ACOT201

SalesGenie

Leveraging generative AI for assisting life insurance sales agents on the field

Sanjay Thawakar

Senior Vice President & Head of AI Works & BPMA
Max Life Insurance

Senthil Jeyachandran

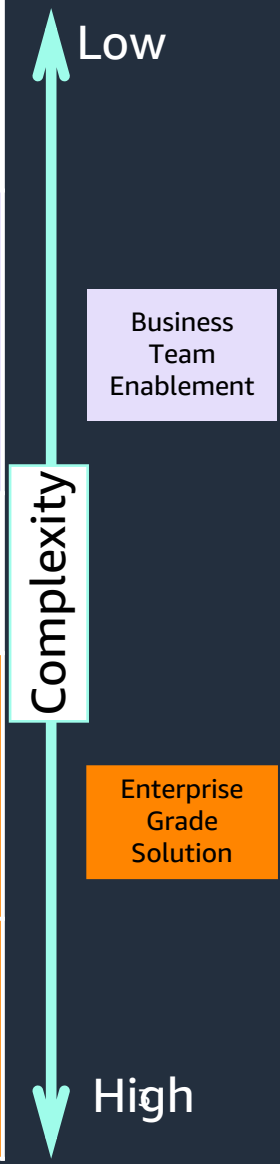
AI & ML Specialist, Solutions Architect
AWS India

Agenda

- Generative AI landscape within life insurance industry
- Maxlife SalesGenie – Solving sales distribution challenges
- Solution development, key elements and learnings
- Solution architecture
- Achievements and way-forward
- Generative AI overview
- Amazon Bedrock
- Customization, Knowledge base, Agents for Bedrock, Guardrails
- Demo

Generative AI landscape within Insurance

Function/ Category*	Sales	Marketing	HR	Operations	Customer Service	Digital, Ecom, AIW	Finance, Legal, & Others
Content Translation & Classification	• Translating sales script & material to language of choice	• Vernacular Campaign design	• Resume Screening		• Script translation	• Code Refactoring	
Content Creation	• Generating Sales Script	• KFD video creation • Seller activation and buzz • Press drafts • Celebrity image generation • CEO Webcast script generation	• Interview Questions • Creating Job Descriptions	• Renewal KFD • Renewal Pitch Generation	• Co-pilots for agents to generate responses on customer queries	• Code Generator • Creating scripts to test code • SEO Meta Description Creation	• Boiler plate contract generator
Content Clarification	• Calls transcripts & summarization	• Analysis for Annual Reports	• Policy Summary	• Condense Lengthy Document into Brief Summaries	• Calls transcripts & summarization	• Annotations to Code • Bug Detection	
Content Analysis, Design & Summarization	• Product Information	• Personalised Product EDM	• Interview feedback • Candidate Ranking • Employee Engagement Survey Analysis	• Search Internal Documents, Optimize, & Recommendations	• Sentiment Analysis & response generator	• Code Optimization	• Contract Summarization • Financial Report Summarization
Act basis Instruction	• Agent Training Bots • Virtual Sales Agent • Virtual Supervisor	• Automate ad targeting	• Virtual Recruiter • Interview Scheduling, processing & summarizing	• UW Twin • Renewal Tele calling assistant • Perform Data Entry	• Email & Chat Bot • Product feature summarization for CS Support • Virtual CS Agent	• Coding co-pilot	• Investment Guide



SalesGenie: Solving sales distribution challenges




Low trainer to FLS ratio	Limited Supervisors	Retention of knowledge	Staying updated with current affairs
Training new force		Disparate skill of sellers	
1 Lac field sales force. High %ge new force every year			

Sales Pitch Assistance

In-depth MLI product information and comparison

Competitor product differentiation

Objection Handling

Awareness of Financial instrument and Current Affairs

Evolution of solution development

1. Business case-pack → 2. Prompt engineering → 3. Retrieval Augmented generation

Inception

1. Business case-pack	2. Prompt engineering	3. Retrieval Augmented generation
Collating questions across business themes + base documents + Videos	Assess the LLM's ability to generate content, summarize information, build on conversations and information	Assess the LLM's ability to generate content, summarize information, build on conversations and information

