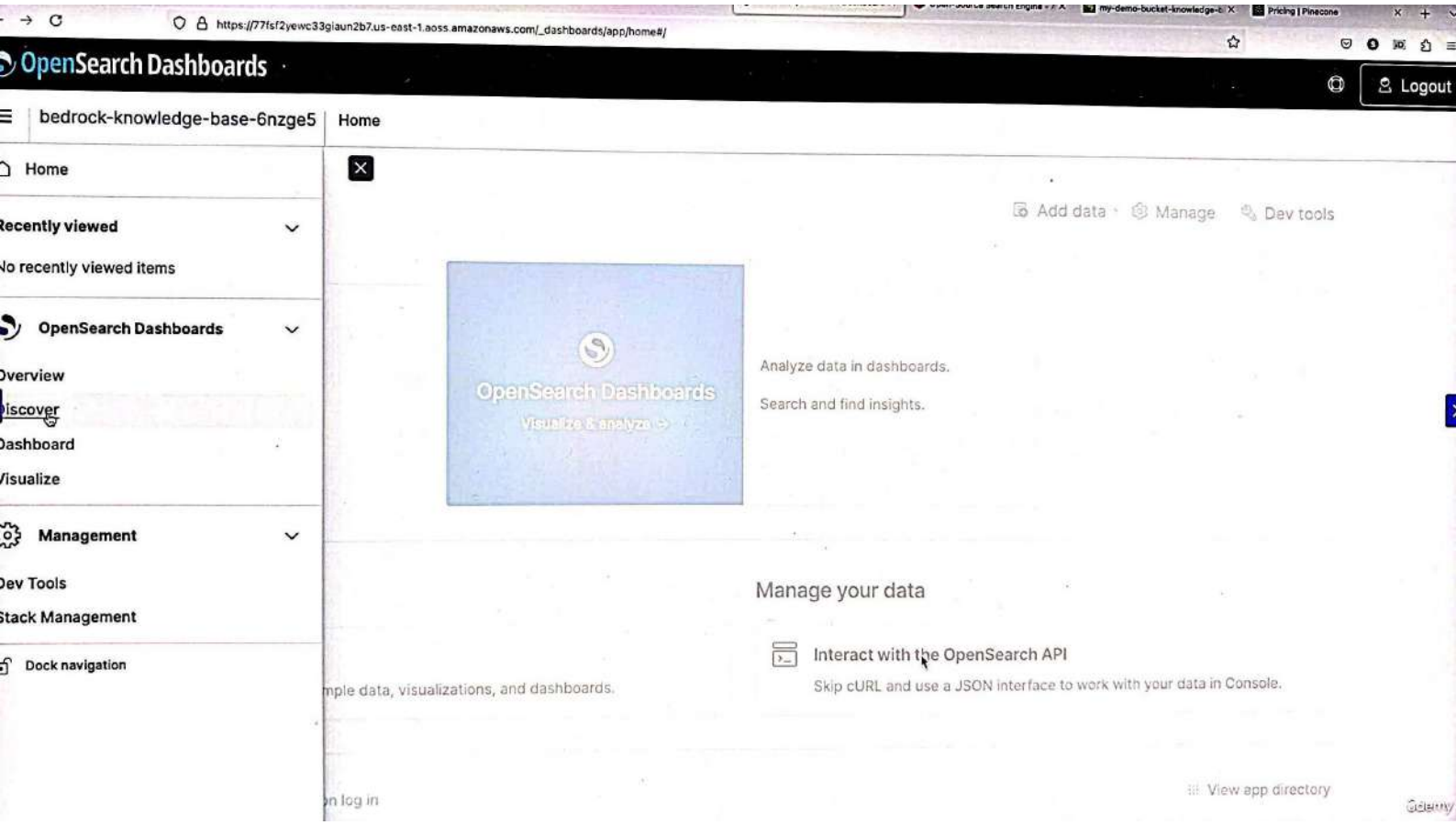

Amazon Bedrock | us-east-1
Create user | IAM | Global
Amazon Bedrock | us-east-1
Open-Source Search Engine -
my-demo-bucket-knowledge-b
Pricing | Pinecone
https://us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/knowledge-bases/create-knowledge-base
aws
Services
Search
[Option+S]
N. Virginia
stephane @ 3814-9195-1425
Amazon Bedrock
Getting started
Overview
Examples
Providers
Foundation models
Base models
Custom models
Imported models Preview
Playgrounds
Chat
Text
Image
Safeguards
Guardrails
Watermark detection
Builder tools
Knowledge bas
Agents
Prompt management Preview
Prompt flows Preview
S3
-
Default
Parsing strategy
Default
Lambda function
-
S3 bucket for Lambda function
-
Data deletion policy
Delete
Step 3: Select embeddings model and configure vector store
Edit
Embeddings model
Model
Titan Text Embeddings v2
Vector dimensions
1024
Vector store
Warning: creating the Knowledge Base will take about 10 minutes to complete
Cancel
Previous
Create knowledge base

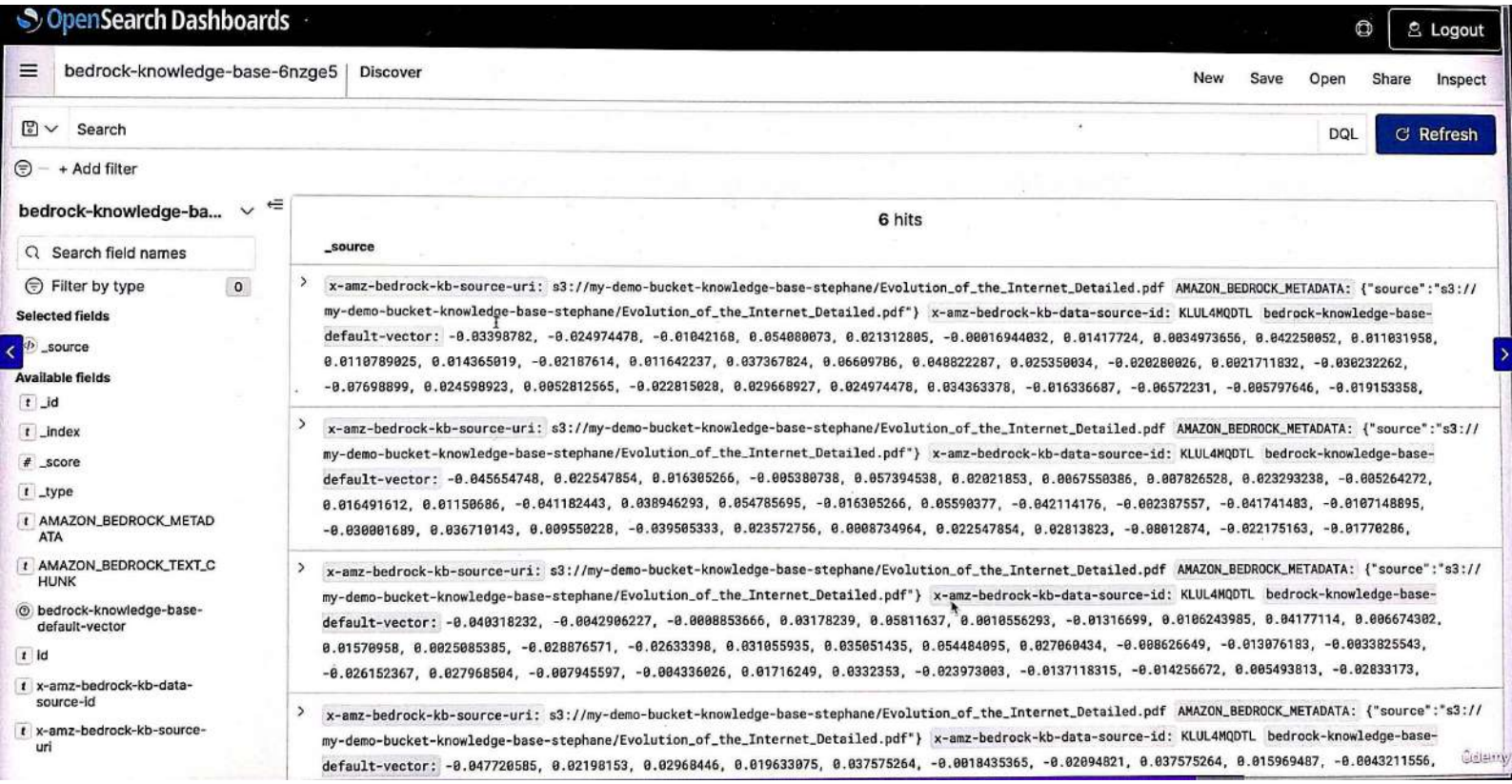

Services
Search
[Option+S]
N. Virginia
stephane @ 3814-9195-1425
Amazon Bedrock
Getting started
Overview
Examples
Providers
Foundation models
Base models
Custom models
Imported models Preview
Playgrounds
Chat
Text
Image
Safeguards
Guardrails
Watermark detection
Builder tools
Knowledge bases
Agents
Prompt management Preview
Prompt flows Preview
Key
Value
No tags
No tags to display
Manage tags
Data source (1)
Add
Edit
Delete
Sync
Data sources contain information returned when querying a Knowledge base.
Find data source
1
Data source name
Status
Data sour...
knowledge-base-quick-start-w3...
Syncing
S3
Embeddings model
Model
Titan Text Embeddings v2
Vector dimensions
1024
Vector database
Test knowledge base
Generate responses
Select model
The system is syncing your data source. Wait for the sync to complete before starting next sync job.
Go to data sources
Configure your retrieval and responses
To customize the search strategy for your knowledge base, select the configurations icon
Test your knowledge base by running a query to generate responses. To disable response generation and only see retrieved information stored from your vector store, turn off Generate responses above.
Please select a model
Run
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

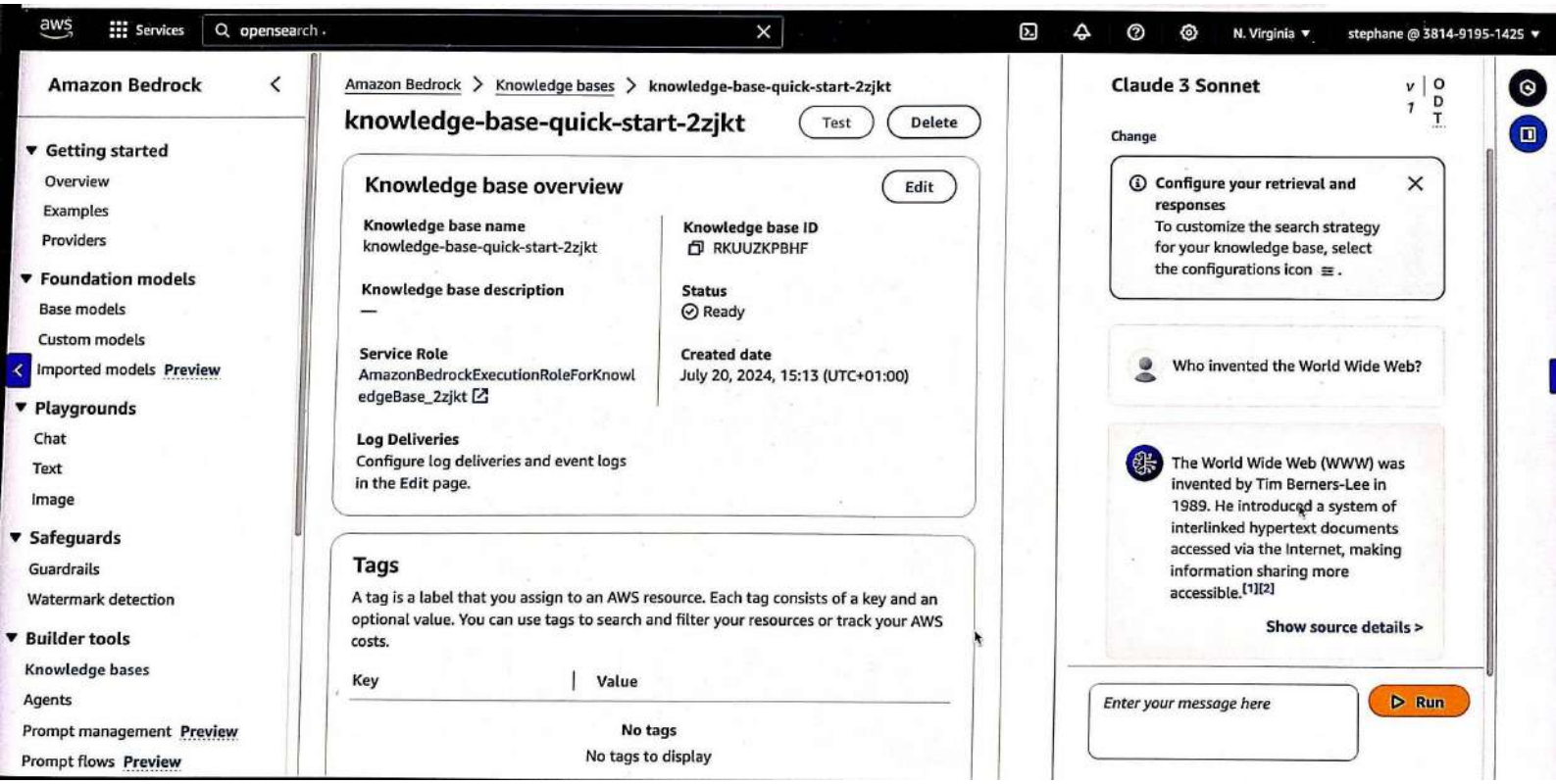

Amazon Bedrock | us-east-1
Create user | IAM | Global
Amazon Bedrock | us-east-1
Amazon OpenSearch Service
Open-Source Search Engine
my-demo-bucket-knowledge-b
Pricing | Pinecone
https://us-east-1.console.aws.amazon.com/aos/home?region=us-east-1#opensearch/collections/bedrock-knowledge-base-6nzge5
aws
Services
Search
[Option+S]
N. Virginia
stephane @ 3814-9195-1425
Amazon OpenSearch Service
Managed clusters
Dashboard
Domains
Reserved Instance leases
Packages
VPC endpoints
Serverless
Dashboard
Collections New
Security
SAML authentication
Data access policies
Encryption policies
Network policies
Data lifecycle policies
VPC endpoints
Collection type
Vectorsearch
Total size
0B
Collection ARN
arn:aws:aoss:us-east-1:381491951425:collection/77fsf2yewc33giaun2b7
Indexes
1
Deployment type
Redundancy not enabled
Endpoint
OpenSearch endpoint
https://77fsf2yewc33giaun2b7.us-east-1.aoss.amazonaws.com
OpenSearch Dashboards URL
https://dashboards.us-east-1.aoss.amazonaws.com/_login/?collectionId=77fsf2yewc33giaun2b7
Network
Manage network access
Associated Policy
bedrock-knowledge-base-6nzge5
Access type
Public
Resource type
Access to Opensearch endpoints
Access to Opensearch Dashboards
Encryption
KMS ARN


Amazon Bedrock | us-east-1
Create user | IAM | Global
Amazon Bedrock | us-east-1
Amazon OpenSearch Service
Home - OpenSearch Dashboard
Open-Source Search Engine
my-demo-bucket-knowledge-b
Pricing | Pinecone
https://us-east-1.console.aws.amazon.com/aos/home?region=us-east-1#opensearch/collections/bedrock-knowledge-base-6nzge5?tabId=collectionIndices
aws
Services
Search
[Option+S]
N. Virginia
stephane @ 3814-9195-1425
Amazon OpenSearch Service
Managed clusters
Dashboard
Domains
Reserved Instance leases
Packages
VPC endpoints
Serverless
Dashboard
Collections New
Security
SAML authentication
Data access policies
Encryption policies
Network policies
Data lifecycle policies
VPC endpoints
As of December 15, 2023, Amazon OpenSearch Serverless updated the default frequency for emitting OpenSearch Compute Unit (OCU) metrics from 1 hour to 1 minute for Indexing and Search. The metrics apply at the account level.
If you have set up any alerts to monitor OCU usage, please ensure that you update them to accommodate the new 1-minute frequency.
Amazon OpenSearch Service > Serverless: Collections > bedrock-knowledge-base-6nzge5
Start indexing your vector data. A vector index consists of vector embeddings alongside metadata that describes the data.
Create vector index
bedrock-knowledge-base-6nzge5 Info
Delete collection
OpenSearch Dashboards
Overview
Monitor
Indexes
Tags
Indexes (1)
Delete
Create vector index
Indexing is the method by which search engines organize data for fast retrieval. Learn more
Search
1
Index name
Total size (bytes)
Total document count
Total vector field count
Created date
bedrock-knowledge-base-default-index
124.6kb
6
1
2024-07-20 02:


https://77fsf2yewc33glaun2b7.us-east-1.aoss.amazonaws.com/_dashboards/app/home#/
OpenSearch Dashboards
Logout
bedrock-knowledge-base-6nzge5
Home
Home
Add data
Manage
Dev tools
OpenSearch Dashboards
Visualize & analyze →
Analyze data in dashboards.
Search and find insights.
Ingest your data
Add sample data
Get started with sample data, visualizations, and dashboards.
Manage your data
Interact with the OpenSearch API
Skip cURL and use a JSON interface to work with your data in Console.
Display a different page on log in
View app directory

