



DEV DAY

SAN FRANCISCO

Patterns to Automatically Scale your Generative AI Solutions with Serverless

Dhiraj Mahapatro

Principal Specialist SA, Amazon Bedrock

AWS

Uma Ramadoss

Principal Specialist SA, Serverless

AWS



Agenda

API based patterns

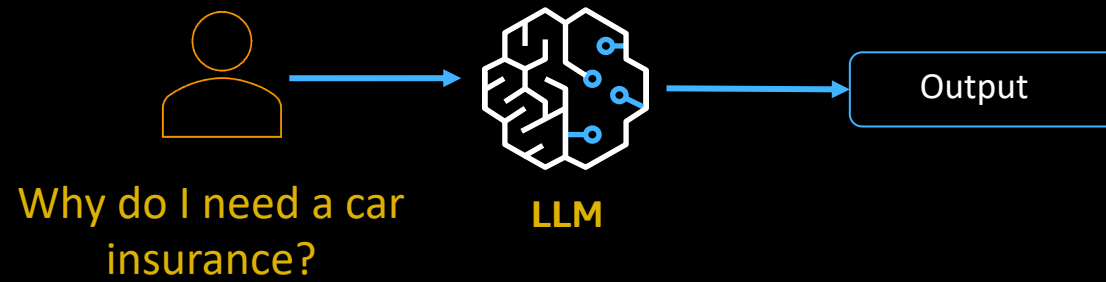