CUG Test release – with feedback

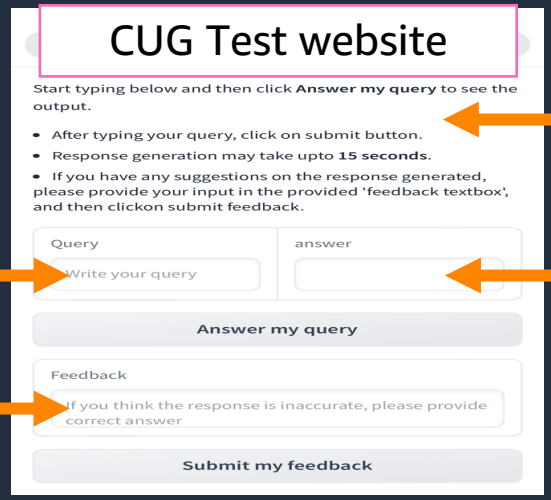
Evolution

1. Business case-pack	2. Prompt engineering	3. Retrieval Augmented generation
Extracting information from images, Speech to text	Build guard-rails Provide sense of timeline Build self QC and verification	Simplifying complex tabular data with merged cells, columns. Building Glossary, common lingo

LLM Evaluation

LLM	Parameters	Access	Information Extraction	Conversational AI	Summarization	Content Generation
Claude 2	135B	API	✓	✓	✓	✓
Falcon-40B	40B	Open Source	⚠	⚠	✓	⚠
Llama	70B	API	⚠	⚠	✓	⚠
GPT-3.5	175B	API	⚠	✓	✓	✓
Llama2	70B	Open Source	⚠	⚠	✓	✓

Key elements and learnings

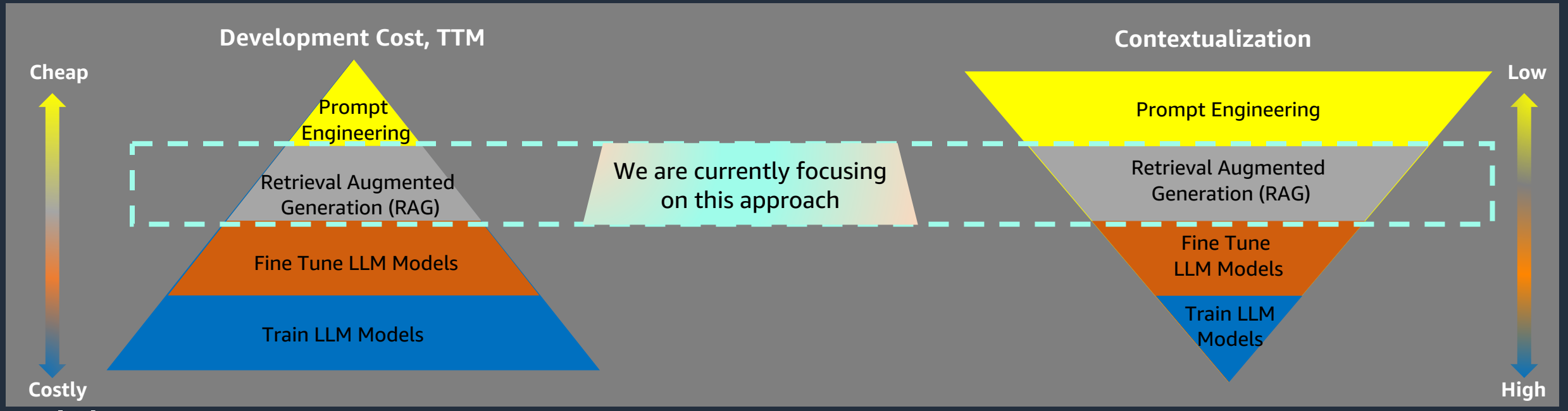


Instructions for new users – How to use and What to expect

Placeholder to ask query

Response generated from intelligence engine

Gather feedback on generated response



Considerations for developing LLM apps

Fraud - jailbreak

- ✓ Strict persona
- ✓ Verify content match to source data
- ✓ Condition user query

Bias - Hallucination

- ✓ Lowest temperature for generation
- ✓ Guard-rails for fact check
- ✓ Verify calculations / inference