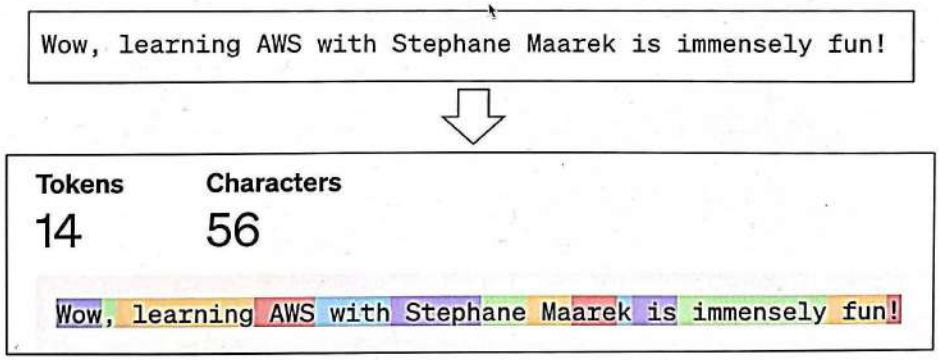

https://77fsf2yewc33giaun2b7.us-east-1.aoss.amazonaws.com/_dashboards/app/home#/
OpenSearch Dashboards
Logout
bedrock-knowledge-base-6nzge5
Home
Home
Recently viewed
No recently viewed items
OpenSearch Dashboards
Overview
Discover
Dashboard
Visualize
Management
Dev Tools
Stack Management
Dock navigation
Add data
Manage
Dev tools
OpenSearch Dashboards
Visualize & analyze
Analyze data in dashboards.
Search and find insights.
Manage your data
Interact with the OpenSearch API
Skip cURL and use a JSON interface to work with your data in Console.
mple data, visualizations, and dashboards.
n log in
View app directory


OpenSearch Dashboards
Logout
bedrock-knowledge-base-6nzge5
Discover
New
Save
Open
Share
Inspect
Search
DQL
Refresh
+ Add filter
bedrock-knowledge-ba...
Search field names
Filter by type
0
Selected fields
_source
Available fields
_id
_index
_score
_type
AMAZON_BEDROCK_METADATA
AMAZON_BEDROCK_TEXT_CHUNK
bedrock-knowledge-base-default-vector
id
x-amz-bedrock-kb-data-source-id
x-amz-bedrock-kb-source-uri
6 hits
_source
x-amz-bedrock-kb-source-uri: s3://my-demo-bucket-knowledge-base-stephane/Evolution_of_the_Internet_Detailed.pdf AMAZON_BEDROCK_METADATA: {"source":"s3://my-demo-bucket-knowledge-base-stephane/Evolution_of_the_Internet_Detailed.pdf"} x-amz-bedrock-kb-data-source-id: KLUL4MQDTL bedrock-knowledge-base-default-vector: -0.03398782, -0.024974478, -0.01042168, 0.054080073, 0.021312805, -0.00016944032, 0.01417724, 0.0034973656, 0.042250052, 0.011031958, 0.0110789025, 0.014365019, -0.02187614, 0.011642237, 0.037367824, 0.06609786, 0.048822287, 0.025350034, -0.020280026, 0.0021711832, -0.030232262, -0.07698899, 0.024598923, 0.0052812565, -0.022815028, 0.029668927, 0.024974478, 0.034363378, -0.016336687, -0.06572231, -0.005797646, -0.019153358,
x-amz-bedrock-kb-source-uri: s3://my-demo-bucket-knowledge-base-stephane/Evolution_of_the_Internet_Detailed.pdf AMAZON_BEDROCK_METADATA: {"source":"s3://my-demo-bucket-knowledge-base-stephane/Evolution_of_the_Internet_Detailed.pdf"} x-amz-bedrock-kb-data-source-id: KLUL4MQDTL bedrock-knowledge-base-default-vector: -0.045654748, 0.022547854, 0.016305266, -0.005380738, 0.057394538, 0.02021853, 0.0067550386, 0.007826528, 0.023293238, -0.005264272, 0.016491612, 0.01150686, -0.041182443, 0.038946293, 0.054785695, -0.016305266, 0.05590377, -0.042114176, -0.002387557, -0.041741483, -0.0107148895, -0.030001689, 0.036710143, 0.009550228, -0.039505333, 0.023572756, 0.0008734964, 0.022547854, 0.02813823, -0.08012874, -0.022175163, -0.01770286,
x-amz-bedrock-kb-source-uri: s3://my-demo-bucket-knowledge-base-stephane/Evolution_of_the_Internet_Detailed.pdf AMAZON_BEDROCK_METADATA: {"source":"s3://my-demo-bucket-knowledge-base-stephane/Evolution_of_the_Internet_Detailed.pdf"} x-amz-bedrock-kb-data-source-id: KLUL4MQDTL bedrock-knowledge-base-default-vector: -0.040318232, -0.0042906227, -0.0008853666, 0.03178239, 0.05811637, 0.0010556293, -0.01316699, 0.0106243985, 0.04177114, 0.006674302, 0.01570958, 0.0025085385, -0.028876571, -0.02633398, 0.031055935, 0.035051435, 0.054484095, 0.027060434, -0.008626649, -0.013076183, -0.0033825543, -0.026152367, 0.027968504, -0.007945597, -0.004336026, 0.01716249, 0.0332353, -0.023973003, -0.0137118315, -0.014256672, 0.005493813, -0.02833173,
x-amz-bedrock-kb-source-uri: s3://my-demo-bucket-knowledge-base-stephane/Evolution_of_the_Internet_Detailed.pdf AMAZON_BEDROCK_METADATA: {"source":"s3://my-demo-bucket-knowledge-base-stephane/Evolution_of_the_Internet_Detailed.pdf"} x-amz-bedrock-kb-data-source-id: KLUL4MQDTL bedrock-knowledge-base-default-vector: -0.047720585, 0.02198153, 0.02968446, 0.019633075, 0.037575264, -0.0018435365, -0.02094821, 0.037575264, 0.015969487, -0.0043211556,


aws
Services
opensearch
N. Virginia
stephane @ 3814-9195-1425
Amazon Bedrock
Getting started
Overview
Examples
Providers
Foundation models
Base models
Custom models
Imported models Preview
Playgrounds
Chat
Text
Image
Safeguards
Guardrails
Watermark detection
Builder tools
Knowledge bases
Agents
Prompt management Preview
Prompt flows Preview
Amazon Bedrock > Knowledge bases > knowledge-base-quick-start-2zjkt
knowledge-base-quick-start-2zjkt
Test
Delete
Knowledge base overview
Edit
Knowledge base name
knowledge-base-quick-start-2zjkt
Knowledge base ID
RKUUZKPBHF
Knowledge base description
—
Status
Ready
Service Role
AmazonBedrockExecutionRoleForKnowledgeBase_2zjkt
Created date
July 20, 2024, 15:13 (UTC+01:00)
Log Deliveries
Configure log deliveries and event logs in the Edit page.
Tags
A tag is a label that you assign to an AWS resource. Each tag consists of a key and an optional value. You can use tags to search and filter your resources or track your AWS costs.
Key
Value
No tags
No tags to display
Claude 3 Sonnet
Change
Configure your retrieval and responses
To customize the search strategy for your knowledge base, select the configurations icon .
Who invented the World Wide Web?
The World Wide Web (WWW) was invented by Tim Berners-Lee in 1989. He introduced a system of interlinked hypertext documents accessed via the Internet, making information sharing more accessible.[1][2]
Show source details >
Enter your message here
Run

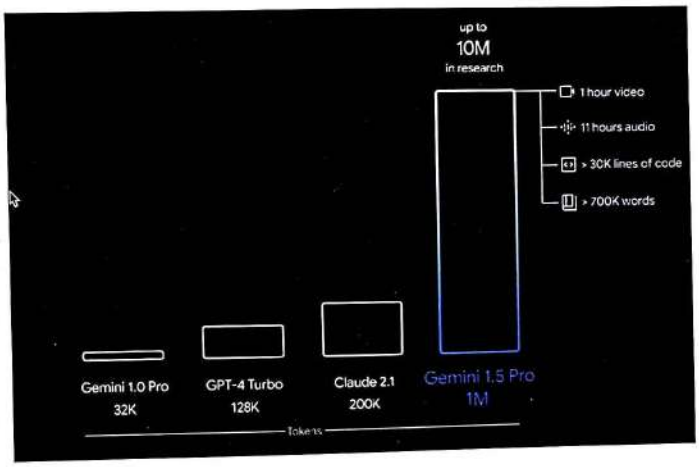
GenAI Concepts – Tokenization

- Tokenization: converting raw text into a sequence of tokens
  - Word-based tokenization: text is split into individual words
  - Subword tokenization: some words can be split too (helpful for long words…)
- Can experiment at: https://platform.openai.com/tokenizer



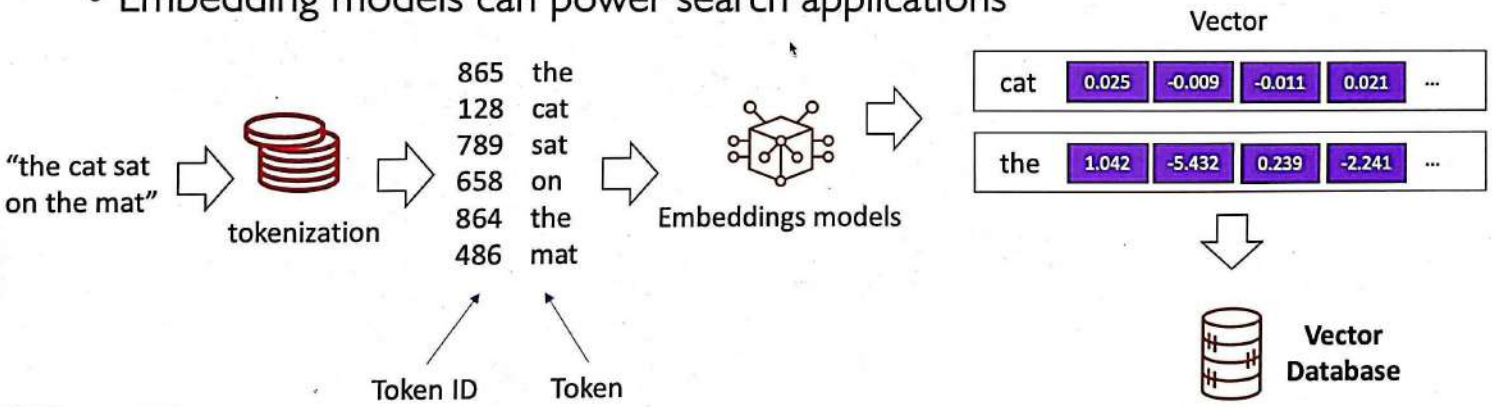
GenAI Concepts – Context Window

- The number of tokens an LLM can consider when generating text
- The larger the context window, the more information and coherence
- Large context windows require more memory and processing power
- First factor to look at when considering a model



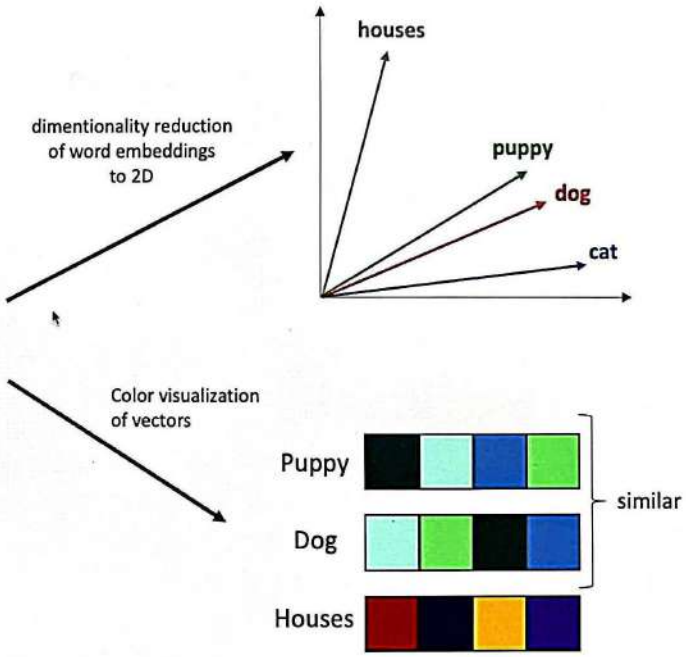
GenAI Concepts – Embeddings

- Create vectors (array of numerical values) out of text, images or audio
- Vectors have a high dimensionality to capture many features for one input token, such as semantic meaning, syntactic role, sentiment
- Embedding models can power search applications



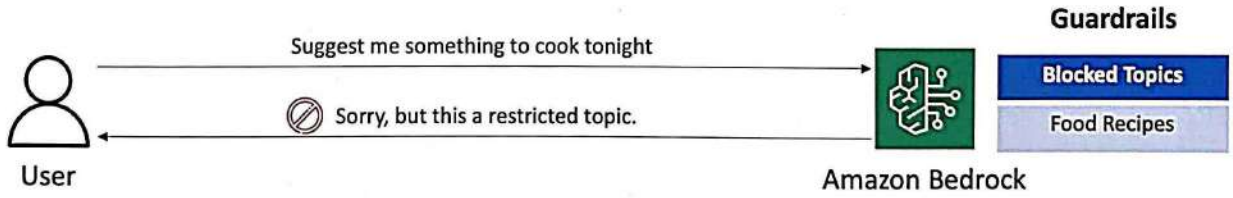
Words that have a Semantic Relationship have Similar Embeddings

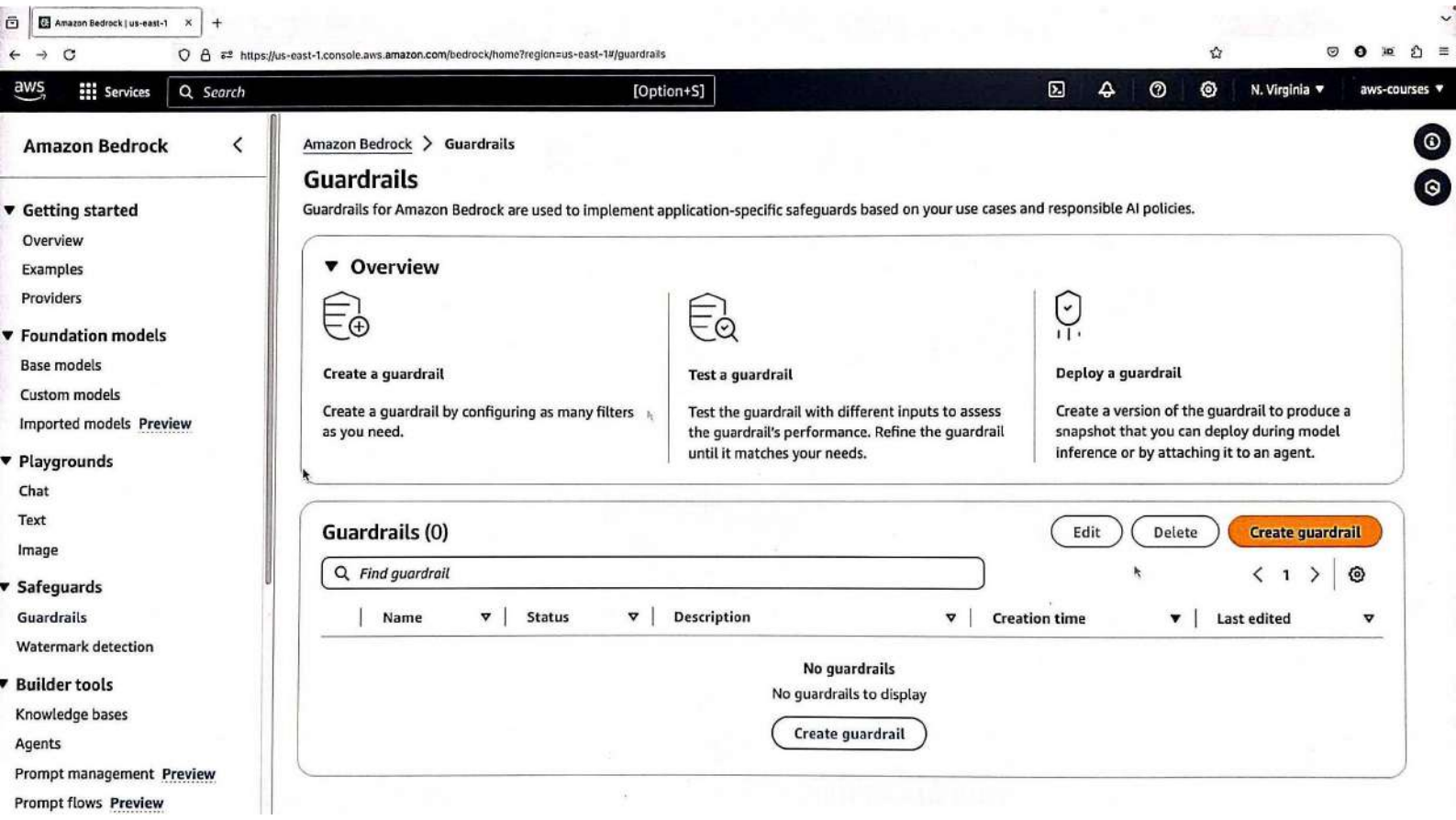
	d1	d2	d3	d4	d5	...	d100
dog	0.6	0.9	0.1	0.4	-0.7	...	-0.2
puppy	0.5	0.8	-0.1	0.2	-0.6	...	-0.1
cat	0.7	-0.1	0.4	0.3	-0.4	...	-0.3
houses	-0.8	-0.4	-0.5	0.1	-0.9	...	0.8