Data processing patterns

Key takeaways

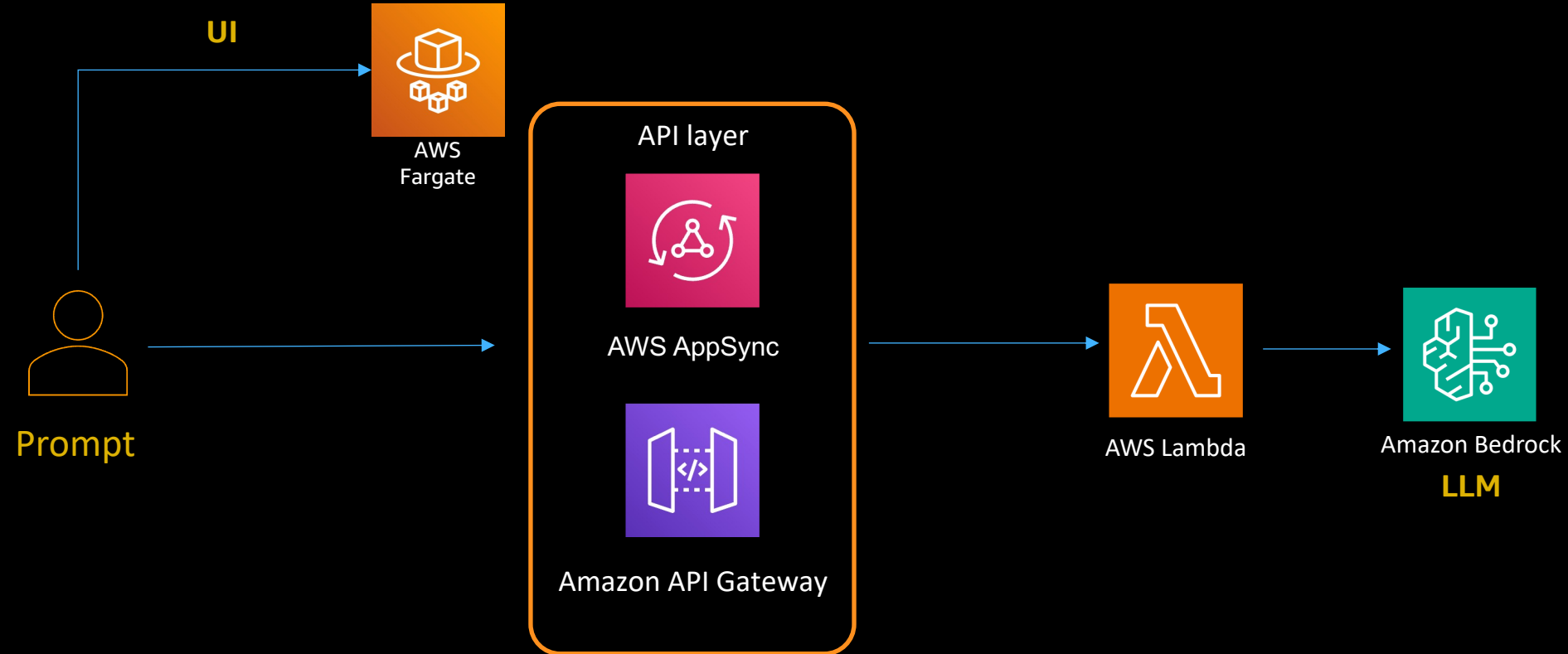
API based patterns



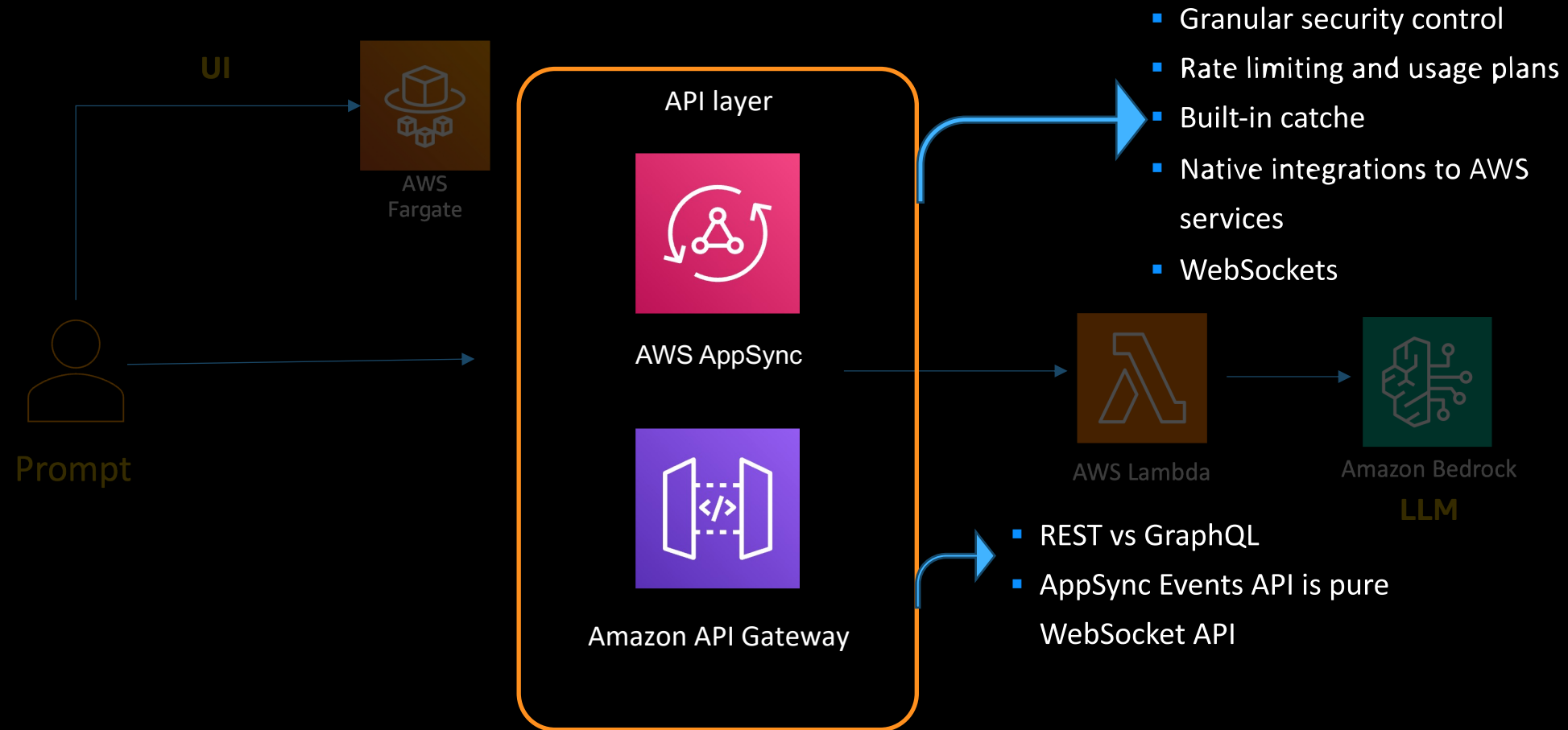
Chatbot



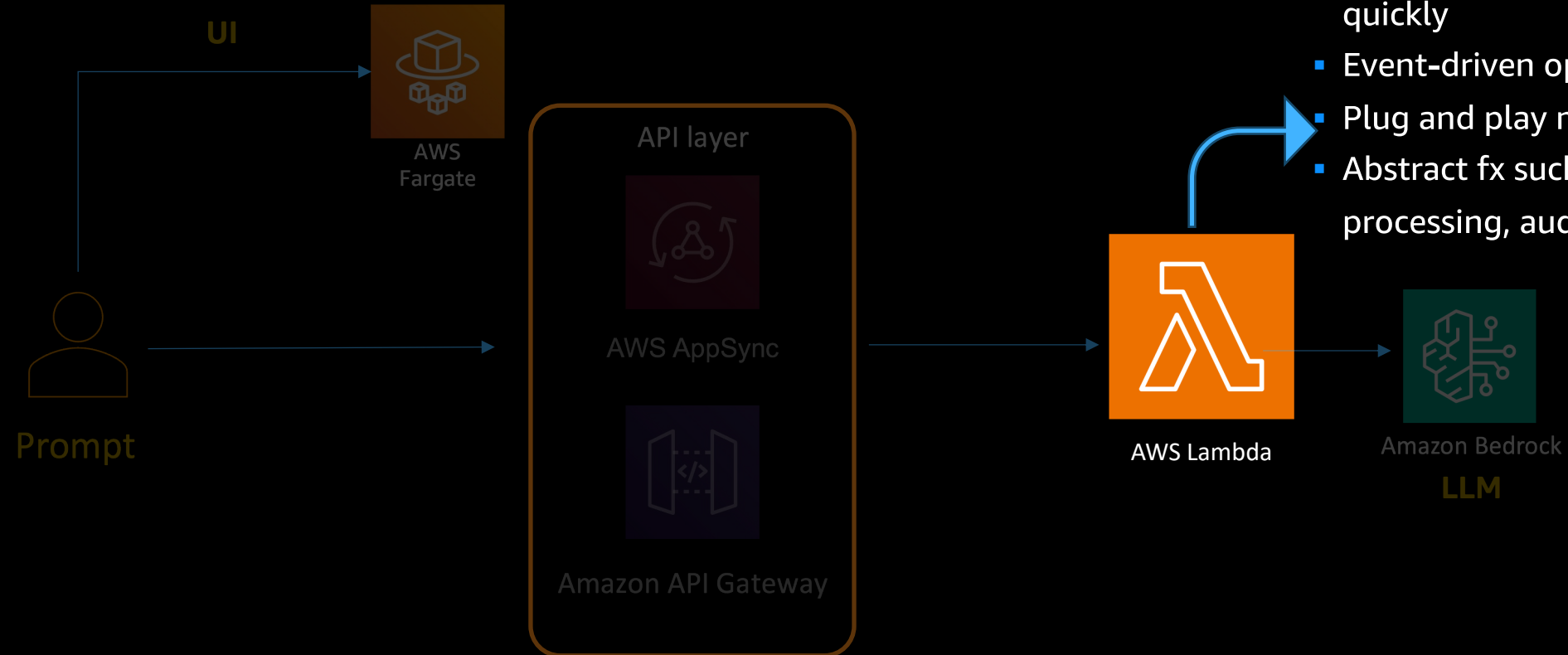
Sync chatbot API implementation



Sync chatbot API implementation

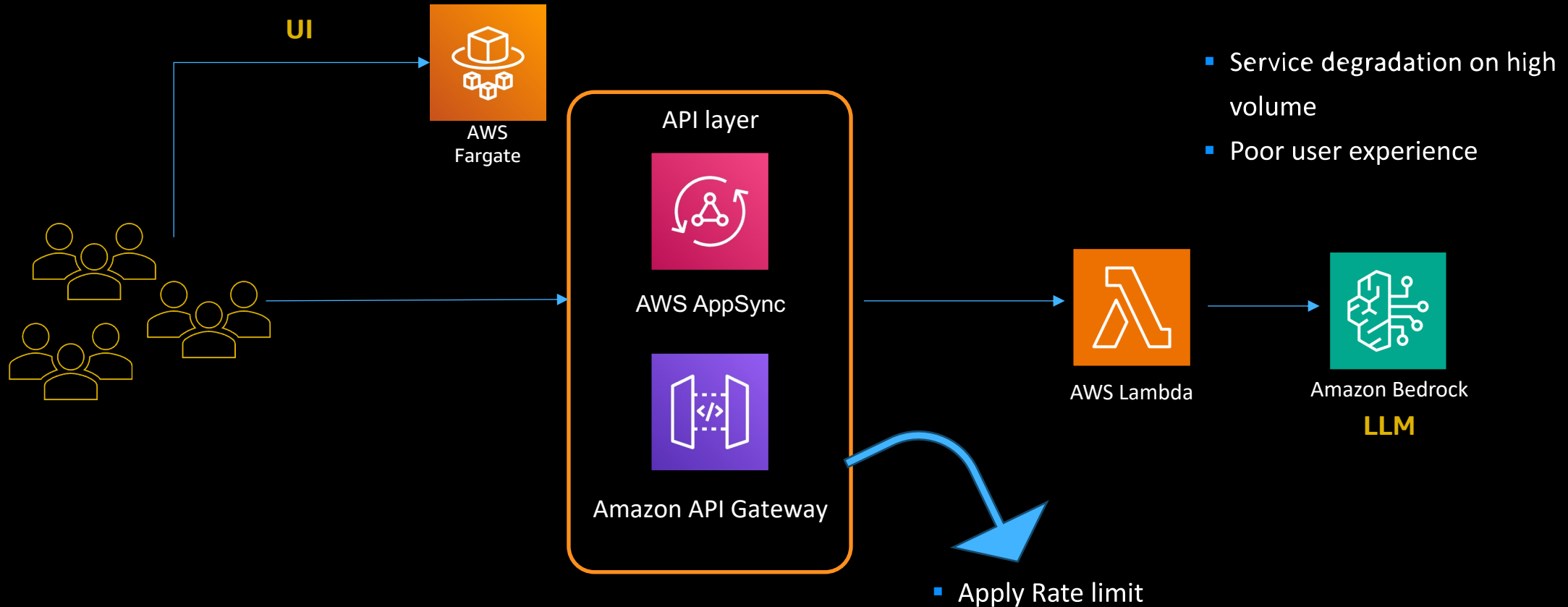