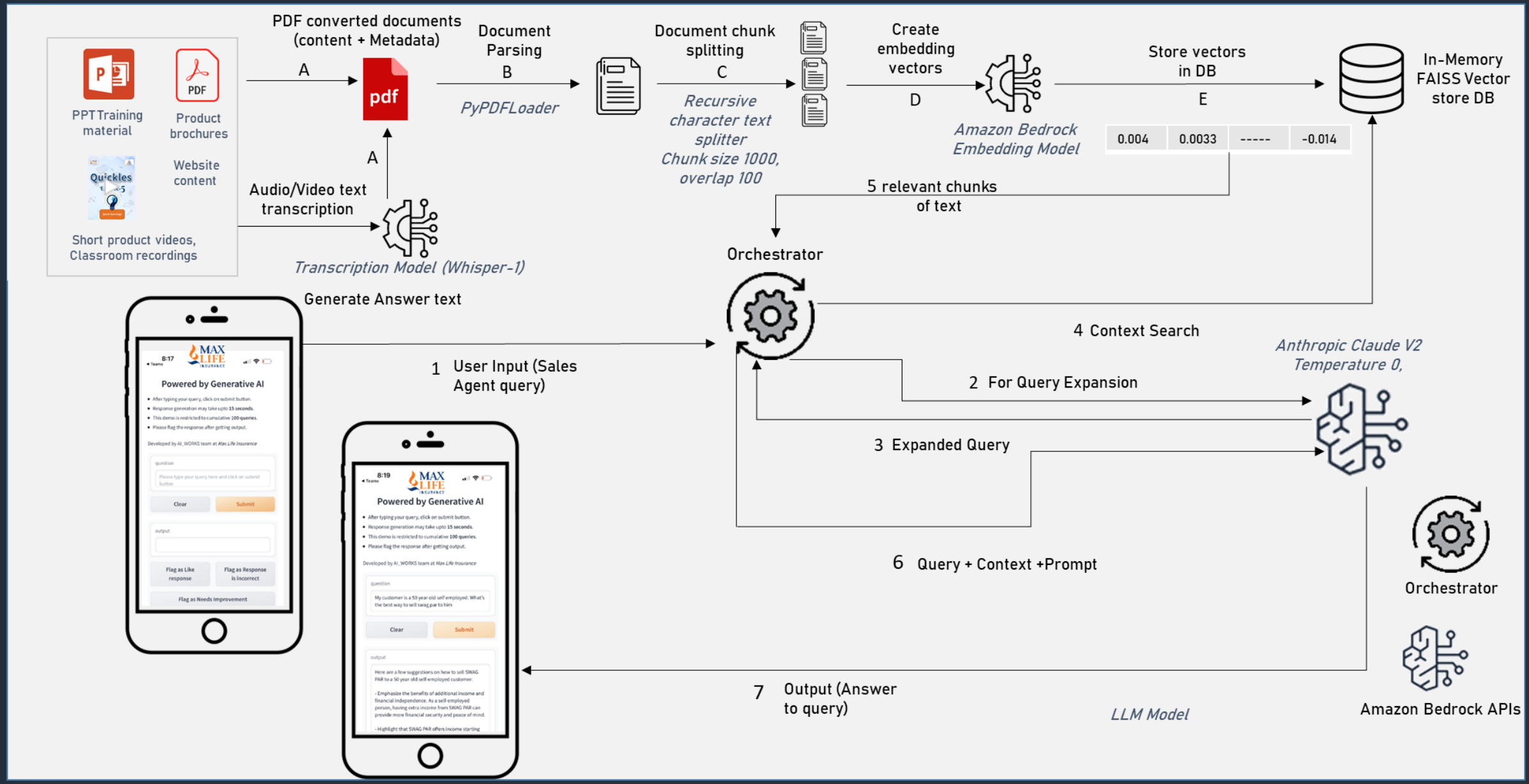
Sensitive data leaks

- ✓ Using private models
- ✓ Using private embedding model

Stability of response

- ✓ Using private instances
- ✓ Model version tracking

Solution Architecture

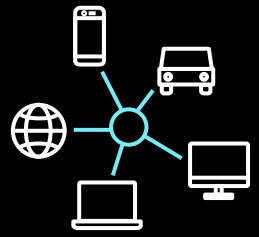


Test results and way forward

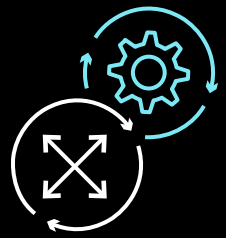
- Amazon bedrock titan text embedding + Anthropic's Claude-V2 outperformed GPT3.5-turbo + Ada text embedding
- Simplifying complex data tables key to improved response generation
- Maintaining knowledgebase and updating knowledge retrieval systems to fetch latest data is key challenge to future sustenance.
- Ability to handle new products, lingo and dialects remain a key challenge when they are presented to LLMs. Ex-SWAG product

Question Category	# Questions	Claude V2		GPT 3.5 Turbo	
		Correct Answers as expected	% Accuracy	Correct Answers as Expected	% Accuracy
Sales and Pitching	216	210	**97%**	210	**97%**
Products Related (Recommendations and variant information)	150	108	**72%**	91	**60%**
Competition comparisons	126	114	**90%**	54	**43%**
Financial knowledge	270	264	**98%**	264	**98%**
Sales pitching, Conversation Starters and Objection Handling	174	168	**97%**	168	**97%**
Total	**936**	**864**	**92%**	**787**	**84%**

What generative AI customers are asking for



Which model should I use?



How can I move quickly?

How can I keep my data secure and private?