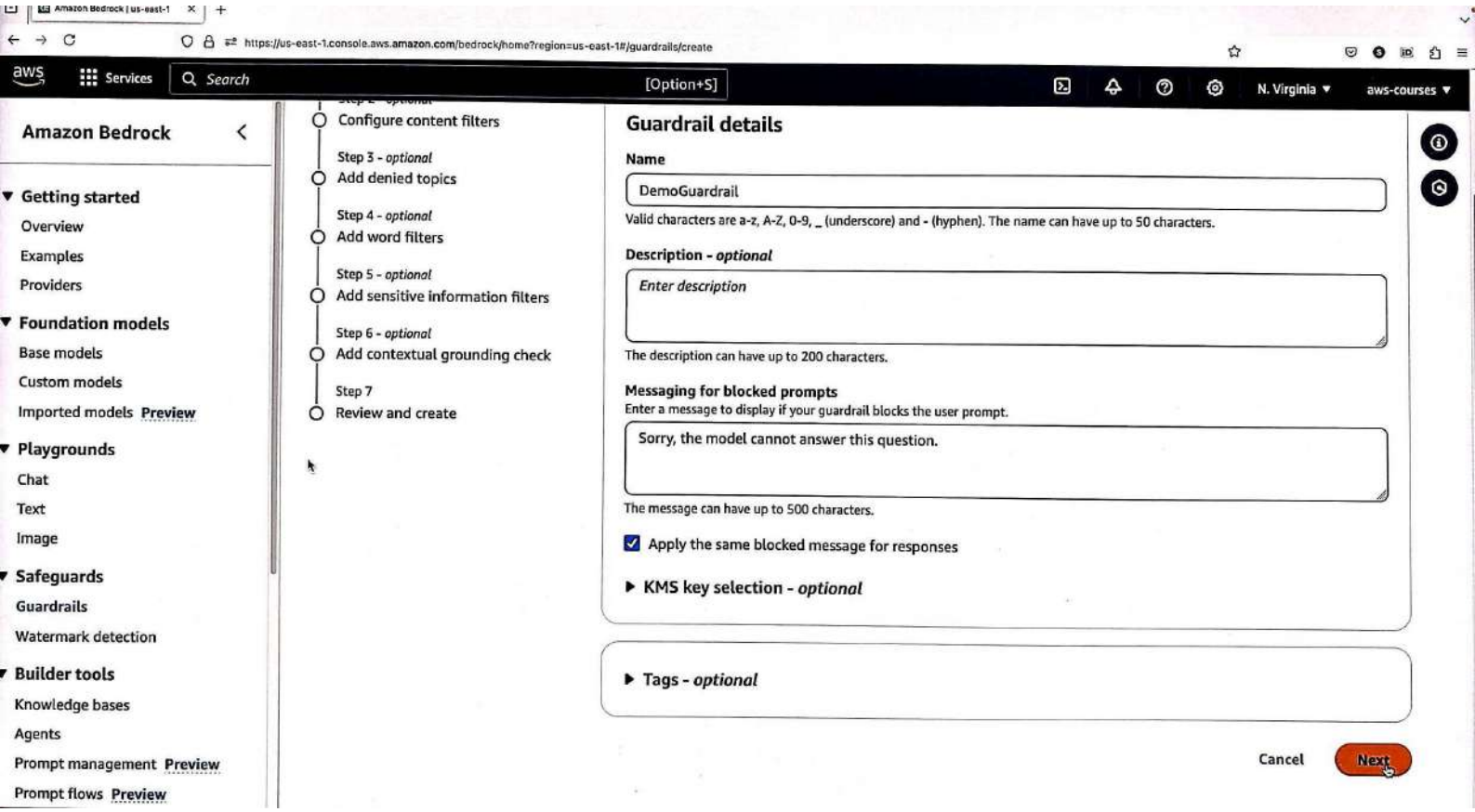


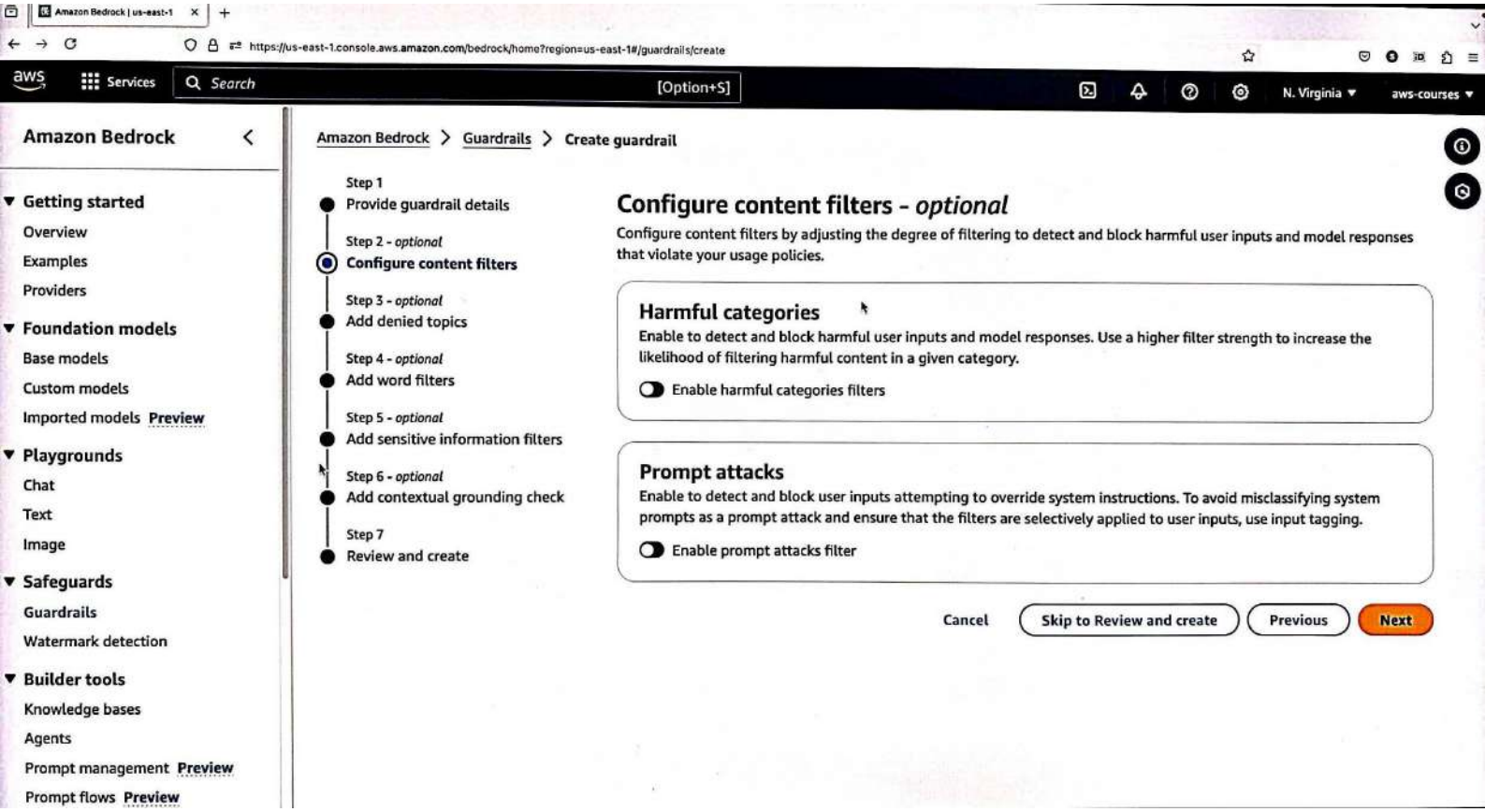
Amazon Bedrock – Guardrails

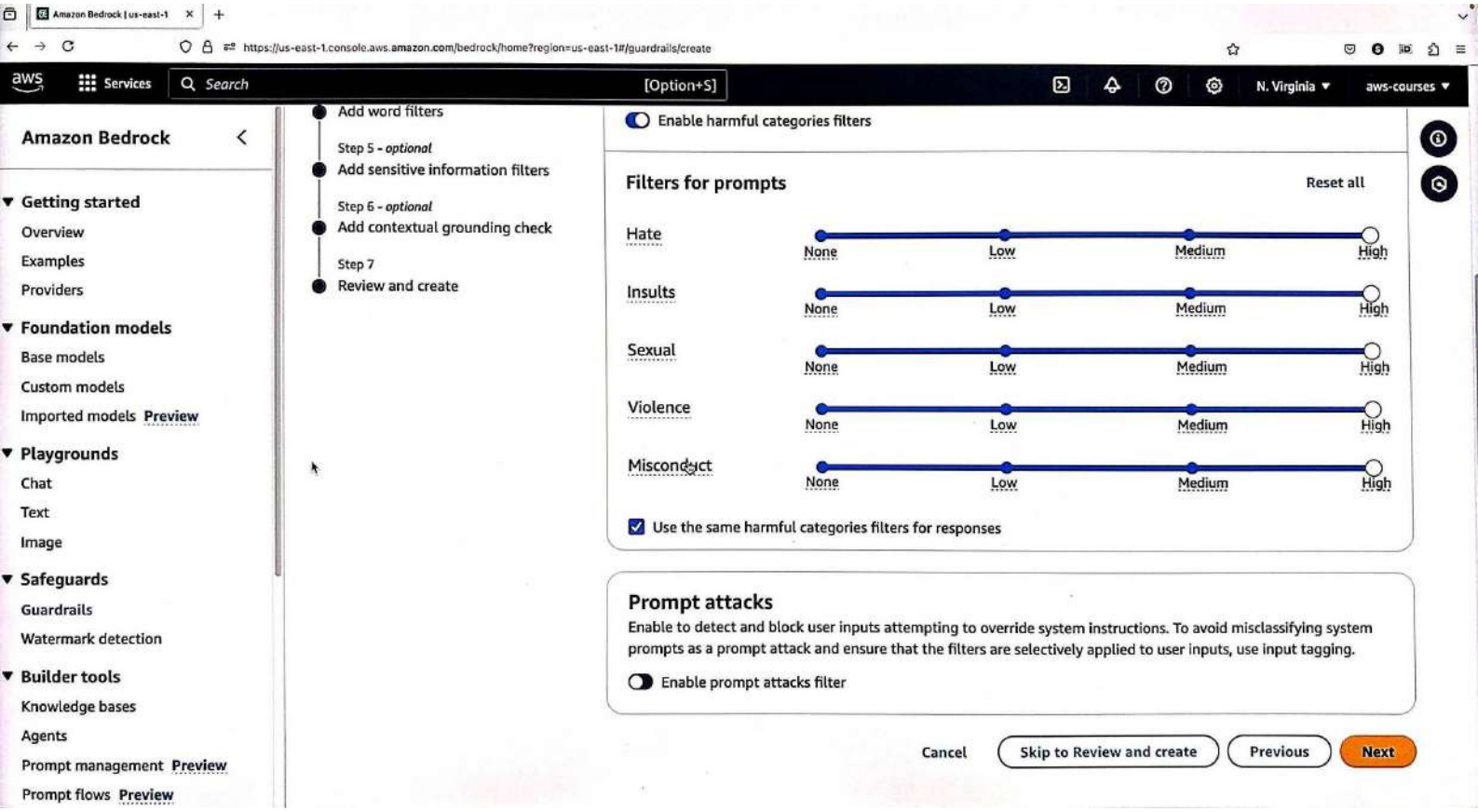
- Control the interaction between users and Foundation Models (FMs)
- Filter undesirable and harmful content
- Remove Personally Identifiable Information (PII)
- Enhanced privacy
- Reduce hallucinations
- Ability to create multiple Guardrails and monitor and analyze user inputs that can violate the Guardrails

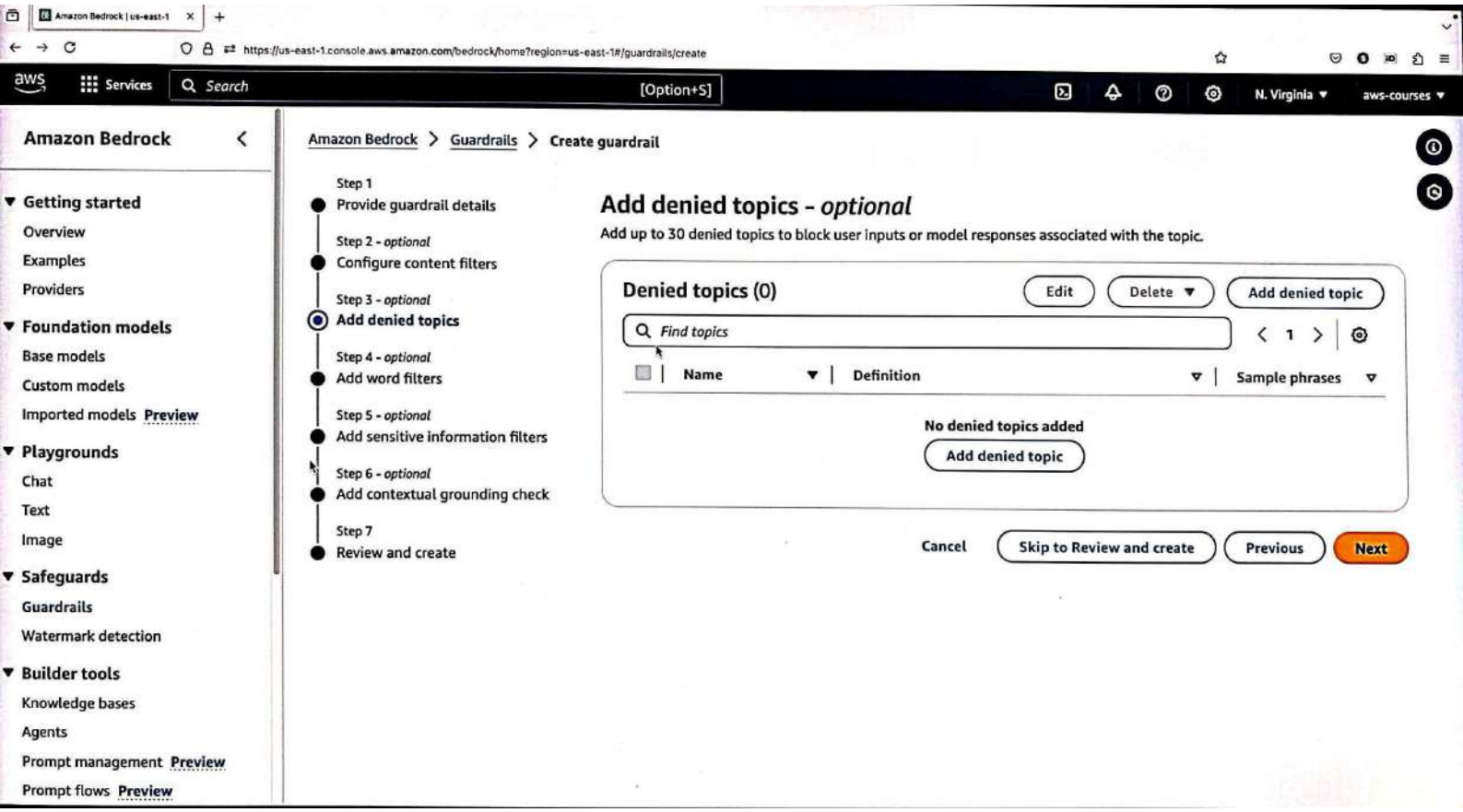



Amazon Bedrock | us-east-1
https://us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/guardrails
aws
Services
Search
[Option+S]
N. Virginia
aws-courses
Amazon Bedrock
Getting started
Overview
Examples
Providers
Foundation models
Base models
Custom models
Imported models Preview
Playgrounds
Chat
Text
Image
Safeguards
Guardrails
Watermark detection
Builder tools
Knowledge bases
Agents
Prompt management Preview
Prompt flows Preview
Amazon Bedrock > Guardrails
Guardrails
Guardrails for Amazon Bedrock are used to implement application-specific safeguards based on your use cases and responsible AI policies.
Overview
Create a guardrail
Create a guardrail by configuring as many filters as you need.
Test a guardrail
Test the guardrail with different inputs to assess the guardrail's performance. Refine the guardrail until it matches your needs.
Deploy a guardrail
Create a version of the guardrail to produce a snapshot that you can deploy during model inference or by attaching it to an agent.
Guardrails (0)
Edit
Delete
Create guardrail
Find guardrail
1
Name
Status
Description
Creation time
Last edited
No guardrails
No guardrails to display
Create guardrail