Sync chatbot API implementation

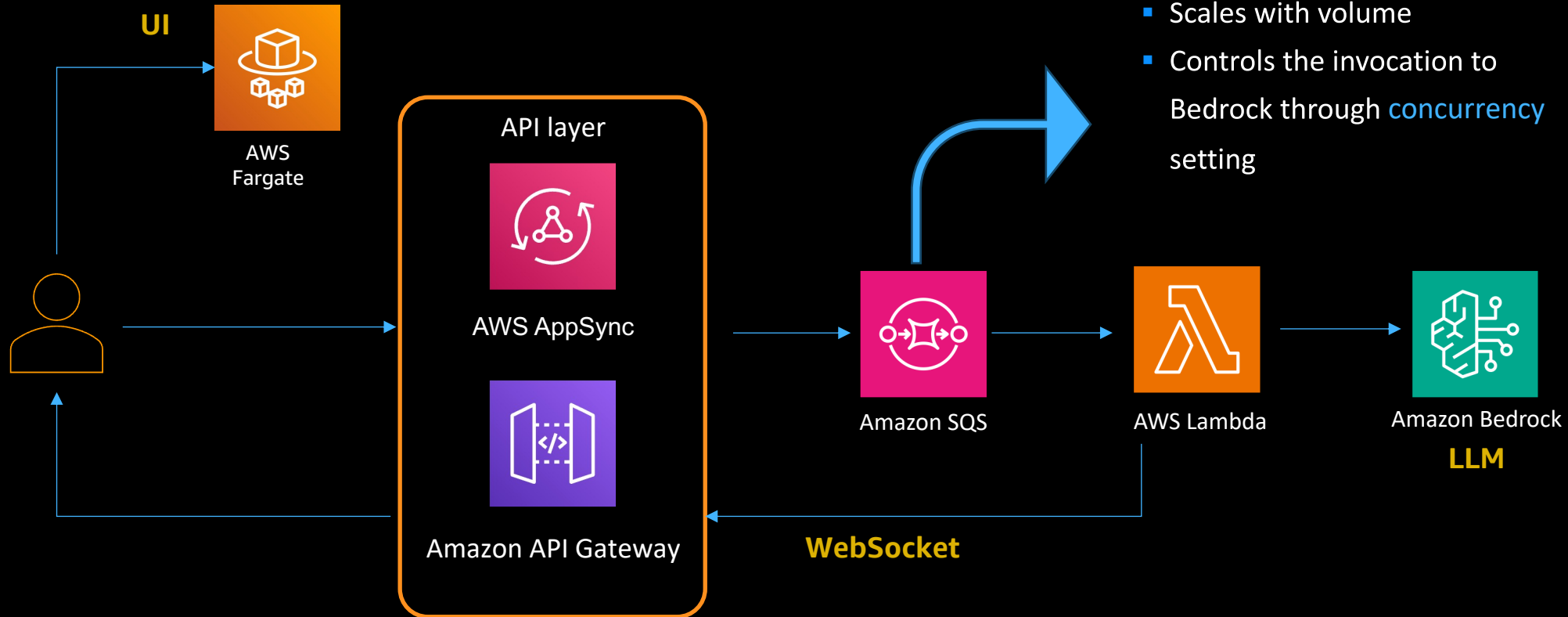


- Scale to zero and ramp up quickly
- Event-driven operating model
- Plug and play models
- Abstract fx such as pre/post processing, auditing

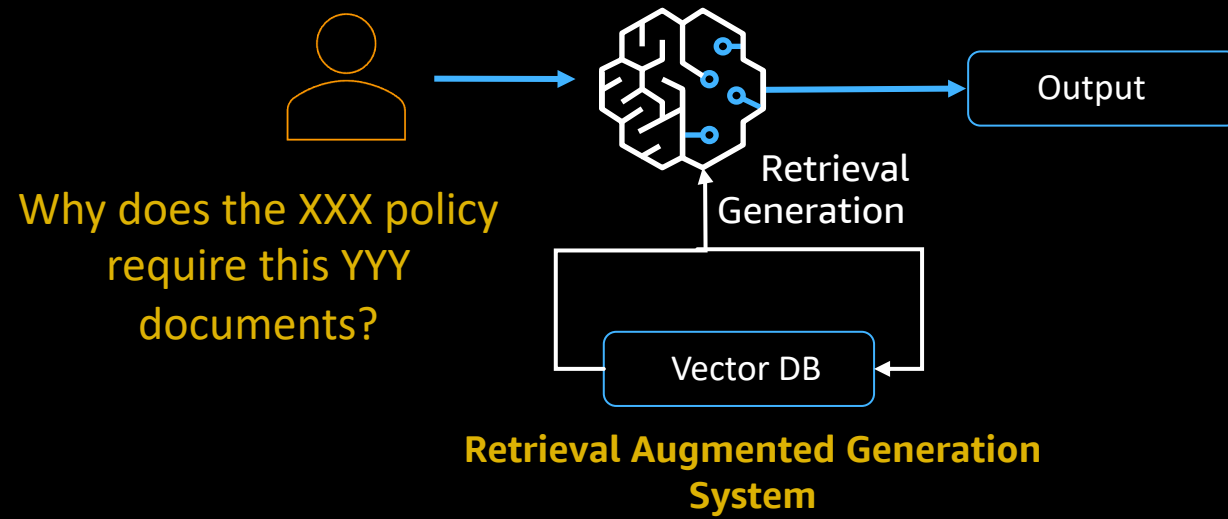
Challenges with sync chatbot implementation