Amazon Bedrock

The easiest way to build and scale generative AI applications with foundation models

- Choice of leading FMs via single API
- Model customization
- Retrieval Augmented Generation (RAG)
- Agents that execute multistep tasks
- Security, privacy, and safety

Amazon Bedrock

Broad choice of models

AI21 labs	amazon	ANTHROPIC	cohere	Meta	stability.ai
Jurassic-2 Ultra	Titan Text Embeddings	Claude 2	Command + Embed	Llama 2	Stable Diffusion XL1.0
Jurassic-2 Mid	Titan Multimodal Embeddings	Claude 2.1	Cohere Command Light	Llama 2 13B	
	Titan Text Lite	Claude Instant	Cohere Embed English	Llama 2 70B	
	Titan Text Express		Cohere Embed Multilingual		
	Titan Image Generator				

Why customize?

Adapt to domain-specific language

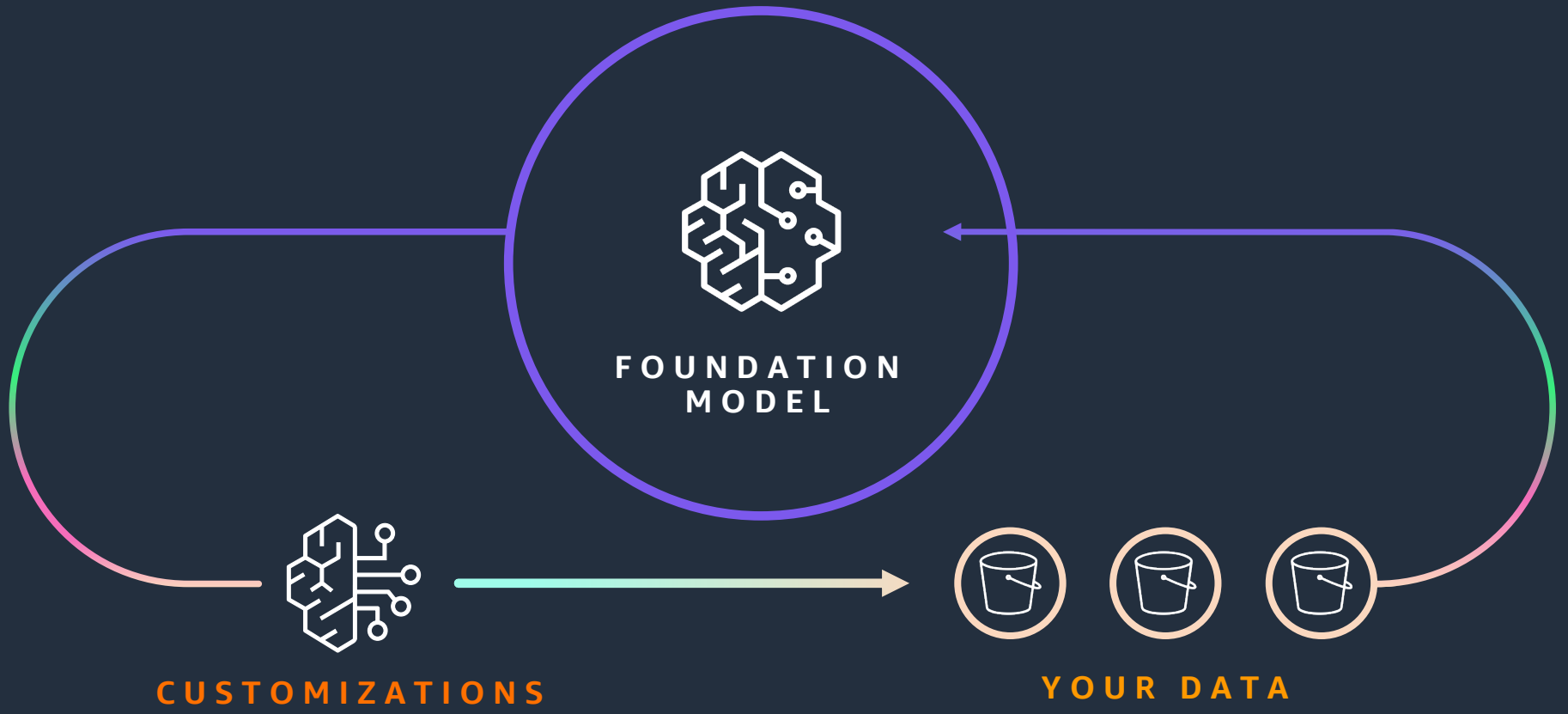
E.g., Healthcare – Understand medical terminology and provide accurate responses related to patient's health

Enhance performance for specific tasks

E.g., Finance – Teach financial & accounting terms to provide good analysis for earnings reports