Amazon Bedrock | us-east-1
https://us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/guardrails/create
aws
Services
Search
[Option+S]
N. Virginia
aws-courses
Amazon Bedrock
Getting started
Overview
Examples
Providers
Foundation models
Base models
Custom models
Imported models Preview
Playgrounds
Chat
Text
Image
Safeguards
Guardrails
Watermark detection
Builder tools
Knowledge bases
Agents
Prompt management Preview
Prompt flows Preview
Configure content filters
Step 3 - optional
Add denied topics
Step 4 - optional
Add word filters
Step 5 - optional
Add sensitive information filters
Step 6 - optional
Add contextual grounding check
Step 7
Review and create
Guardrail details
Name
DemoGuardrail
Valid characters are a-z, A-Z, 0-9, _ (underscore) and - (hyphen). The name can have up to 50 characters.
Description - optional
Enter description
The description can have up to 200 characters.
Messaging for blocked prompts
Enter a message to display if your guardrail blocks the user prompt.
Sorry, the model cannot answer this question.
The message can have up to 500 characters.
Apply the same blocked message for responses
KMS key selection - optional
Tags - optional
Cancel
Next


Amazon Bedrock | us-east-1
https://us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/guardrails/create
aws
Services
Search
[Option+S]
N. Virginia
aws-courses
Amazon Bedrock
Getting started
Overview
Examples
Providers
Foundation models
Base models
Custom models
Imported models Preview
Playgrounds
Chat
Text
Image
Safeguards
Guardrails
Watermark detection
Builder tools
Knowledge bases
Agents
Prompt management Preview
Prompt flows Preview
Amazon Bedrock > Guardrails > Create guardrail
Step 1
Provide guardrail details
Step 2 - optional
Configure content filters
Step 3 - optional
Add denied topics
Step 4 - optional
Add word filters
Step 5 - optional
Add sensitive information filters
Step 6 - optional
Add contextual grounding check
Step 7
Review and create
Configure content filters - optional
Configure content filters by adjusting the degree of filtering to detect and block harmful user inputs and model responses that violate your usage policies.
Harmful categories
Enable to detect and block harmful user inputs and model responses. Use a higher filter strength to increase the likelihood of filtering harmful content in a given category.
Enable harmful categories filters
Prompt attacks
Enable to detect and block user inputs attempting to override system instructions. To avoid misclassifying system prompts as a prompt attack and ensure that the filters are selectively applied to user inputs, use input tagging.
Enable prompt attacks filter
Cancel
Skip to Review and create
Previous
Next


Amazon Bedrock | us-east-1
https://us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/guardrails/create
aws
Services
Search
[Option+S]
N. Virginia
aws-courses
Amazon Bedrock
Getting started
Overview
Examples
Providers
Foundation models
Base models
Custom models
Imported models Preview
Playgrounds
Chat
Text
Image
Safeguards
Guardrails
Watermark detection
Builder tools
Knowledge bases
Agents
Prompt management Preview
Prompt flows Preview
Add word filters
Step 5 - optional
Add sensitive information filters
Step 6 - optional
Add contextual grounding check
Step 7
Review and create
Enable harmful categories filters
Filters for prompts
Reset all
Hate
Insults
Sexual
Violence
Misconduct
None
Low
Medium
High
Use the same harmful categories filters for responses
Prompt attacks
Enable to detect and block user inputs attempting to override system instructions. To avoid misclassifying system prompts as a prompt attack and ensure that the filters are selectively applied to user inputs, use input tagging.
Enable prompt attacks filter
Cancel
Skip to Review and create
Previous
Next


Amazon Bedrock | us-east-1
https://us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/guardrails/create
aws
Services
Search
[Option+S]
N. Virginia
aws-courses
Amazon Bedrock
Getting started
Overview
Examples
Providers
Foundation models
Base models
Custom models
Imported models Preview
Playgrounds
Chat
Text
Image
Safeguards
Guardrails
Watermark detection
Builder tools
Knowledge bases
Agents
Prompt management Preview
Prompt flows Preview
Amazon Bedrock > Guardrails > Create guardrail
Step 1
Provide guardrail details
Step 2 - optional
Configure content filters
Step 3 - optional
Add denied topics
Step 4 - optional
Add word filters
Step 5 - optional
Add sensitive information filters
Step 6 - optional
Add contextual grounding check
Step 7
Review and create
Add denied topics - optional
Add up to 30 denied topics to block user inputs or model responses associated with the topic.
Denied topics (0)
Edit
Delete
Add denied topic
Find topics
1
Name
Definition
Sample phrases
No denied topics added
Add denied topic
Cancel
Skip to Review and create
Previous
Next

Add denied topic

Name

Recipes

Valid characters are a-z, A-Z, 0-9, underscore (_), hyphen (-), space, exclamation point (!), question mark (?), and period (.). The name can have up to 100 characters.

Definition for topic

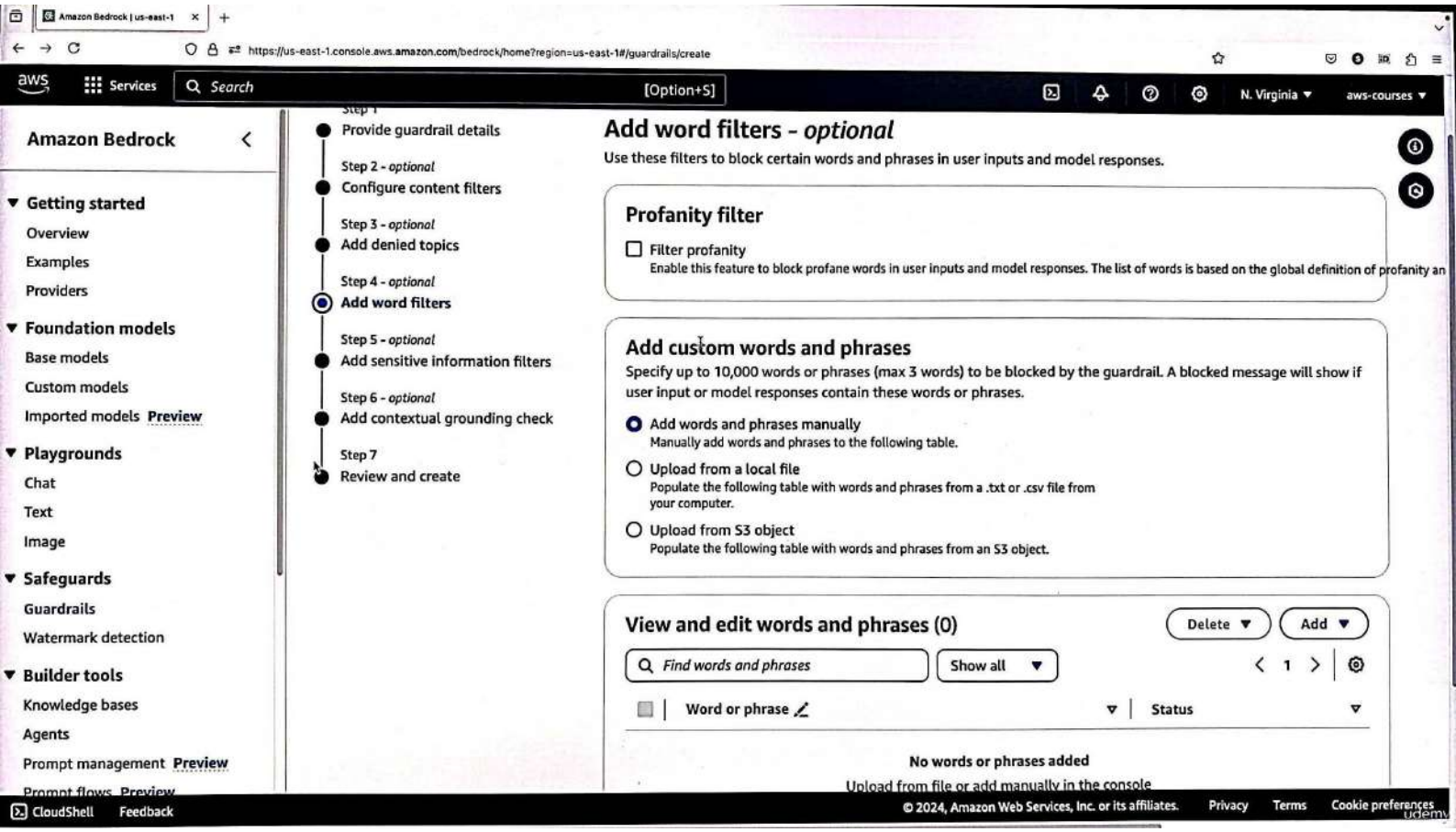
Provide a clear definition to detect and block user inputs and FM responses that fall into this topic. Avoid starting with "don't".

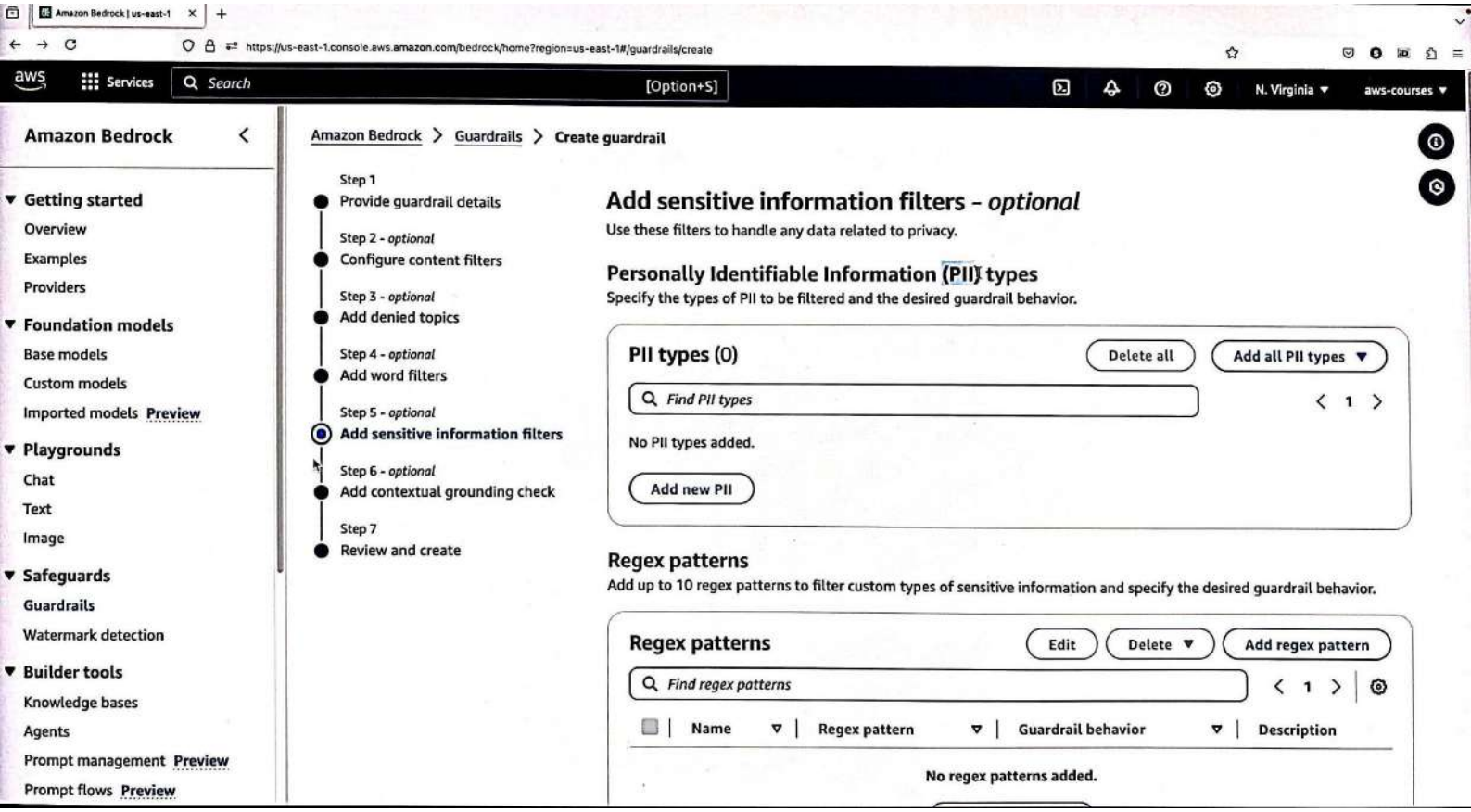
Food recipes are instructions on how to cook specific dishes.

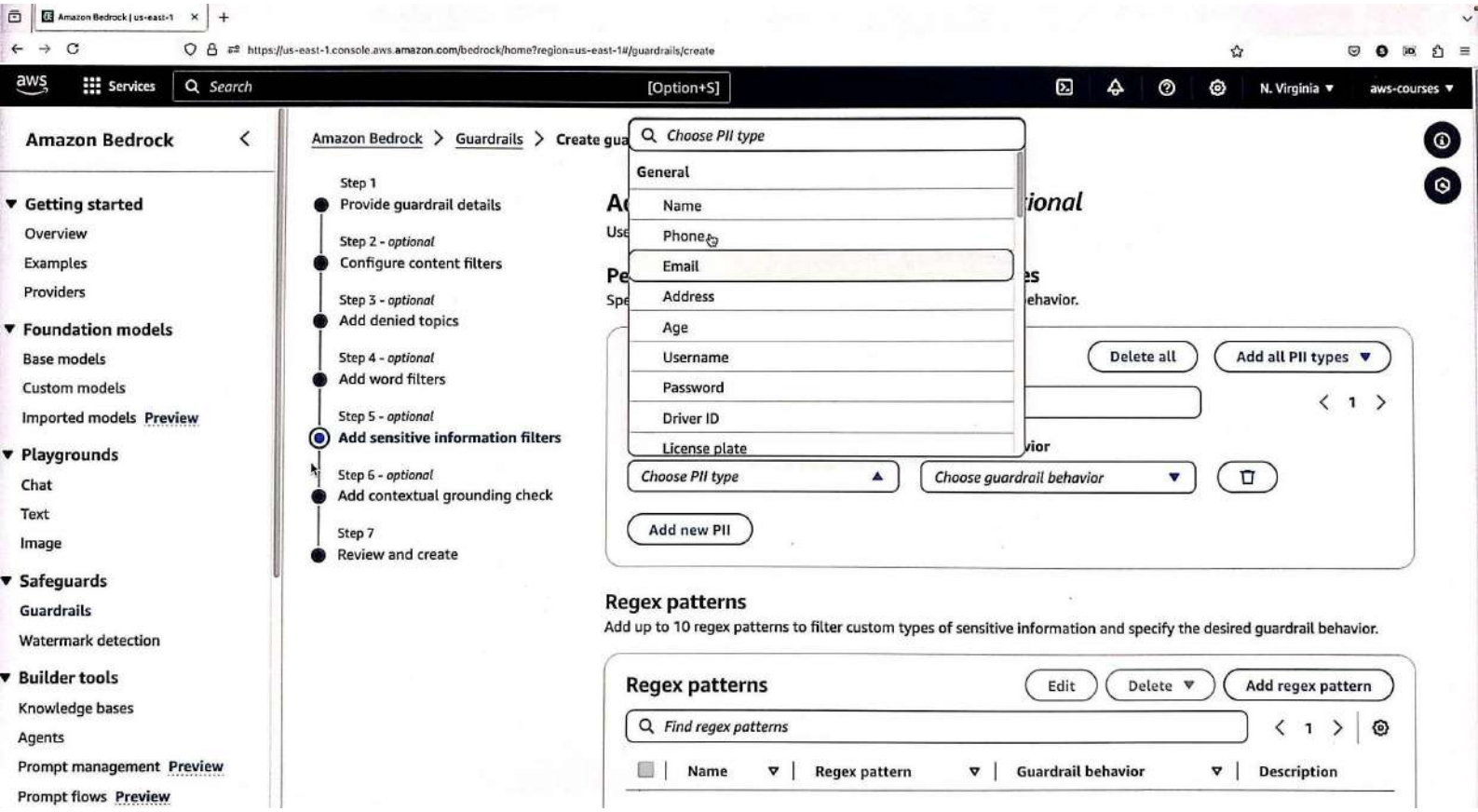
The definition can have up to 200 characters.