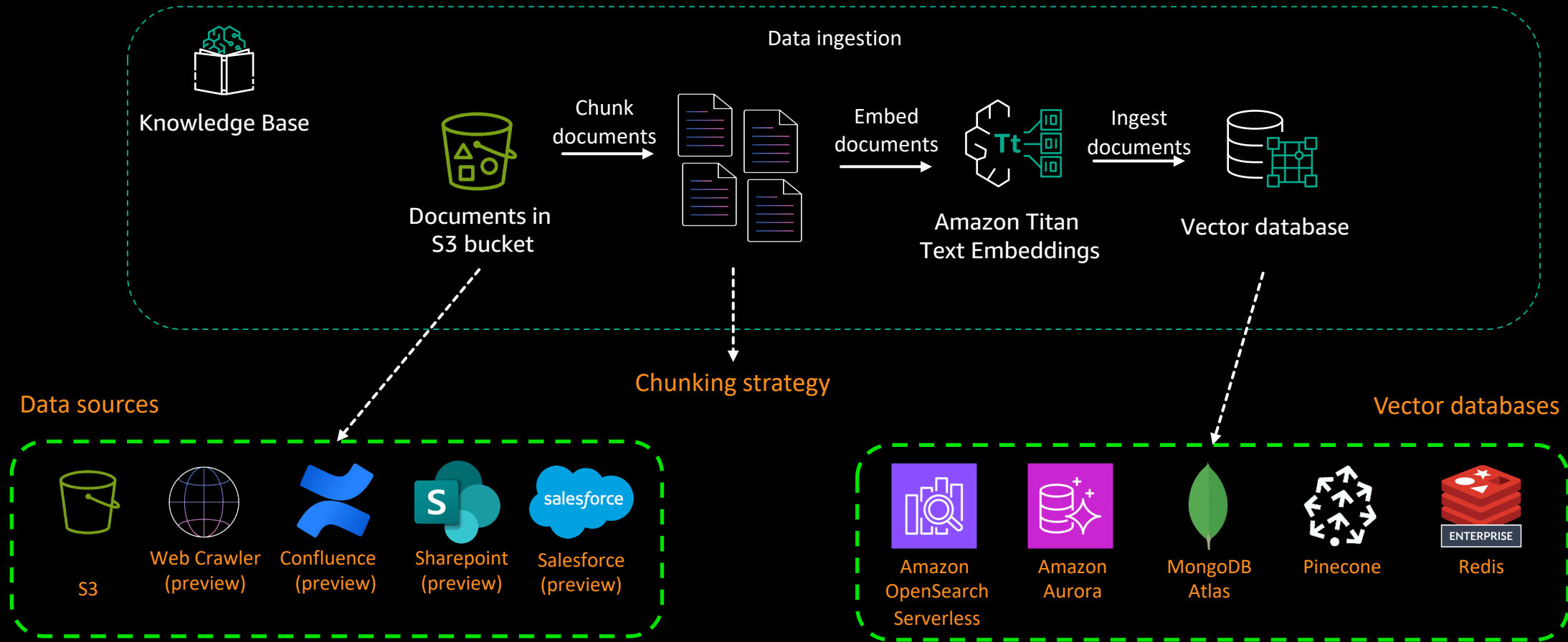
Async chatbot API implementation



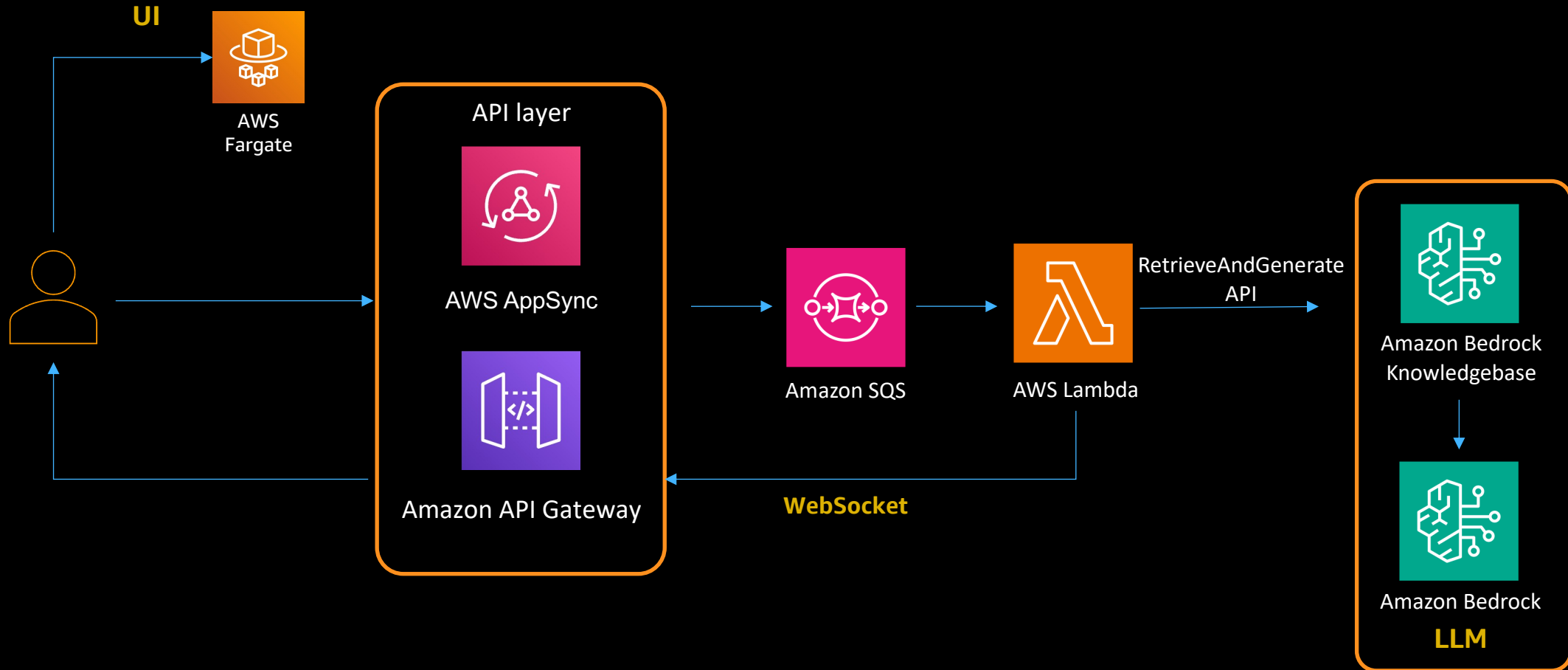
Domain specific chatbot



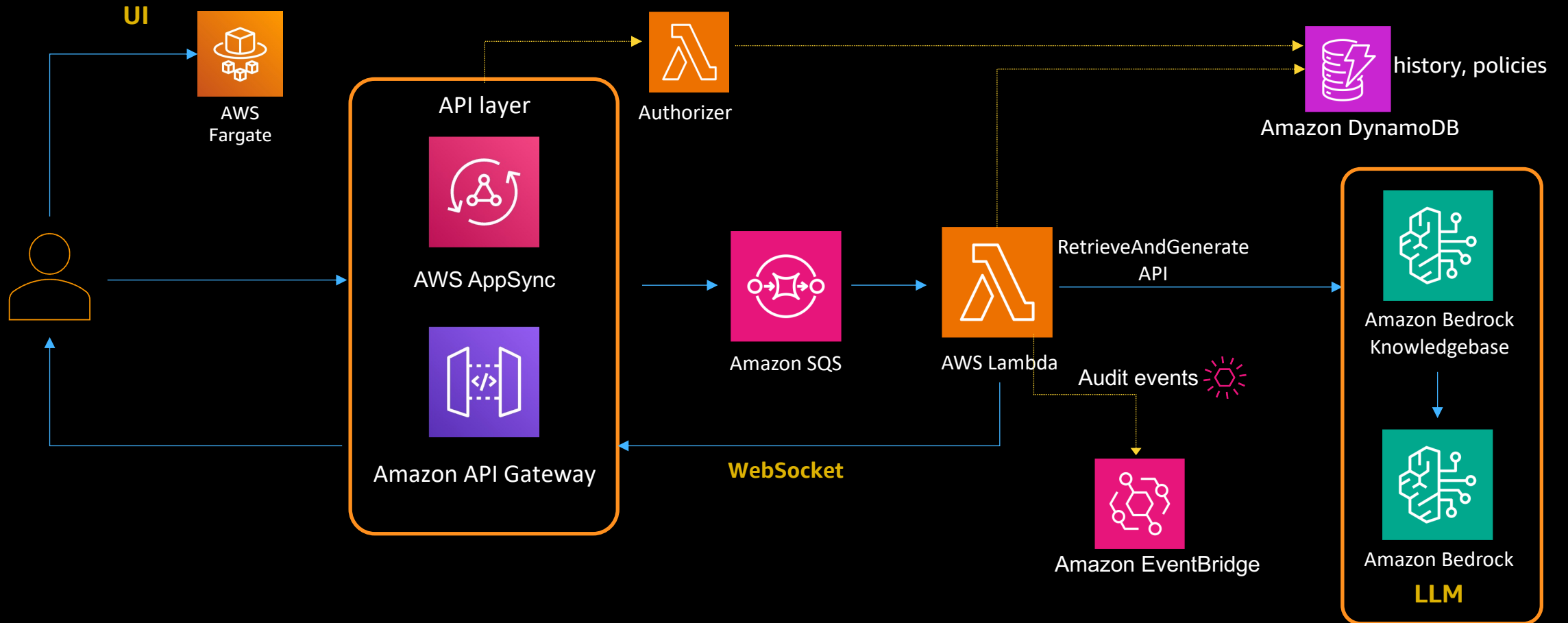
Amazon Bedrock knowledge bases



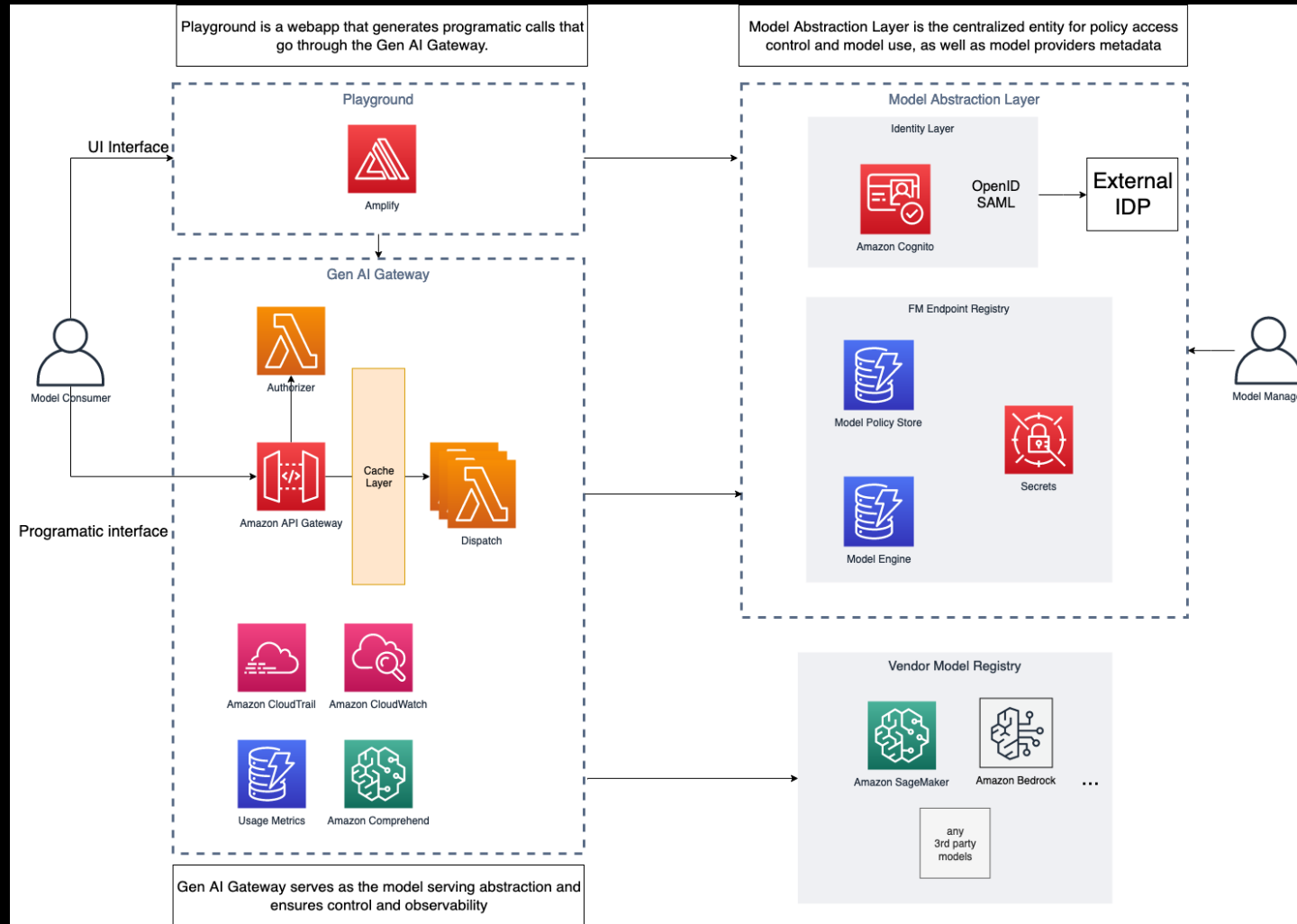
API Augmented with domain specific data



API Augmented with domain specific data