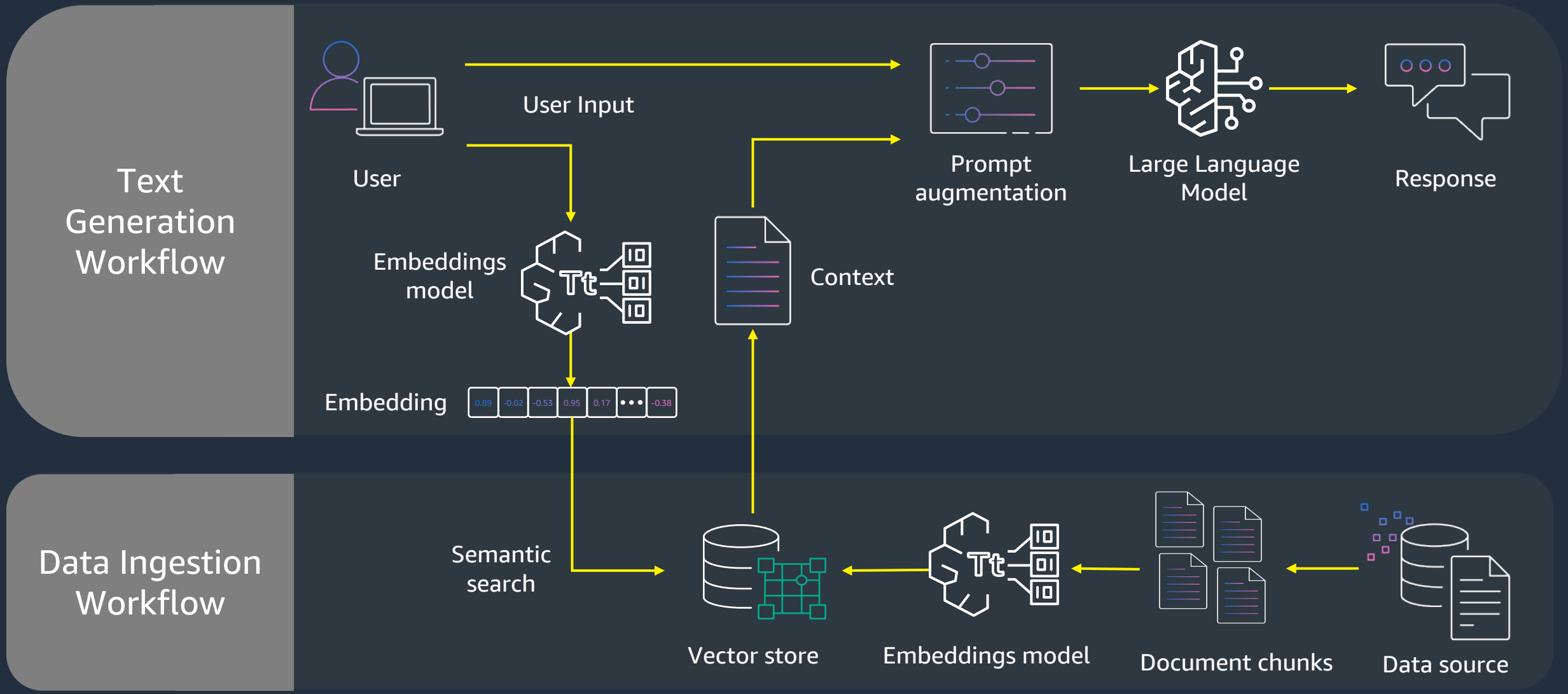
Improve context-awareness in responses

E.g., Customer Service – Improve ability to understand and respond to customer's inquires and complaints



Fine tuning | Retrieval Augmented Generation (RAG) | Continued Pre-training

RAG in Action

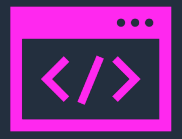


However, when it comes to implementing RAG, there are challenges…

Managing multiple data sources

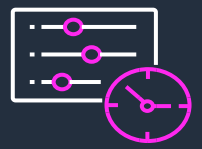
Creating vector embeddings for large volumes of data