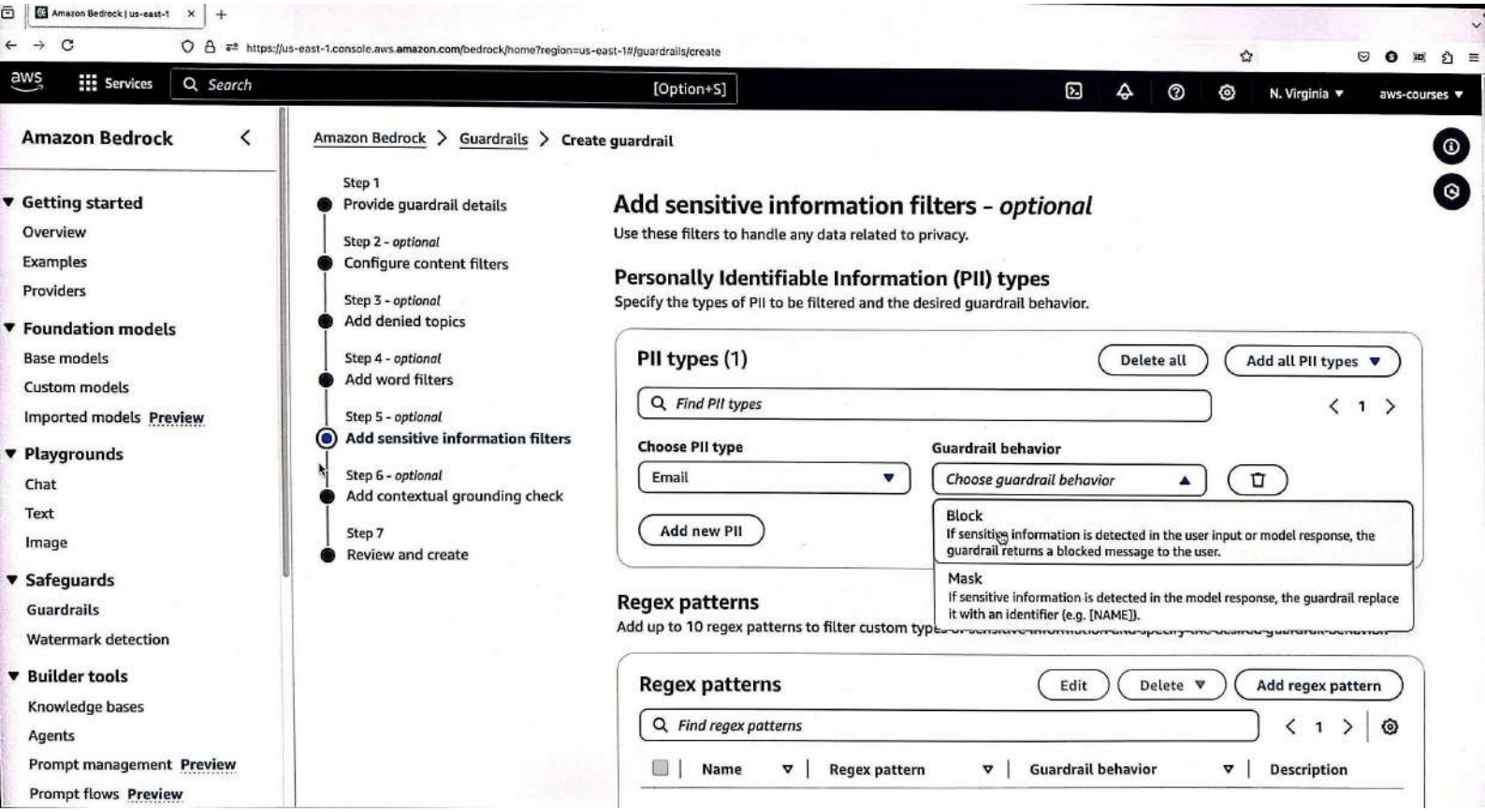
▶ **Add sample phrases - *optional***

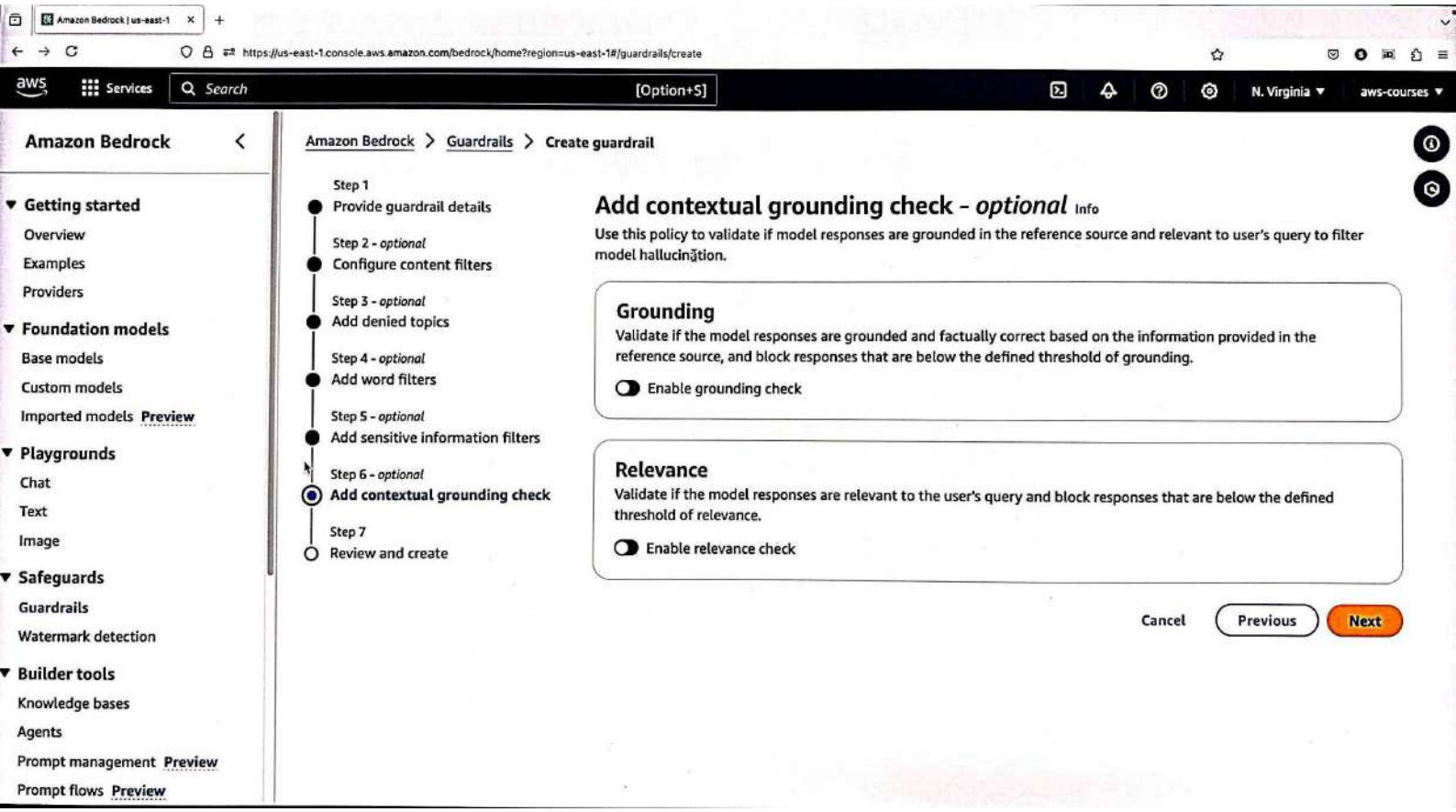
Cancel Confirm

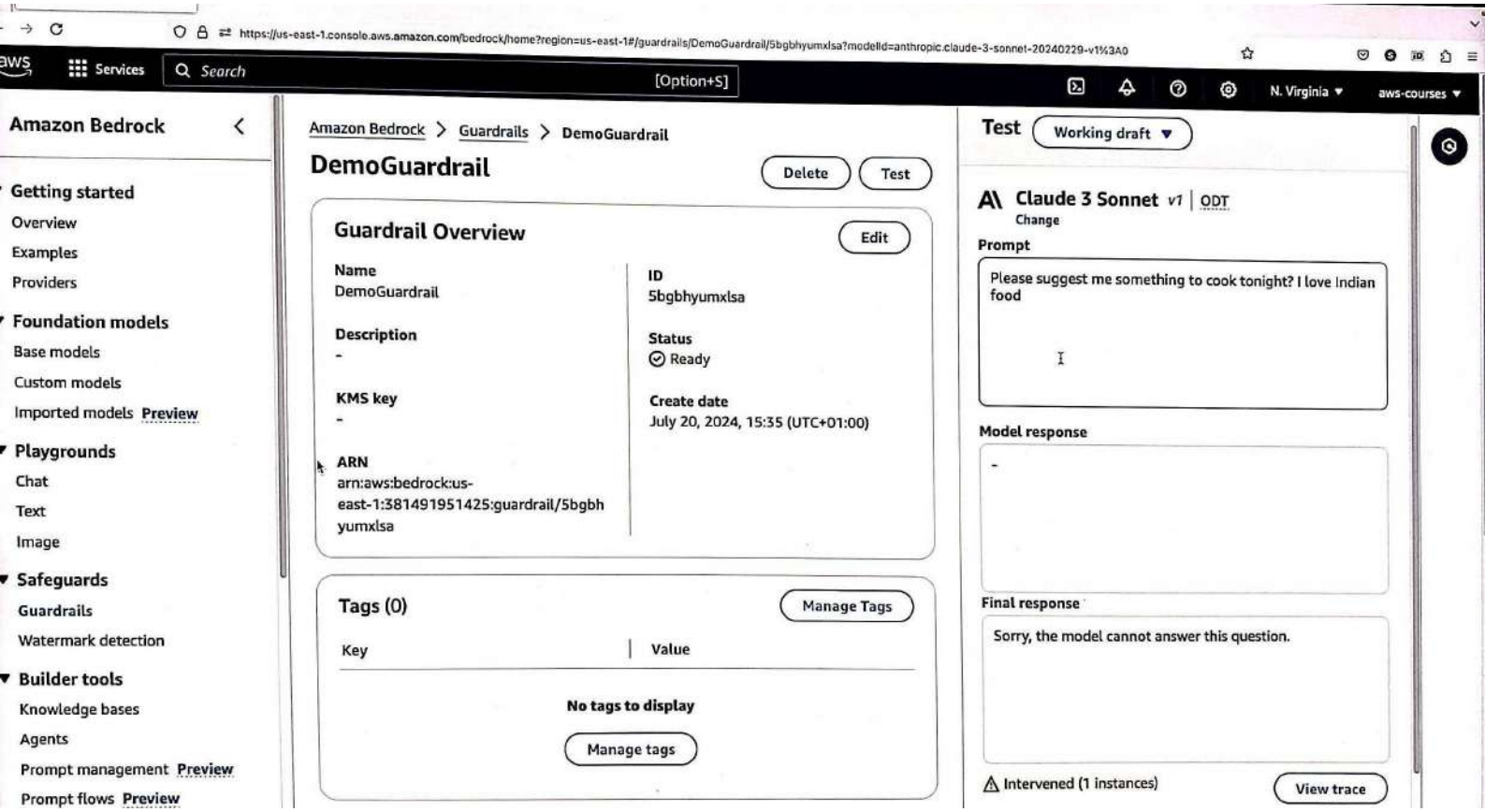

Amazon Bedrock | us-east-1
https://us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/guardrails/create
aws
Services
Search
[Option+S]
N. Virginia
aws-courses
Amazon Bedrock
Getting started
Overview
Examples
Providers
Foundation models
Base models
Custom models
Imported models Preview
Playgrounds
Chat
Text
Image
Safeguards
Guardrails
Watermark detection
Builder tools
Knowledge bases
Agents
Prompt management Preview
Prompt flows Preview
Step 1
Provide guardrail details
Step 2 - optional
Configure content filters
Step 3 - optional
Add denied topics
Step 4 - optional
Add word filters
Step 5 - optional
Add sensitive information filters
Step 6 - optional
Add contextual grounding check
Step 7
Review and create
Add word filters - optional
Use these filters to block certain words and phrases in user inputs and model responses.
Profanity filter
Filter profanity
Enable this feature to block profane words in user inputs and model responses. The list of words is based on the global definition of profanity an
Add custom words and phrases
Specify up to 10,000 words or phrases (max 3 words) to be blocked by the guardrail. A blocked message will show if user input or model responses contain these words or phrases.
Add words and phrases manually
Manually add words and phrases to the following table.
Upload from a local file
Populate the following table with words and phrases from a .txt or .csv file from your computer.
Upload from S3 object
Populate the following table with words and phrases from an S3 object.
View and edit words and phrases (0)
Delete
Add
Find words and phrases
Show all
1
Word or phrase
Status
No words or phrases added
Upload from file or add manually in the console
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

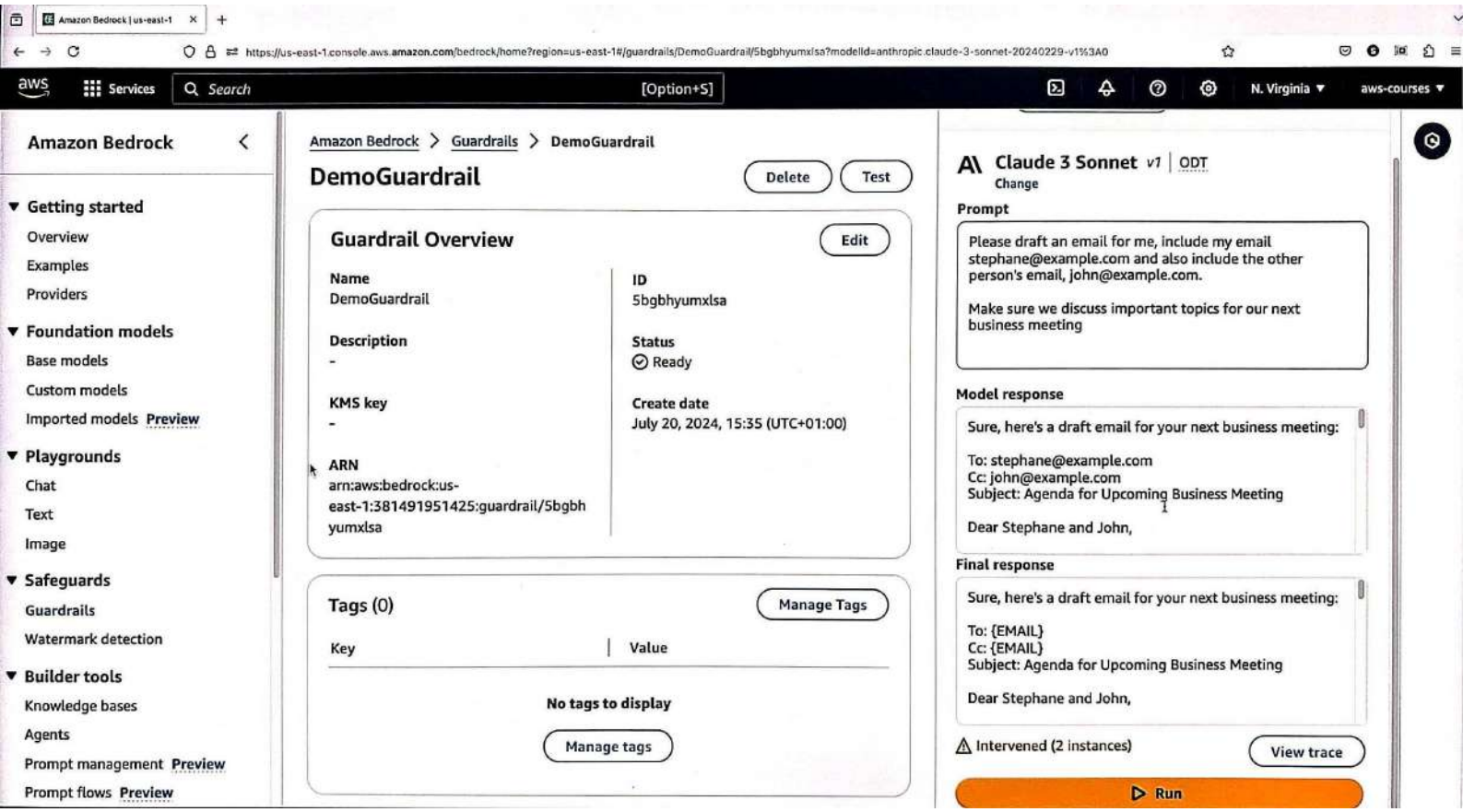

Amazon Bedrock | us-east-1
https://us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/guardrails/create
aws
Services
Search
[Option+S]
N. Virginia
aws-courses
Amazon Bedrock
Getting started
Overview
Examples
Providers
Foundation models
Base models
Custom models
Imported models Preview
Playgrounds
Chat
Text
Image
Safeguards
Guardrails
Watermark detection
Builder tools
Knowledge bases
Agents
Prompt management Preview
Prompt flows Preview
Amazon Bedrock > Guardrails > Create guardrail
Step 1
Provide guardrail details
Step 2 - optional
Configure content filters
Step 3 - optional
Add denied topics
Step 4 - optional
Add word filters
Step 5 - optional
Add sensitive information filters
Step 6 - optional
Add contextual grounding check
Step 7
Review and create
Add sensitive information filters - optional
Use these filters to handle any data related to privacy.
Personally Identifiable Information (PII) types
Specify the types of PII to be filtered and the desired guardrail behavior.
PII types (0)
Delete all
Add all PII types
Find PII types
1
No PII types added.
Add new PII
Regex patterns
Add up to 10 regex patterns to filter custom types of sensitive information and specify the desired guardrail behavior.
Regex patterns
Edit
Delete
Add regex pattern
Find regex patterns
1
Name
Regex pattern
Guardrail behavior
Description
No regex patterns added.