Extend your architecture - Generative AI Gateway

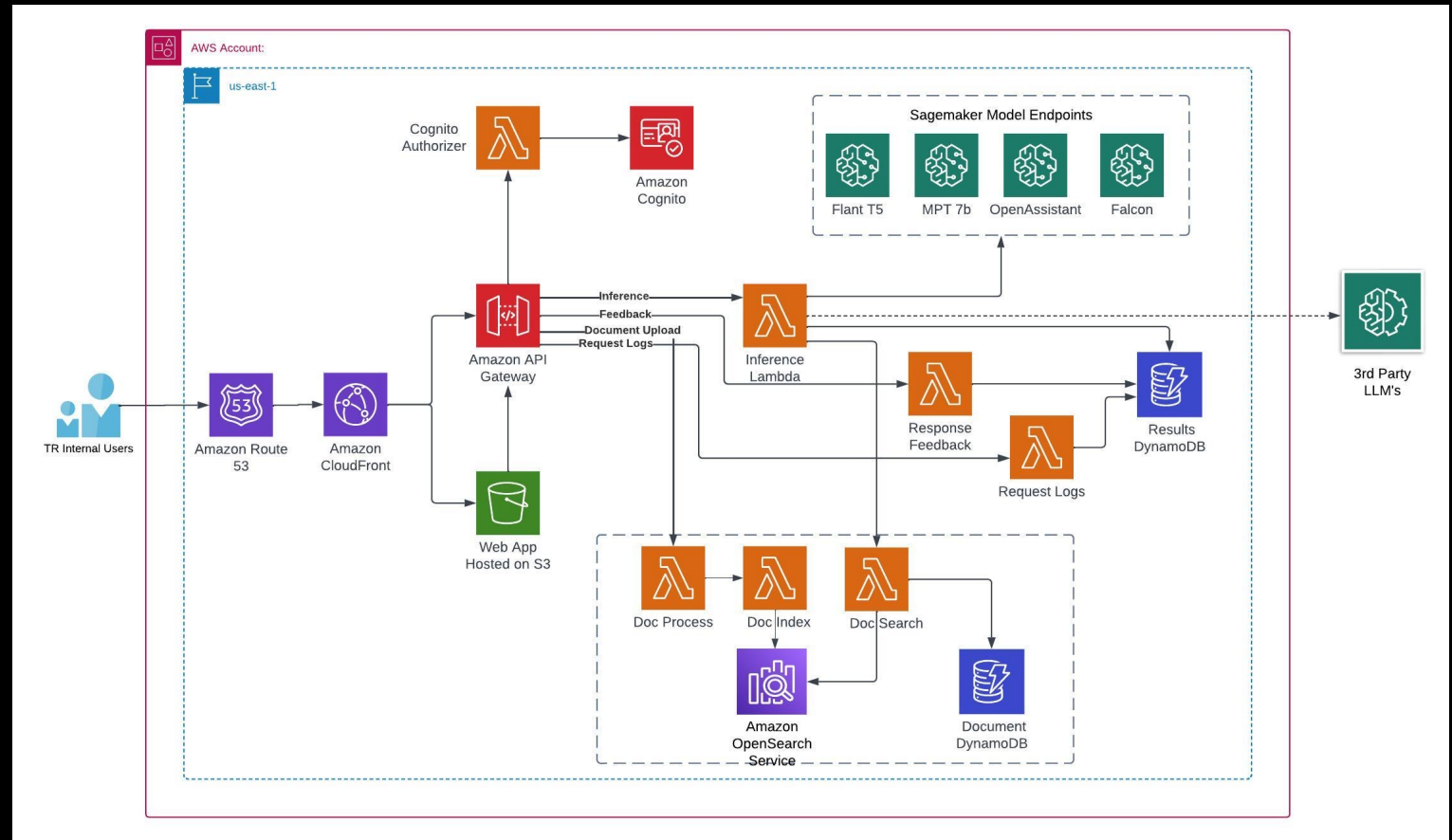


- A unified API interface across all FMs
- API quota, limits, and usage management
- Access control and model governance
- Content, privacy, and responsible AI policy enforcement

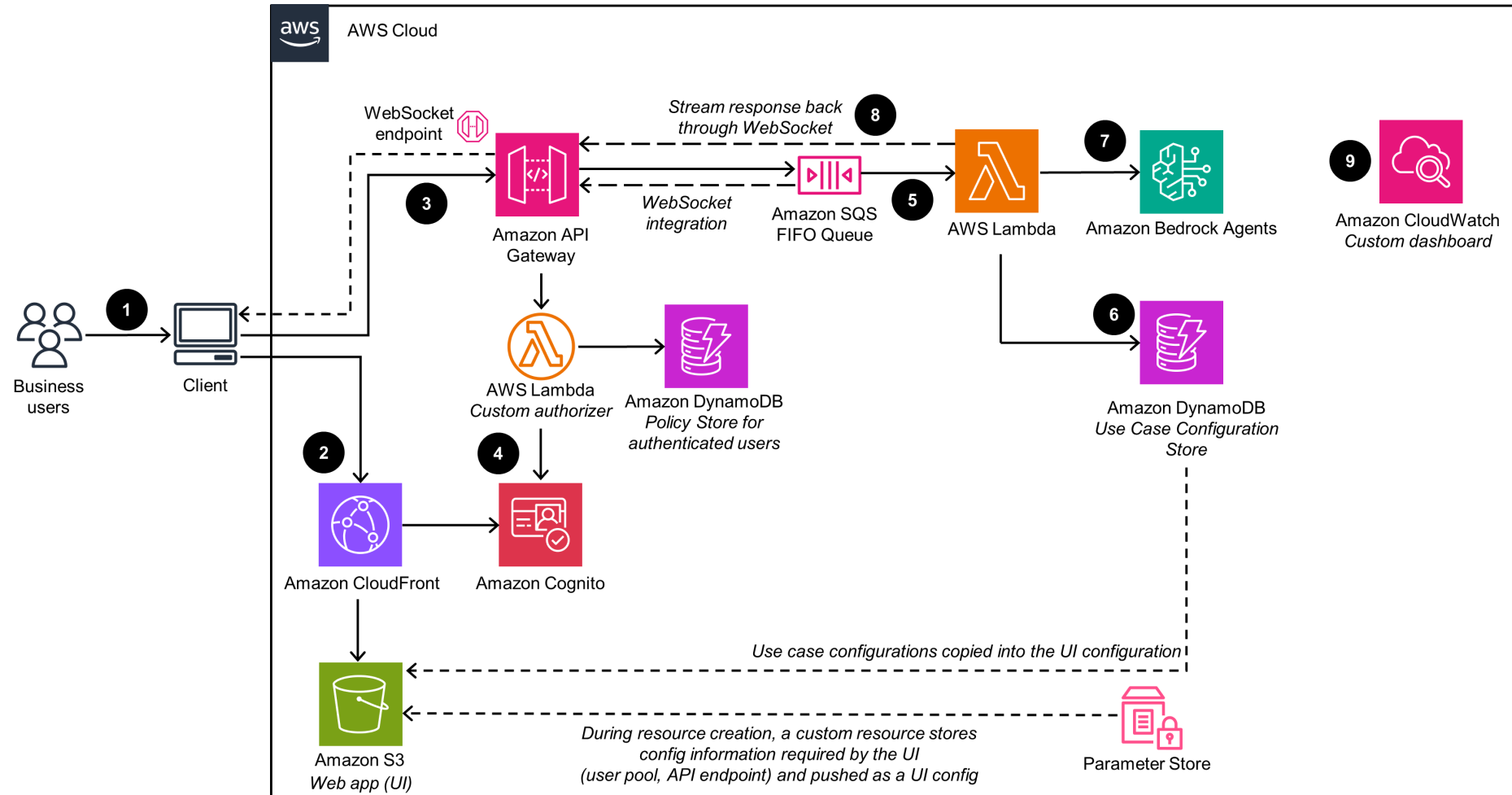