Incremental updates to vector store



Coding effort

Scaling retrieval mechanism



Orchestration

Knowledge bases for Amazon Bedrock

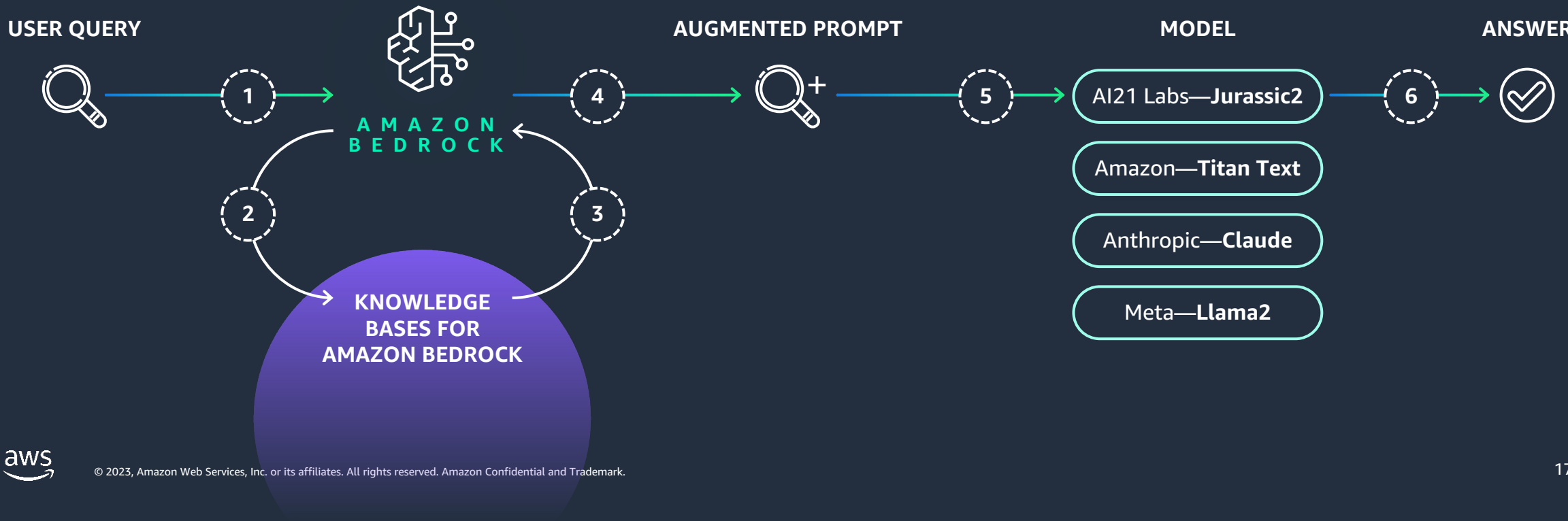
NATIVE SUPPORT FOR RETRIEVAL AUGMENTED GENERATION (RAG)

Securely connect FMs to data sources for RAG to deliver more relevant responses

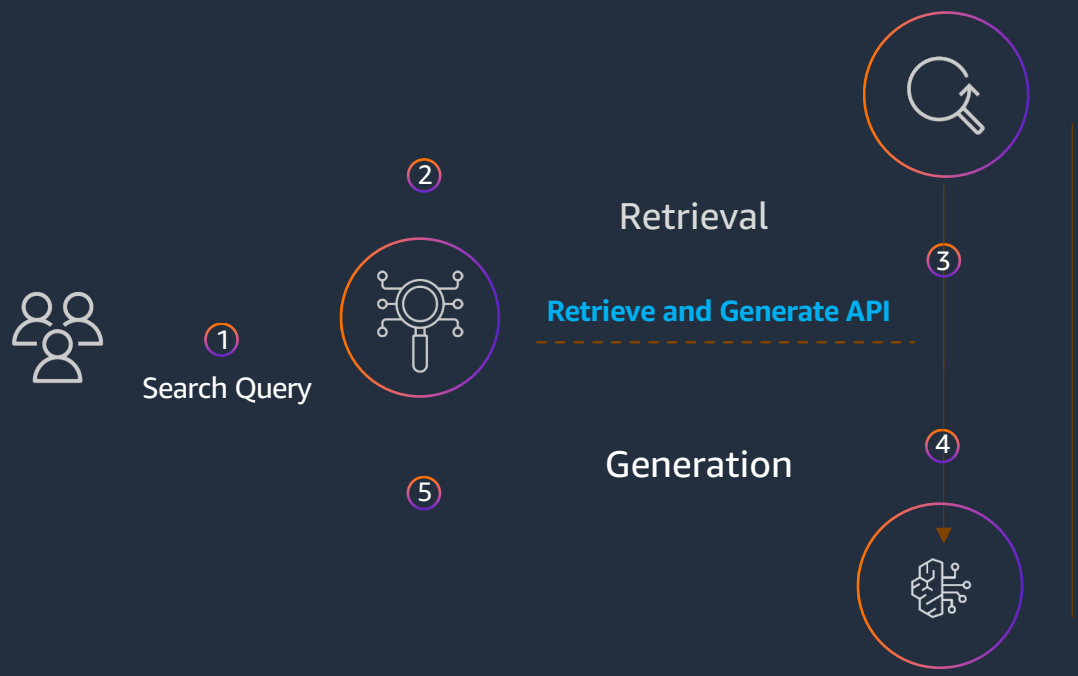
Fully managed RAG workflow including ingestion, retrieval, and augmentation

Built-in session context management for multi-turn conversations

Automatic citations with retrievals to improve transparency



Retrieve and Generate API



- Pinecone
- redis
- OpenSearch
- Aurora — *Coming soon*
- MongoDB — *Coming soon*