Amazon Bedrock | us-east-1
https://us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/guardrails/create
aws
Services
Search
[Option+S]
N. Virginia
aws-courses
Amazon Bedrock
Getting started
Overview
Examples
Providers
Foundation models
Base models
Custom models
Imported models Preview
Playgrounds
Chat
Text
Image
Safeguards
Guardrails
Watermark detection
Builder tools
Knowledge bases
Agents
Prompt management Preview
Prompt flows Preview
Amazon Bedrock > Guardrails > Create gua
Step 1
Provide guardrail details
Step 2 - optional
Configure content filters
Step 3 - optional
Add denied topics
Step 4 - optional
Add word filters
Step 5 - optional
Add sensitive information filters
Step 6 - optional
Add contextual grounding check
Step 7
Review and create
Choose PII type
General
Name
Phone
Email
Address
Age
Username
Password
Driver ID
License plate
Delete all
Add all PII types
1
Choose PII type
Choose guardrail behavior
Add new PII
Regex patterns
Add up to 10 regex patterns to filter custom types of sensitive information and specify the desired guardrail behavior.
Regex patterns
Edit
Delete
Add regex pattern
Find regex patterns
1
Name
Regex pattern
Guardrail behavior
Description

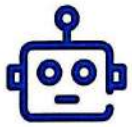

Amazon Bedrock | us-east-1
https://us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/guardrails/create
aws
Services
Search
[Option+S]
N. Virginia
aws-courses
Amazon Bedrock
Getting started
Overview
Examples
Providers
Foundation models
Base models
Custom models
Imported models Preview
Playgrounds
Chat
Text
Image
Safeguards
Guardrails
Watermark detection
Builder tools
Knowledge bases
Agents
Prompt management Preview
Prompt flows Preview
Amazon Bedrock > Guardrails > Create guardrail
Step 1
Provide guardrail details
Step 2 - optional
Configure content filters
Step 3 - optional
Add denied topics
Step 4 - optional
Add word filters
Step 5 - optional
Add sensitive information filters
Step 6 - optional
Add contextual grounding check
Step 7
Review and create
Add sensitive information filters - optional
Use these filters to handle any data related to privacy.
Personally Identifiable Information (PII) types
Specify the types of PII to be filtered and the desired guardrail behavior.
PII types (1)
Delete all
Add all PII types
Find PII types
1
Choose PII type
Email
Guardrail behavior
Choose guardrail behavior
Add new PII
Block
If sensitive information is detected in the user input or model response, the guardrail returns a blocked message to the user.
Mask
If sensitive information is detected in the model response, the guardrail replace it with an identifier (e.g. [NAME]).
Regex patterns
Add up to 10 regex patterns to filter custom types
Regex patterns
Edit
Delete
Add regex pattern
Find regex patterns
1
Name
Regex pattern
Guardrail behavior
Description


Amazon Bedrock | us-east-1
https://us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/guardrails/create
aws
Services
Search
[Option+S]
N. Virginia
aws-courses
Amazon Bedrock
Getting started
Overview
Examples
Providers
Foundation models
Base models
Custom models
Imported models Preview
Playgrounds
Chat
Text
Image
Safeguards
Guardrails
Watermark detection
Builder tools
Knowledge bases
Agents
Prompt management Preview
Prompt flows Preview
Amazon Bedrock > Guardrails > Create guardrail
Step 1
Provide guardrail details
Step 2 - optional
Configure content filters
Step 3 - optional
Add denied topics
Step 4 - optional
Add word filters
Step 5 - optional
Add sensitive information filters
Step 6 - optional
Add contextual grounding check
Step 7
Review and create
Add contextual grounding check - optional Info
Use this policy to validate if model responses are grounded in the reference source and relevant to user's query to filter model hallucination.
Grounding
Validate if the model responses are grounded and factually correct based on the information provided in the reference source, and block responses that are below the defined threshold of grounding.
Enable grounding check
Relevance
Validate if the model responses are relevant to the user's query and block responses that are below the defined threshold of relevance.
Enable relevance check
Cancel
Previous
Next


https://us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/guardrails/DemoGuardrail/5bgbhyumxlsa?modelId=anthropic.claude-3-sonnet-20240229-v1%3A0
Services
Search
[Option+S]
N. Virginia
aws-courses
Amazon Bedrock
Getting started
Overview
Examples
Providers
Foundation models
Base models
Custom models
Imported models Preview
Playgrounds
Chat
Text
Image
Safeguards
Guardrails
Watermark detection
Builder tools
Knowledge bases
Agents
Prompt management Preview
Prompt flows Preview
Amazon Bedrock > Guardrails > DemoGuardrail
DemoGuardrail
Delete
Test
Guardrail Overview
Edit
Name
DemoGuardrail
Description
-
KMS key
-
ARN
arn:aws:bedrock:us-east-1:381491951425:guardrail/5bgbhyumxlsa
ID
5bgbhyumxlsa
Status
Ready
Create date
July 20, 2024, 15:35 (UTC+01:00)
Tags (0)
Manage Tags
Key
Value
No tags to display
Manage tags
Test
Working draft
Claude 3 Sonnet v1 | ODT
Change
Prompt
Please suggest me something to cook tonight? I love Indian food
Model response
-
Final response
Sorry, the model cannot answer this question.
Intervened (1 instances)
View trace


Amazon Bedrock | us-east-1
https://us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/guardrails/DemoGuardrail/5bgbhyumxlsa?modelId=anthropic.claude-3-sonnet-20240229-v1%3A0
aws
Services
Search
[Option+S]
N. Virginia
aws-courses
Amazon Bedrock
Getting started
Overview
Examples
Providers
Foundation models
Base models
Custom models
Imported models Preview
Playgrounds
Chat
Text
Image
Safeguards
Guardrails
Watermark detection
Builder tools
Knowledge bases
Agents
Prompt management Preview
Prompt flows Preview
Amazon Bedrock > Guardrails > DemoGuardrail
DemoGuardrail
Delete
Test
Guardrail Overview
Edit
Name
DemoGuardrail
ID
5bgbhyumxlsa
Description
-
Status
Ready
KMS key
-
Create date
July 20, 2024, 15:35 (UTC+01:00)
ARN
arn:aws:bedrock:us-east-1:381491951425:guardrail/5bgbhyumxlsa
Tags (0)
Manage Tags
Key
Value
No tags to display
Manage tags
Claude 3 Sonnet v1 | ODT
Change
Prompt
Please draft an email for me, include my email stephane@example.com and also include the other person's email, john@example.com.
Make sure we discuss important topics for our next business meeting
Model response
Sure, here's a draft email for your next business meeting:
To: stephane@example.com
Cc: john@example.com
Subject: Agenda for Upcoming Business Meeting
Dear Stephane and John,
Final response
Sure, here's a draft email for your next business meeting:
To: {EMAIL}
Cc: {EMAIL}
Subject: Agenda for Upcoming Business Meeting
Dear Stephane and John,
Intervened (2 instances)
View trace
Run

Amazon Bedrock – Agents

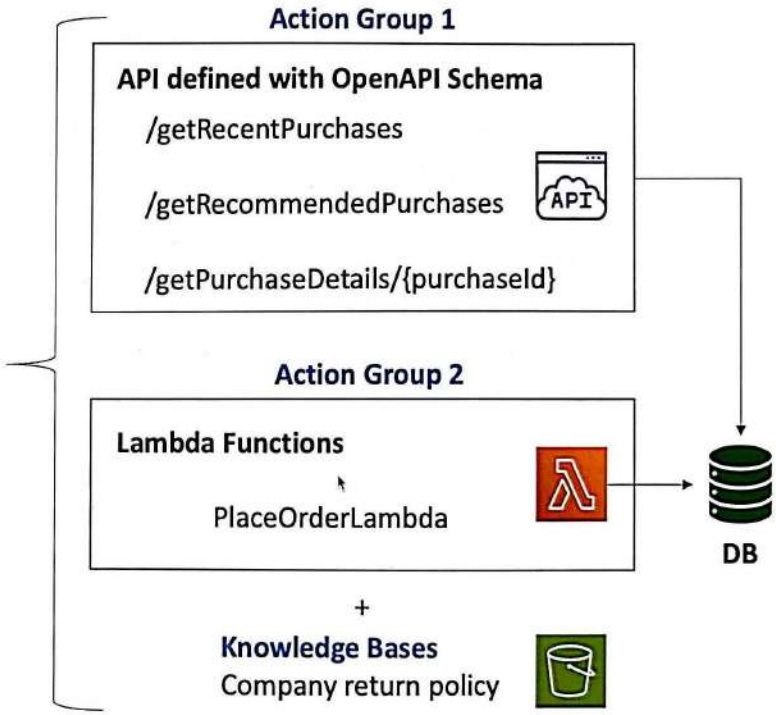


- Manage and carry out **various multi-step tasks** related to infrastructure provisioning, application deployment, and operational activities
- Task coordination: perform tasks in the correct order and ensure information is passed correctly between tasks
- Agents are configured to perform specific pre-defined action groups
- Integrate with other systems, services, databases and API to exchange data or initiate actions
- Leverage RAG to retrieve information when necessary

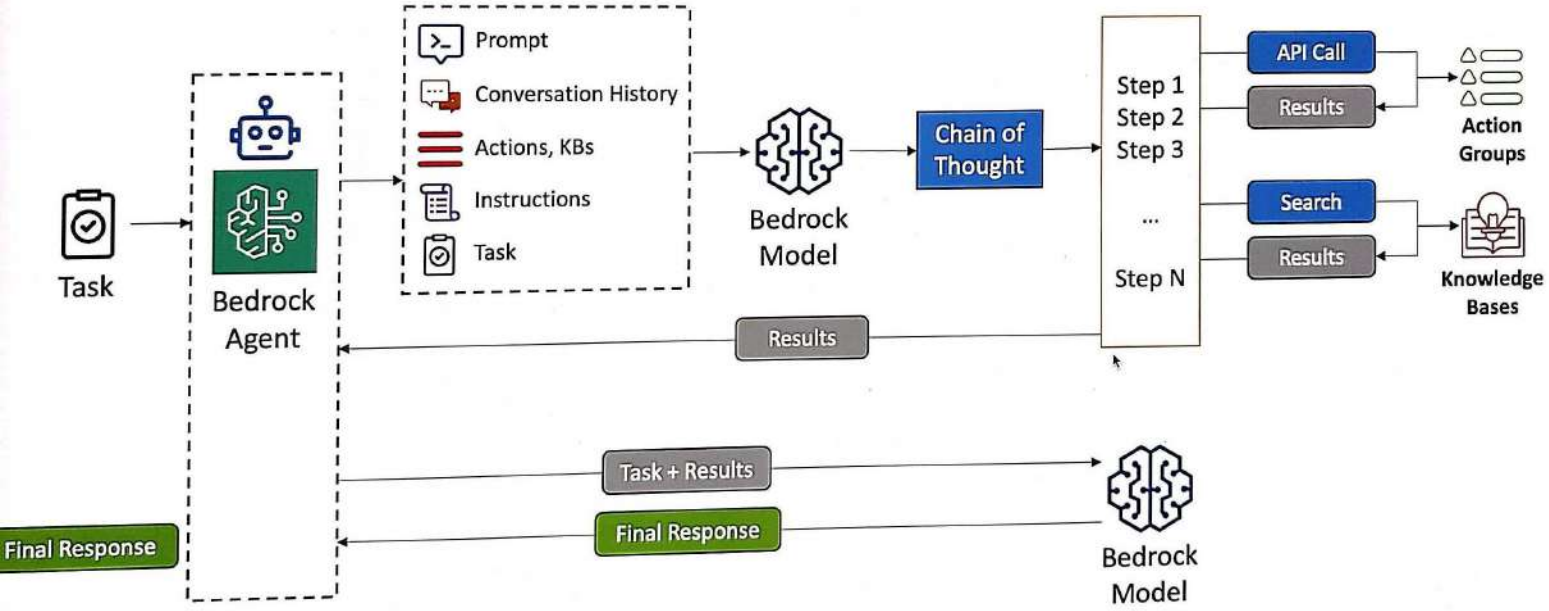
Bedrock Agent Setup

Instructions for the Agent

You are an agent responsible for accessing purchase history for our customers, as well as recommendations into what they can purchase next. You are also responsible for placing new orders.