Democratize AI to Fuel Innovation – Thomson Reuters

Open arena – an enterprise-wide LLM playground

- Accelerate model deployment
- Safe and secure enterprise-wide platform
- Accurate data management
- Simplify user experience



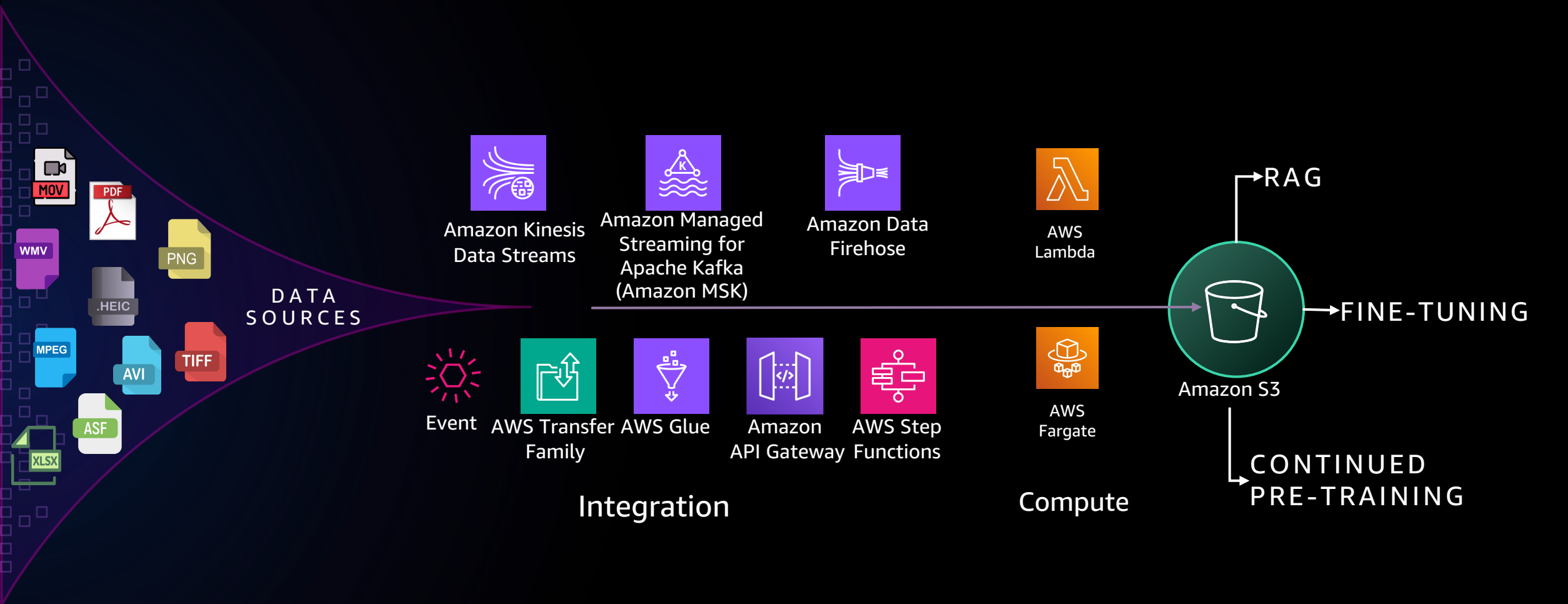
Get started today with a Solution Library