Single API	One API for RAG
Comprehensive	Multiple knowledge-bases
Session	Built-in session management

Retrieve and generate API will enable a simplified RAG solution

Vector databases for **Amazon Bedrock**

Vector Engine For Amazon OpenSearch Serverless

Redis Enterprise Cloud

Pinecone

COMING SOON

Amazon Aurora

COMING SOON

MongoDB

Agents for Amazon Bedrock

ENABLE GENERATIVE AI APPLICATIONS TO EXECUTE MULTISTEP TASKS USING COMPANY SYSTEMS AND DATA SOURCES

1. SELECT YOUR FOUNDATION MODEL
2. PROVIDE BASIC INSTRUCTIONS
3. SELECT RELEVANT DATA SOURCES
4. SPECIFY AVAILABLE ACTIONS

- Breaks down and orchestrates tasks
- Securely accesses and retrieves company data for RAG
- Takes action by invoking API calls on your behalf
- Chain-of-thought trace and ability to modify agent prompts

Guardrails for Amazon Bedrock

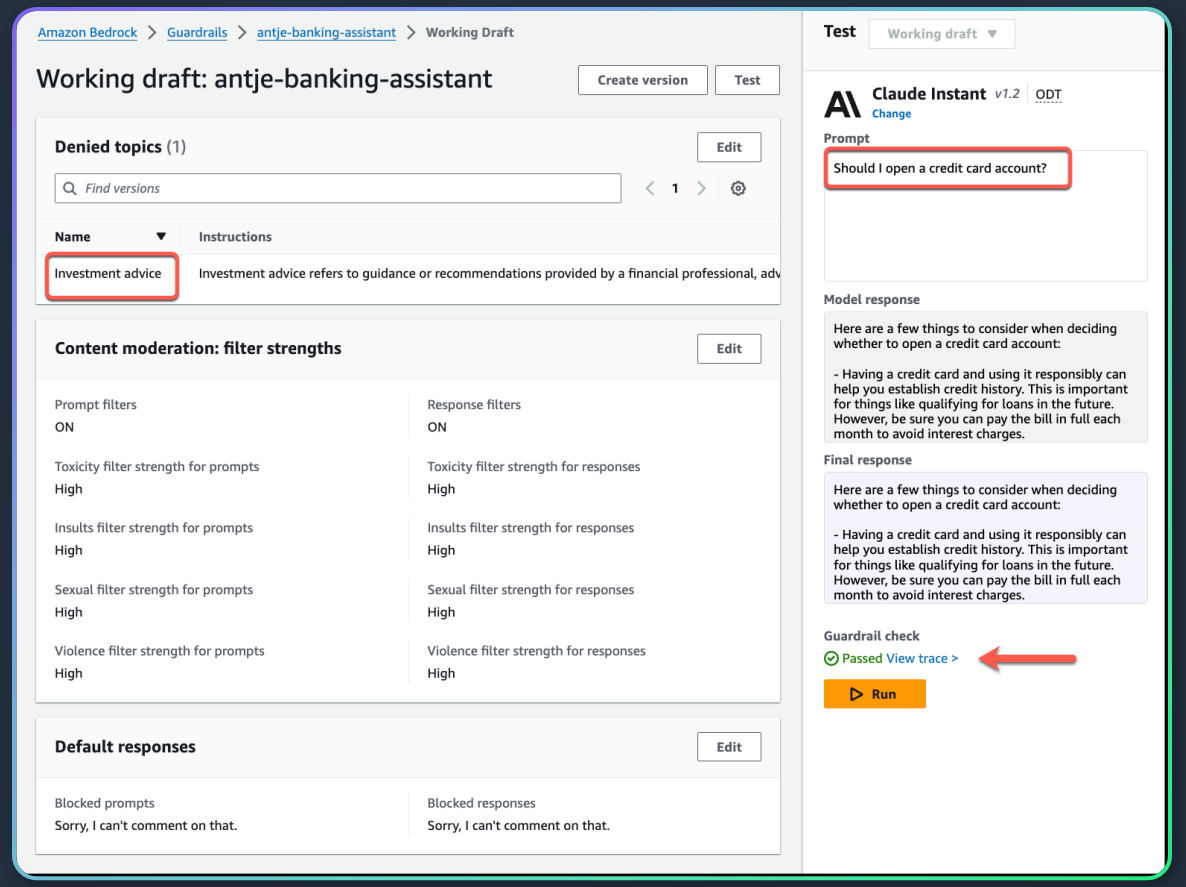
IMPLEMENT SAFEGUARDS TAILORED TO YOUR APPLICATION REQUIREMENTS AND RESPONSIBLE AI POLICIES