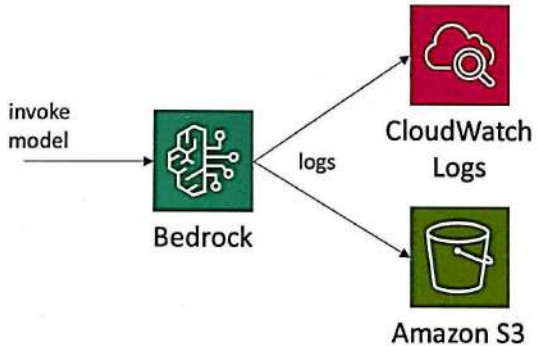


Agent - Diagram

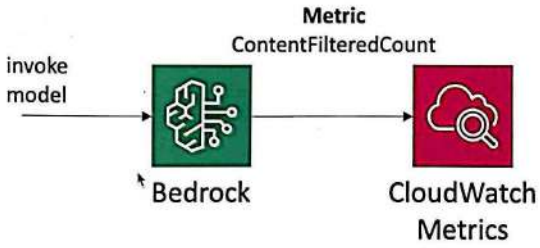


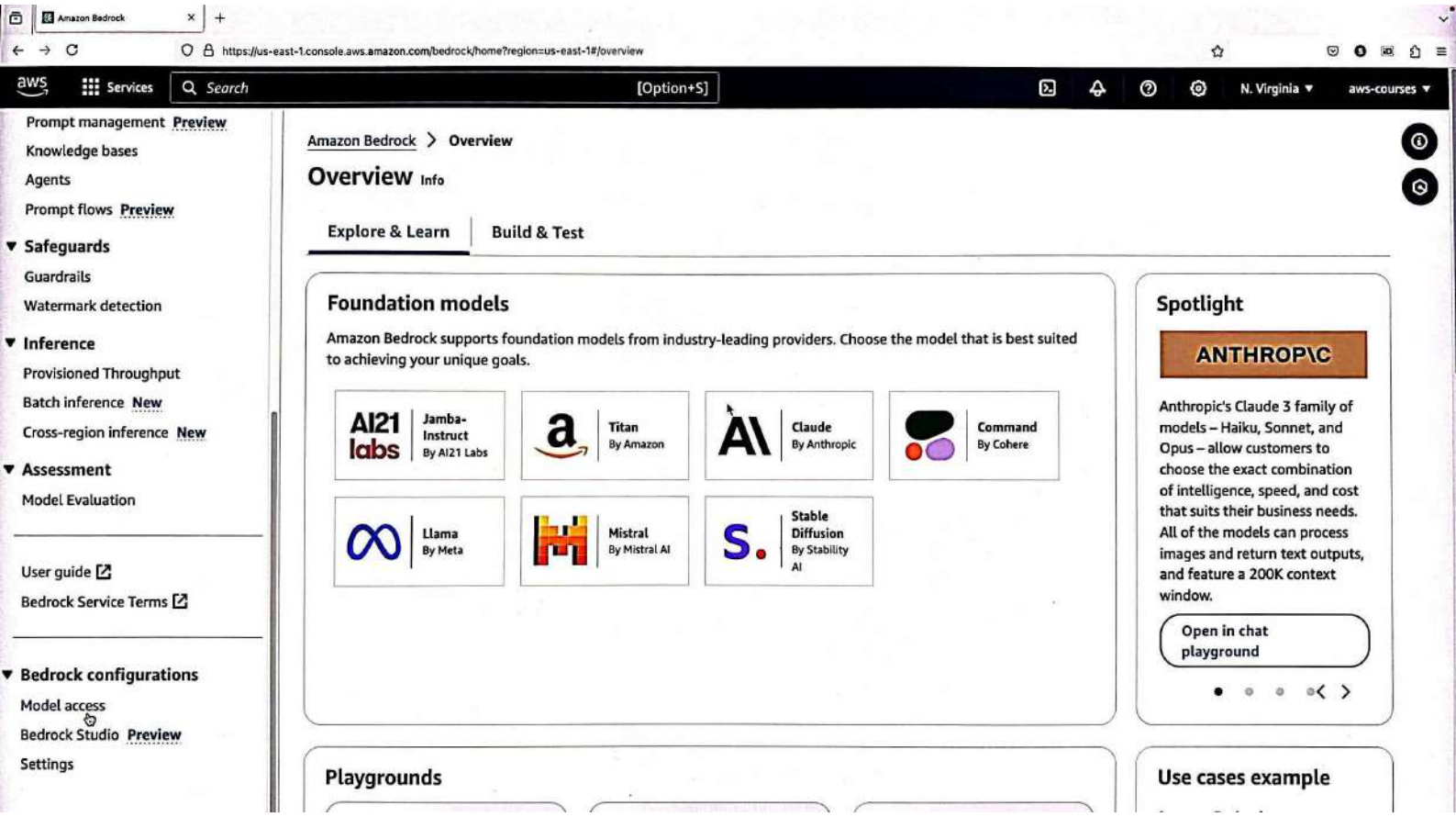
Amazon Bedrock & CloudWatch

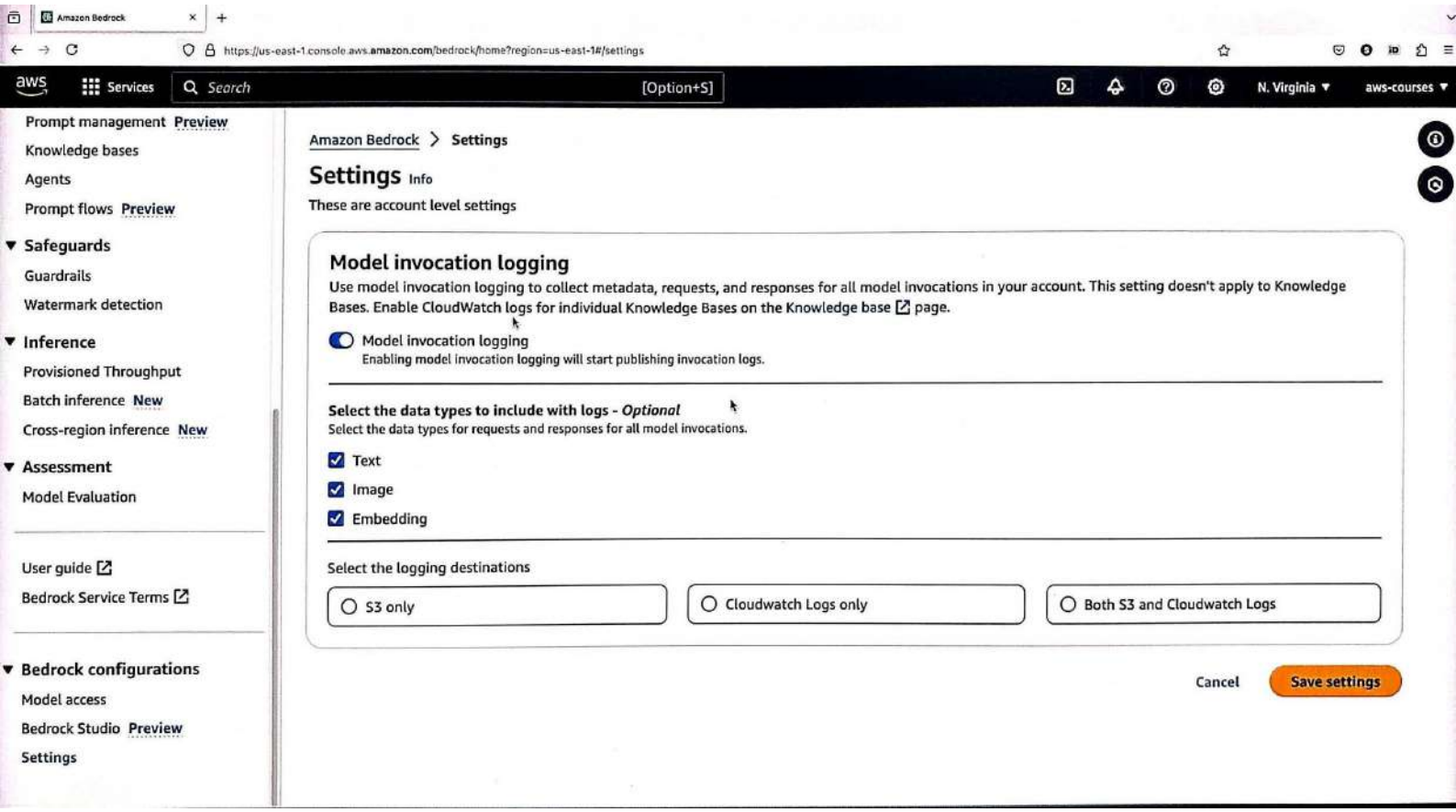
- Model Invocation Logging
  - Send logs of all invocations to Amazon CloudWatch and S3
  - Can include text, images and embeddings
  - Analyze further and build alerting thanks to CloudWatch Logs Insights

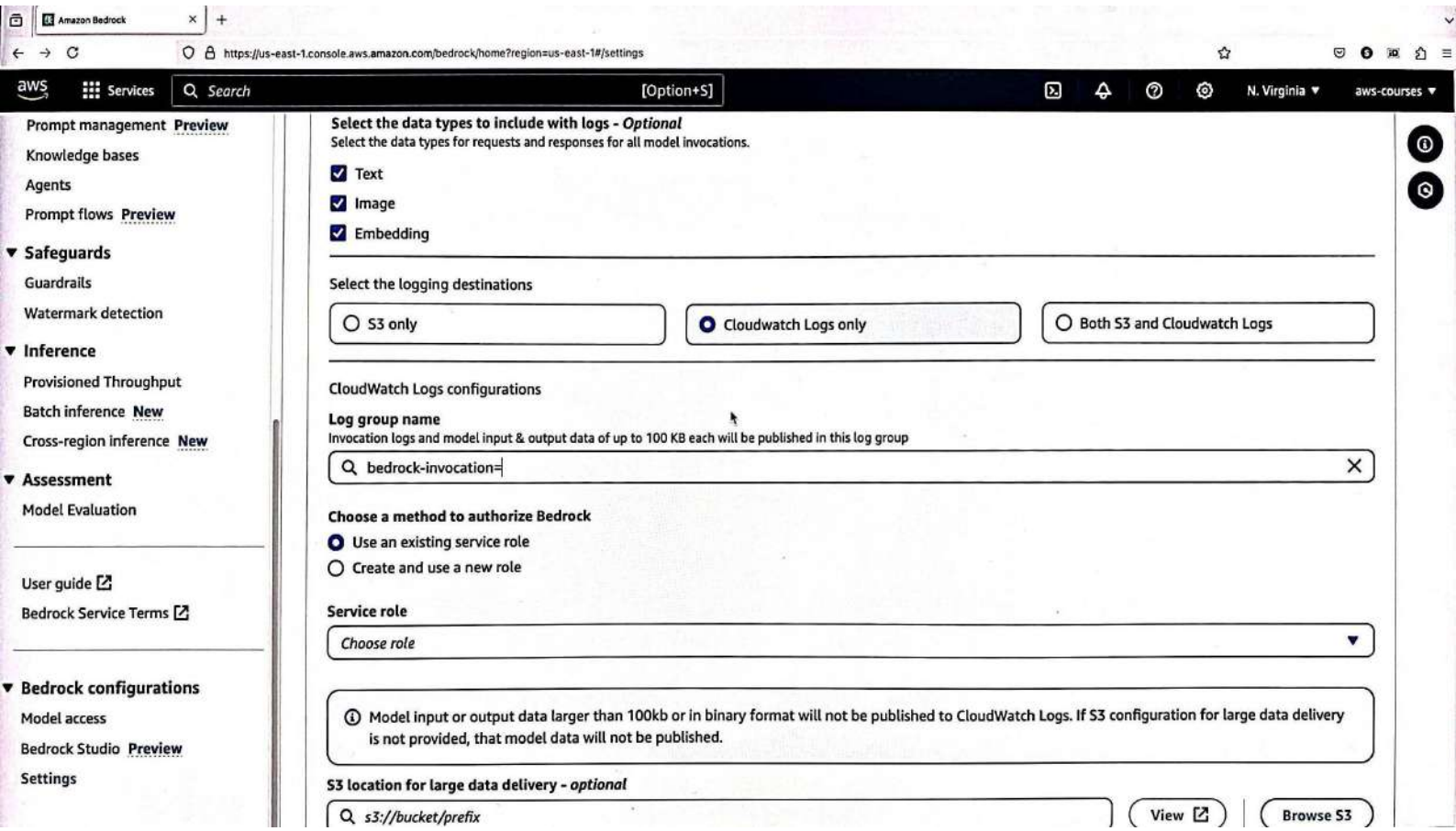


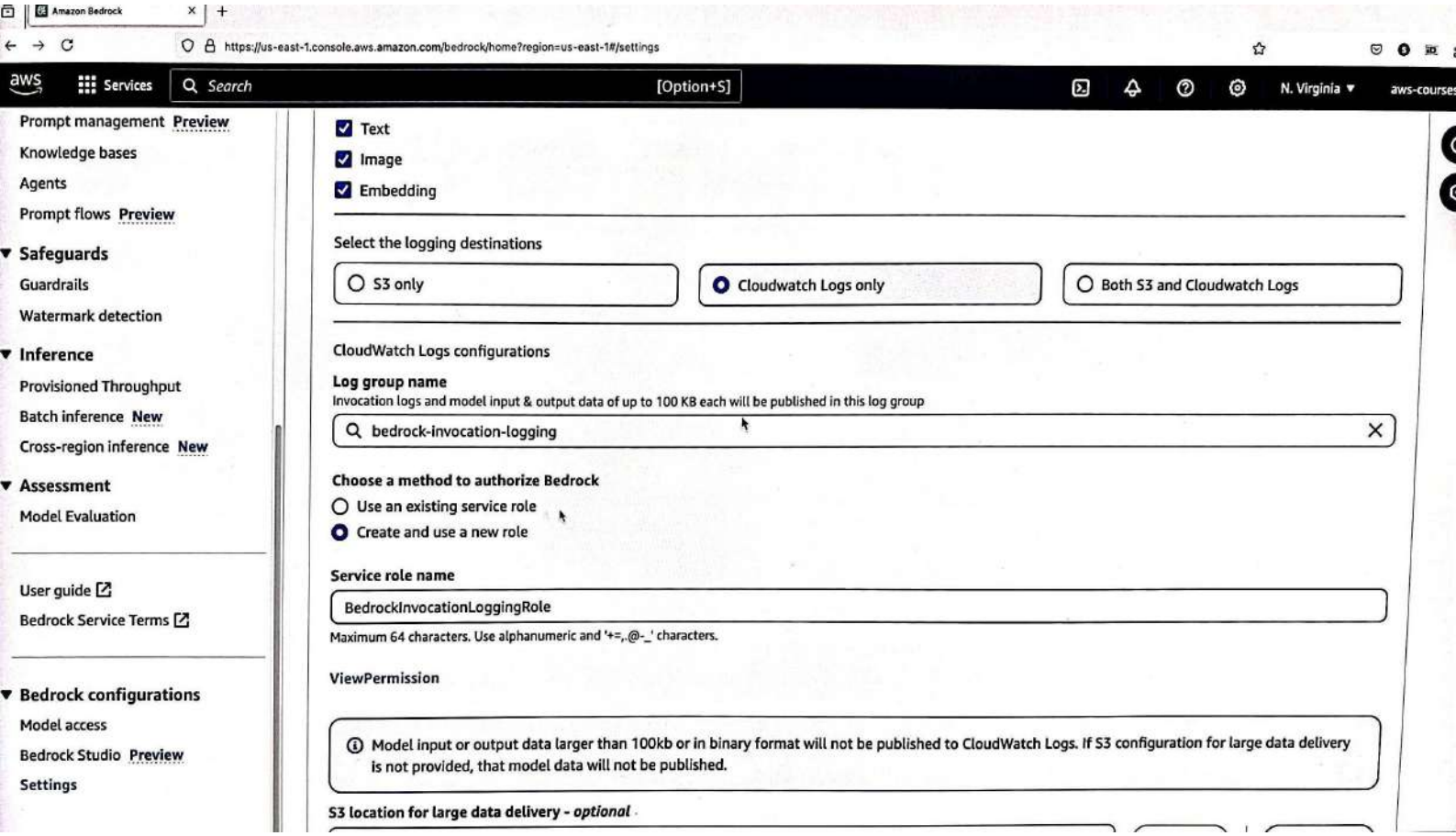
- CloudWatch Metrics
  - Published metrics from Bedrock to CloudWatch
    - Including *ContentFilteredCount*, which helps to see if Guardrails are functioning
  - Can build CloudWatch Alarms on top of Metrics