Data processing patterns

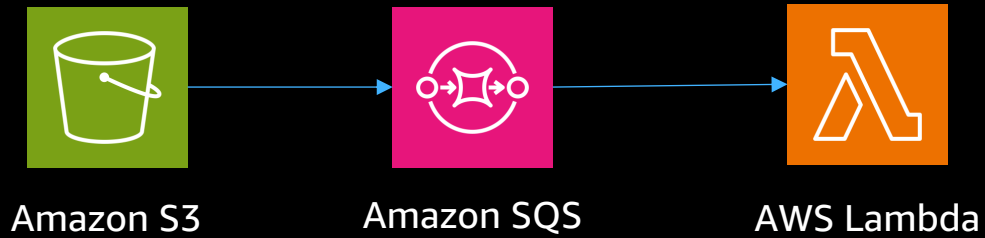


Building your data pipeline

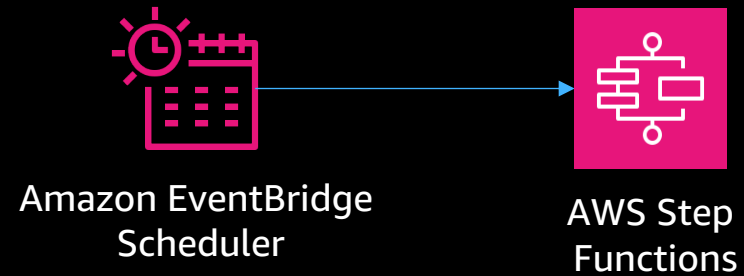


Data processing patterns

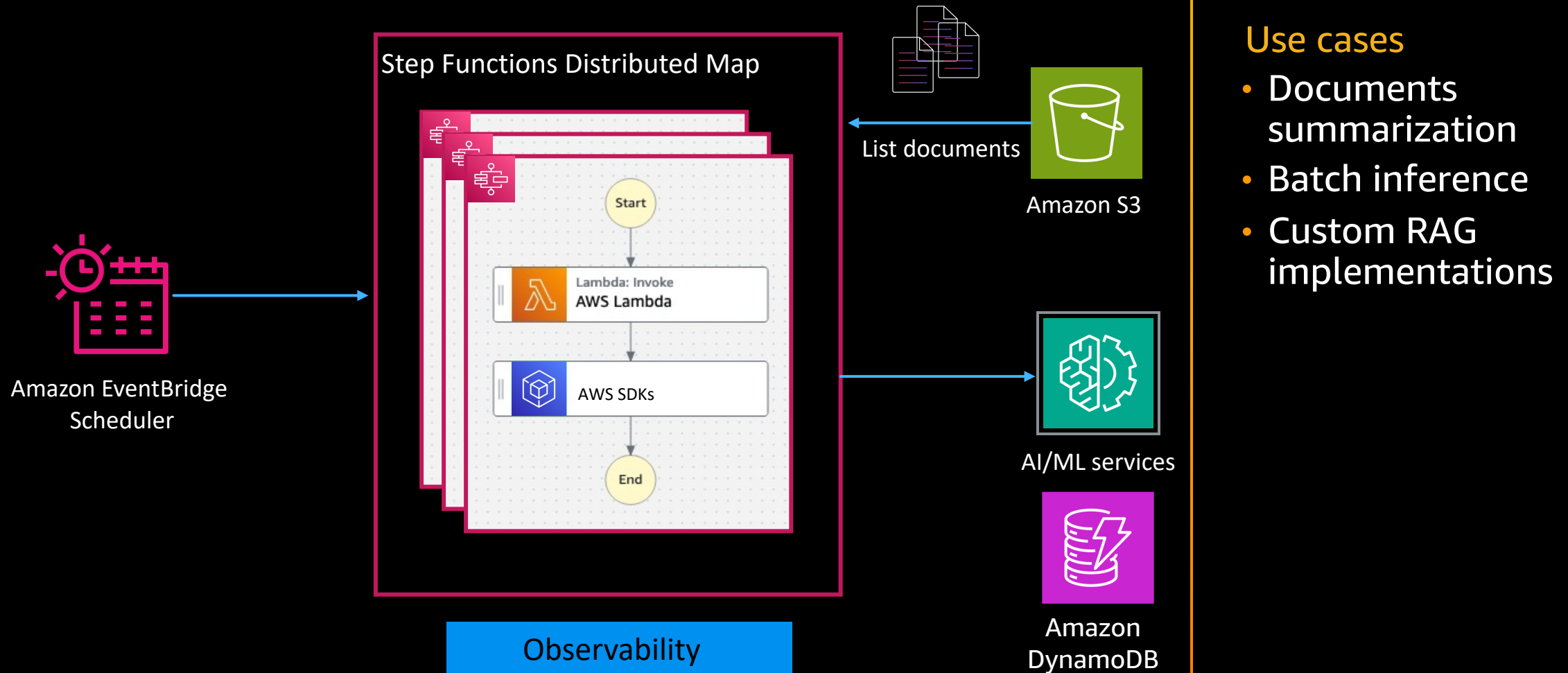
Event-drive data processing



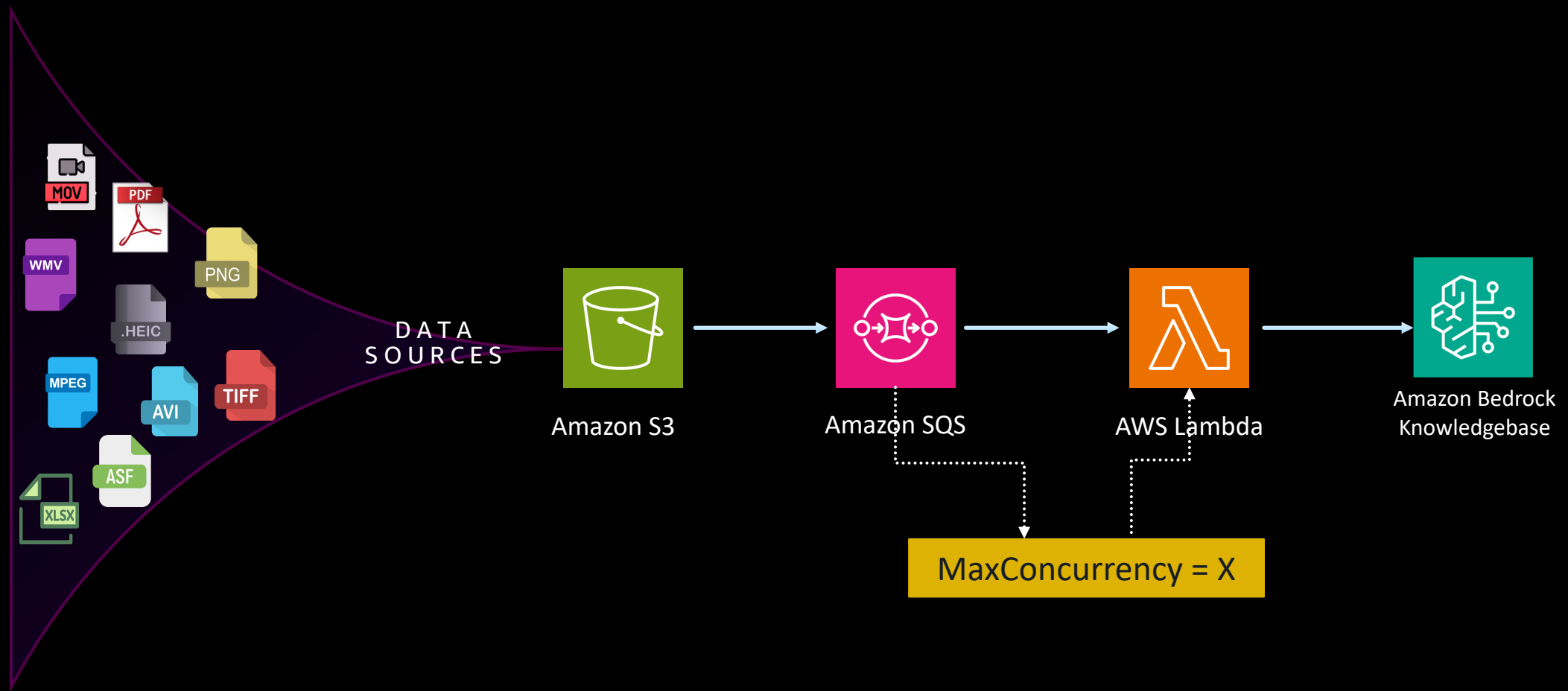
Batch data processing



Large scale document processing

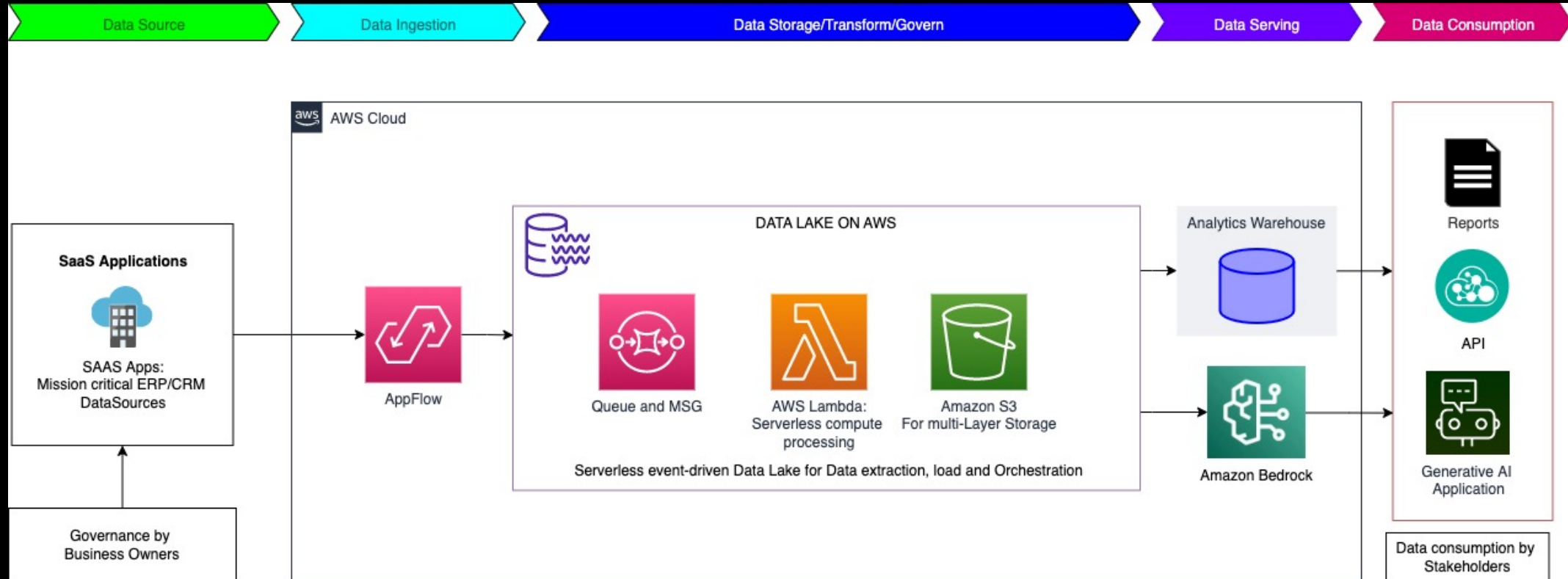