Preview

Apply guardrails consistently across FMs including fine-tuned models and agents

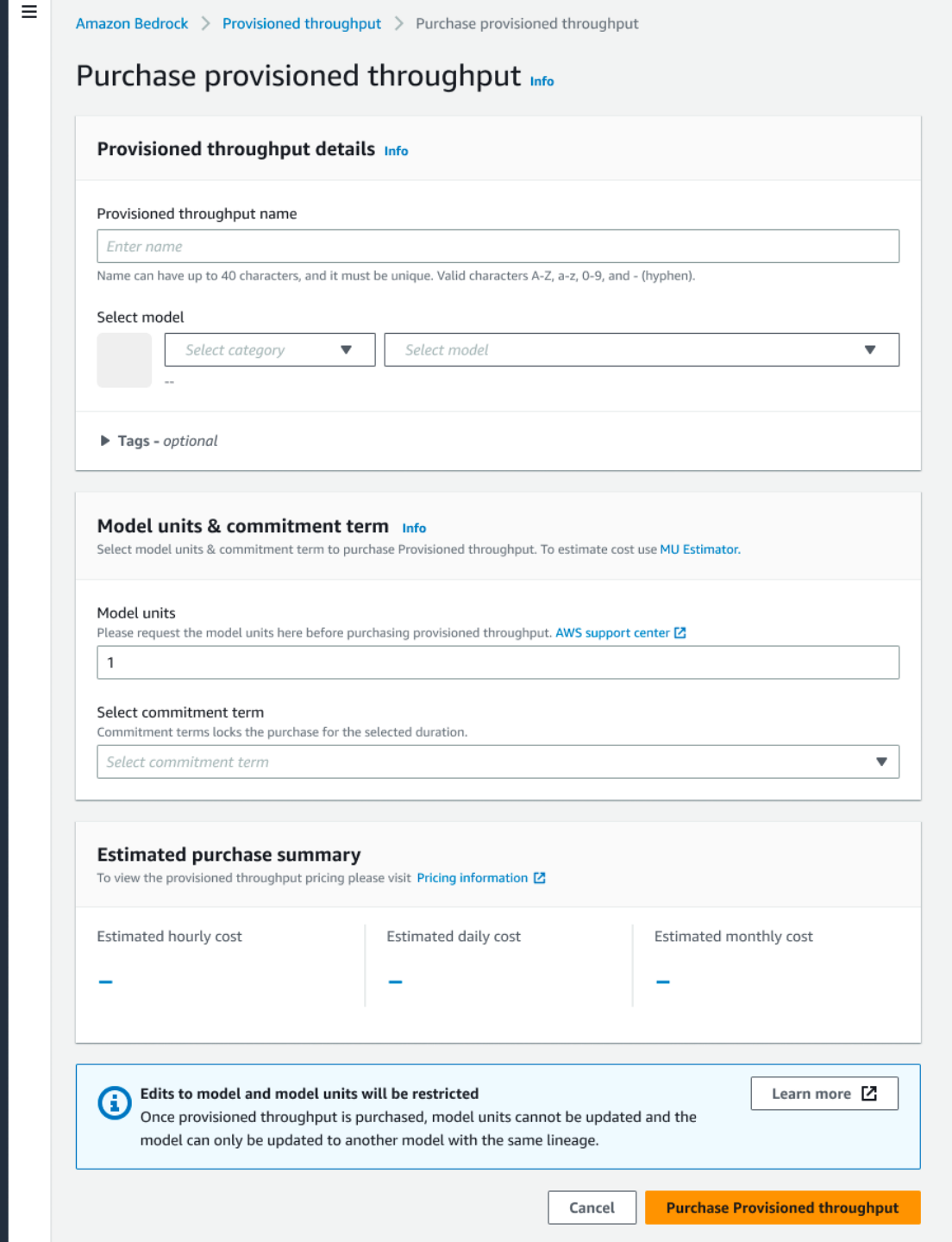
Configure filtering of harmful content and topics to avoid based on your responsible AI policies

Redact personally identifiable information (coming soon)



Provisioned throughput

- Reserve throughput (input/output tokens per minute)
- Ensure consistent user experience during traffic spikes
- Purchase with commitment term of one month or six months
- Pay hourly rate, discounted for extended commitment



Demo

Start typing below and then click **Answer my query** to see the output.

SalesGenie

- After typing your query, click on submit button.
- Response generation may take upto **15 seconds**.
- This demo is restricted to cumulative **100 queries**.
- If you have any suggestions on the response generated, please provide your input in the provided 'feedback textbox', and then clickon submit feedback.

Query

Write your query

answer

Answer my query

Feedback

If you think the response is inaccurate, please provide correct answer

Submit my feedback

Thank you!

Sanjay Thawakar

Senior Vice President & Head – AI Works & BPMA

Max Life Insurance

Senthil Jeyachandran

AI & ML Specialist, Solutions Architect

AWS India