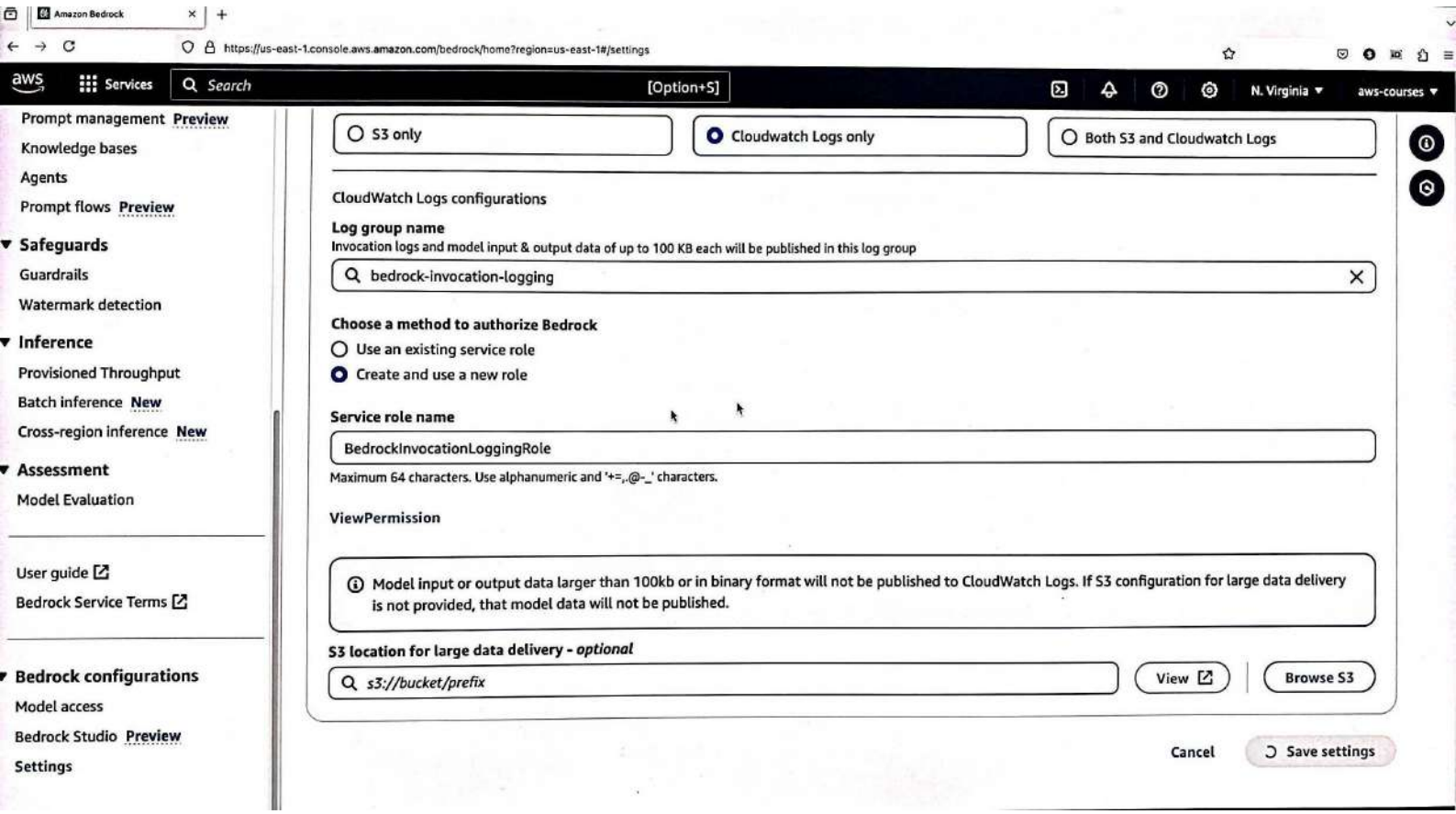


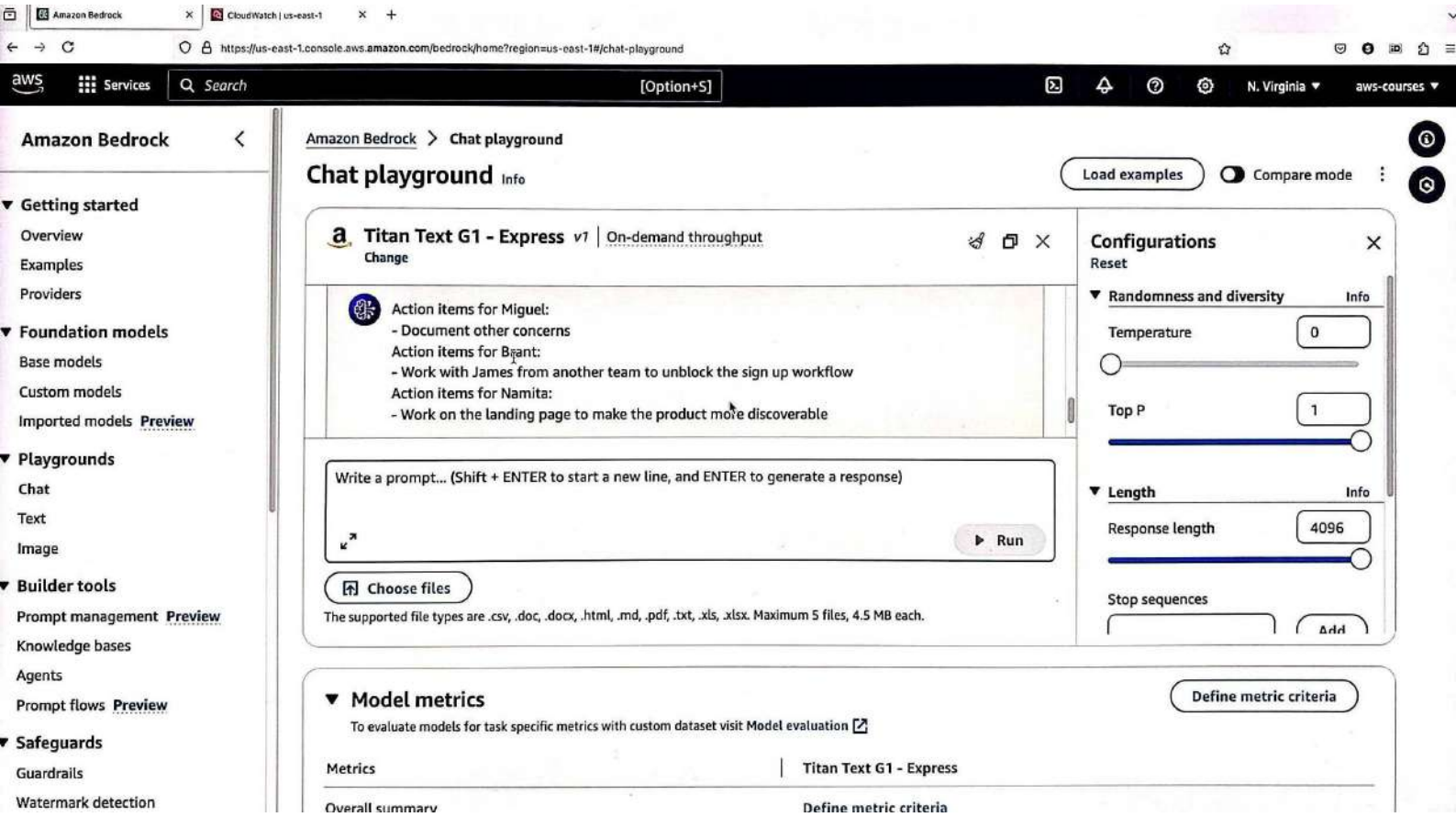

Amazon Bedrock
https://us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/overview
aws
Services
Search
[Option+S]
N. Virginia
aws-courses
Prompt management Preview
Knowledge bases
Agents
Prompt flows Preview
Safeguards
Guardrails
Watermark detection
Inference
Provisioned Throughput
Batch inference New
Cross-region inference New
Assessment
Model Evaluation
User guide
Bedrock Service Terms
Bedrock configurations
Model access
Bedrock Studio Preview
Settings
Amazon Bedrock > Overview
Overview Info
Explore & Learn
Build & Test
Foundation models
Amazon Bedrock supports foundation models from industry-leading providers. Choose the model that is best suited to achieving your unique goals.
AI21 labs
Jamba-Instruct
By AI21 Labs
Titan
By Amazon
Claude
By Anthropic
Command
By Cohere
Llama
By Meta
Mistral
By Mistral AI
Stable Diffusion
By Stability AI
Spotlight
ANTHROPIC
Anthropic's Claude 3 family of models – Haiku, Sonnet, and Opus – allow customers to choose the exact combination of intelligence, speed, and cost that suits their business needs. All of the models can process images and return text outputs, and feature a 200K context window.
Open in chat playground
Playgrounds
Use cases example

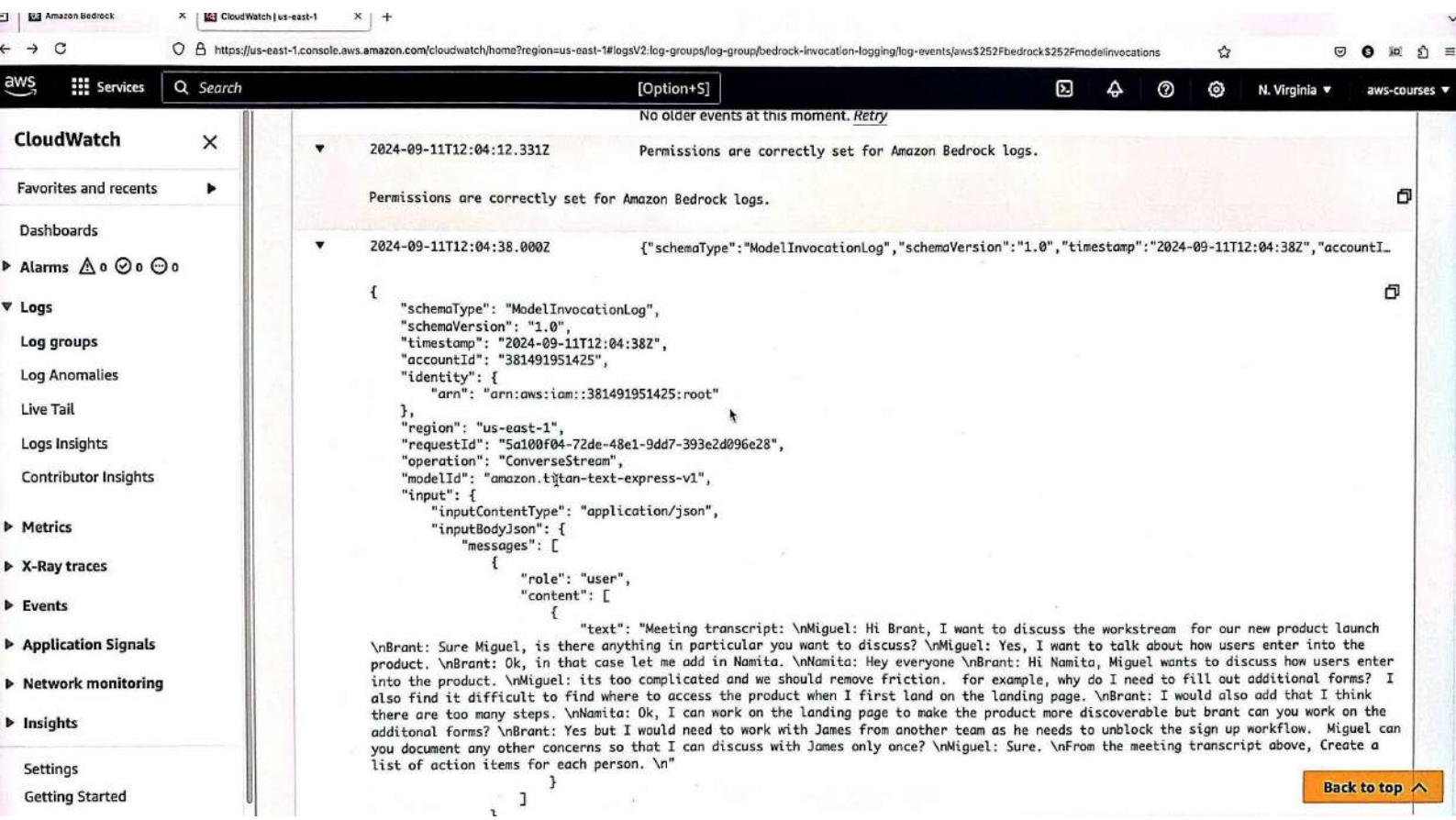

Amazon Bedrock
https://us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/settings
aws
Services
Search
[Option+S]
N. Virginia
aws-courses
Prompt management Preview
Knowledge bases
Agents
Prompt flows Preview
Safeguards
Guardrails
Watermark detection
Inference
Provisioned Throughput
Batch inference New
Cross-region inference New
Assessment
Model Evaluation
User guide
Bedrock Service Terms
Bedrock configurations
Model access
Bedrock Studio Preview
Settings
Amazon Bedrock > Settings
Settings Info
These are account level settings
Model invocation logging
Use model invocation logging to collect metadata, requests, and responses for all model invocations in your account. This setting doesn't apply to Knowledge Bases. Enable CloudWatch logs for individual Knowledge Bases on the Knowledge base page.
Model invocation logging
Enabling model invocation logging will start publishing invocation logs.
Select the data types to include with logs - Optional
Select the data types for requests and responses for all model invocations.
Text
Image
Embedding
Select the logging destinations
S3 only
Cloudwatch Logs only
Both S3 and Cloudwatch Logs
Cancel
Save settings


Amazon Bedrock
https://us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/settings
aws
Services
Search
[Option+S]
N. Virginia
aws-courses
Prompt management Preview
Knowledge bases
Agents
Prompt flows Preview
Safeguards
Guardrails
Watermark detection
Inference
Provisioned Throughput
Batch inference New
Cross-region inference New
Assessment
Model Evaluation
User guide
Bedrock Service Terms
Bedrock configurations
Model access
Bedrock Studio Preview
Settings
Select the data types to include with logs - Optional
Select the data types for requests and responses for all model invocations.
Text
Image
Embedding
Select the logging destinations
S3 only
Cloudwatch Logs only
Both S3 and Cloudwatch Logs
CloudWatch Logs configurations
Log group name
Invocation logs and model input & output data of up to 100 KB each will be published in this log group
bedrock-invocation=
Choose a method to authorize Bedrock
Use an existing service role
Create and use a new role
Service role
Choose role
Model input or output data larger than 100kb or in binary format will not be published to CloudWatch Logs. If S3 configuration for large data delivery is not provided, that model data will not be published.
S3 location for large data delivery - optional
s3://bucket/prefix
View
Browse S3


Amazon Bedrock
https://us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/settings
aws
Services
Search
[Option+S]
N. Virginia
aws-courses
Prompt management Preview
Knowledge bases
Agents
Prompt flows Preview
Safeguards
Guardrails
Watermark detection
Inference
Provisioned Throughput
Batch inference New
Cross-region inference New
Assessment
Model Evaluation
User guide
Bedrock Service Terms
Bedrock configurations
Model access
Bedrock Studio Preview
Settings
Text
Image
Embedding
Select the logging destinations
S3 only
Cloudwatch Logs only
Both S3 and Cloudwatch Logs
CloudWatch Logs configurations
Log group name
Invocation logs and model input & output data of up to 100 KB each will be published in this log group
bedrock-invocation-logging
Choose a method to authorize Bedrock
Use an existing service role
Create and use a new role
Service role name
BedrockInvocationLoggingRole
Maximum 64 characters. Use alphanumeric and '+=,.@-_' characters.
ViewPermission
Model input or output data larger than 100kb or in binary format will not be published to CloudWatch Logs. If S3 configuration for large data delivery is not provided, that model data will not be published.
S3 location for large data delivery - optional


Amazon Bedrock
https://us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/settings
aws
Services
Search
[Option+S]
N. Virginia
aws-courses
Prompt management Preview
Knowledge bases
Agents
Prompt flows Preview
Safeguards
Guardrails
Watermark detection
Inference
Provisioned Throughput
Batch inference New
Cross-region inference New
Assessment
Model Evaluation
User guide
Bedrock Service Terms
Bedrock configurations
Model access
Bedrock Studio Preview
Settings
S3 only
Cloudwatch Logs only
Both S3 and Cloudwatch Logs
CloudWatch Logs configurations
Log group name
Invocation logs and model input & output data of up to 100 KB each will be published in this log group
bedrock-invocation-logging
Choose a method to authorize Bedrock
Use an existing service role
Create and use a new role
Service role name
BedrockInvocationLoggingRole
Maximum 64 characters. Use alphanumeric and '+=,.@-_' characters.
ViewPermission
Model input or output data larger than 100kb or in binary format will not be published to CloudWatch Logs. If S3 configuration for large data delivery is not provided, that model data will not be published.
S3 location for large data delivery - optional
s3://bucket/prefix
View
Browse S3
Cancel
Save settings


Amazon Bedrock
CloudWatch | us-east-1
https://us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/chat-playground
aws
Services
Search
[Option+S]
N. Virginia
aws-courses
Amazon Bedrock
Getting started
Overview
Examples
Providers
Foundation models
Base models
Custom models
Imported models Preview
Playgrounds
Chat
Text
Image
Builder tools
Prompt management Preview
Knowledge bases
Agents
Prompt flows Preview
Safeguards
Guardrails
Watermark detection
Amazon Bedrock > Chat playground
Chat playground Info
Load examples
Compare mode
Titan Text G1 - Express v1 | On-demand throughput
Change
Action items for Miguel:
- Document other concerns
Action items for Brant:
- Work with James from another team to unblock the sign up workflow
Action items for Namita:
- Work on the landing page to make the product more discoverable
Write a prompt... (Shift + ENTER to start a new line, and ENTER to generate a response)
Run
Choose files
The supported file types are .csv, .doc, .docx, .html, .md, .pdf, .txt, .xls, .xlsx. Maximum 5 files, 4.5 MB each.
Configurations
Reset
Randomness and diversity Info
Temperature
0
Top P
1
Length Info
Response length
4096
Stop sequences
Add
Model metrics
Define metric criteria
To evaluate models for task specific metrics with custom dataset visit Model evaluation
Metrics
Titan Text G1 - Express
Overall summary
Define metric criteria


Amazon Bedrock
CloudWatch | us-east-1
https://us-east-1.console.aws.amazon.com/cloudwatch/home?region=us-east-1#logsV2:log-groups/log-group/bedrock-invocation-logging/log-events/aws$252Fbedrock$252Fmodelinvocations
aws
Services
Search
[Option+S]
N. Virginia
aws-courses
CloudWatch
Favorites and recents
Dashboards
Alarms
Logs
Log groups
Log Anomalies
Live Tail
Logs Insights
Contributor Insights
Metrics
X-Ray traces
Events
Application Signals
Network monitoring
Insights
Settings
Getting Started
No older events at this moment. Retry
2024-09-11T12:04:12.331Z
Permissions are correctly set for Amazon Bedrock logs.
Permissions are correctly set for Amazon Bedrock logs.
2024-09-11T12:04:38.000Z
{"schemaType":"ModelInvocationLog","schemaVersion":"1.0","timestamp":"2024-09-11T12:04:38Z","accountI…
{
"schemaType": "ModelInvocationLog",
"schemaVersion": "1.0",
"timestamp": "2024-09-11T12:04:38Z",
"accountId": "381491951425",
"identity": {
"arn": "arn:aws:iam::381491951425:root"
},
"region": "us-east-1",
"requestId": "5a100f04-72de-48e1-9dd7-393e2d096e28",
"operation": "ConverseStream",
"modelId": "amazon.titan-text-express-v1",
"input": {
"inputContentType": "application/json",
"inputBodyJson": {
"messages": [
{
"role": "user",
"content": [
{
"text": "Meeting transcript: \nMiguel: Hi Brant, I want to discuss the workstream for our new product launch \nBrant: Sure Miguel, is there anything in particular you want to discuss? \nMiguel: Yes, I want to talk about how users enter into the product. \nBrant: Ok, in that case let me add in Namita. \nNamita: Hey everyone \nBrant: Hi Namita, Miguel wants to discuss how users enter into the product. \nMiguel: its too complicated and we should remove friction. for example, why do I need to fill out additional forms? I also find it difficult to find where to access the product when I first land on the landing page. \nBrant: I would also add that I think there are too many steps. \nNamita: Ok, I can work on the landing page to make the product more discoverable but brant can you work on the additonal forms? \nBrant: Yes but I would need to work with James from another team as he needs to unblock the sign up workflow. Miguel can you document any other concerns so that I can discuss with James only once? \nMiguel: Sure. \nFrom the meeting transcript above, Create a list of action items for each person. \n"
}
]
Back to top