Event-driven re-indexing knowledge bases



OMRON Data and Analytics Platform

[OMRON Corporation](#) is a leading technology provider in industrial automation, healthcare, and electronic components.



- integrate data from multiple ERP and CRM systems
- Real-time transformations and calculations as data arrives
- Data democratization and business agility

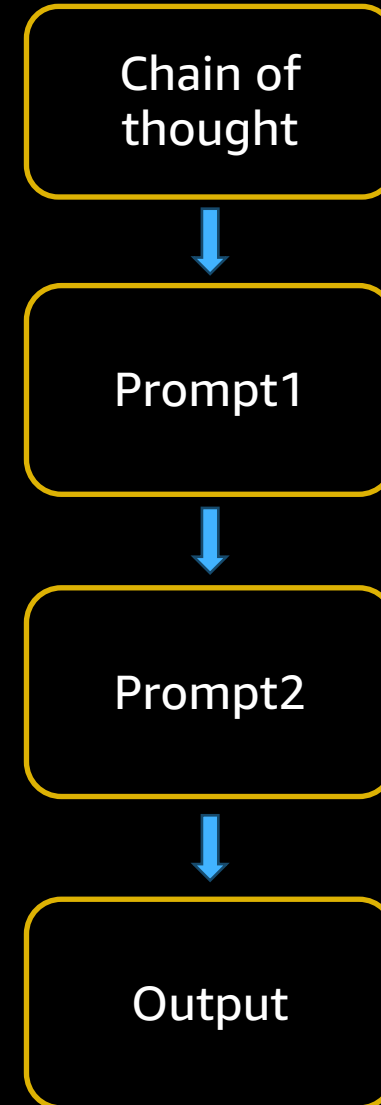


Prompt chaining

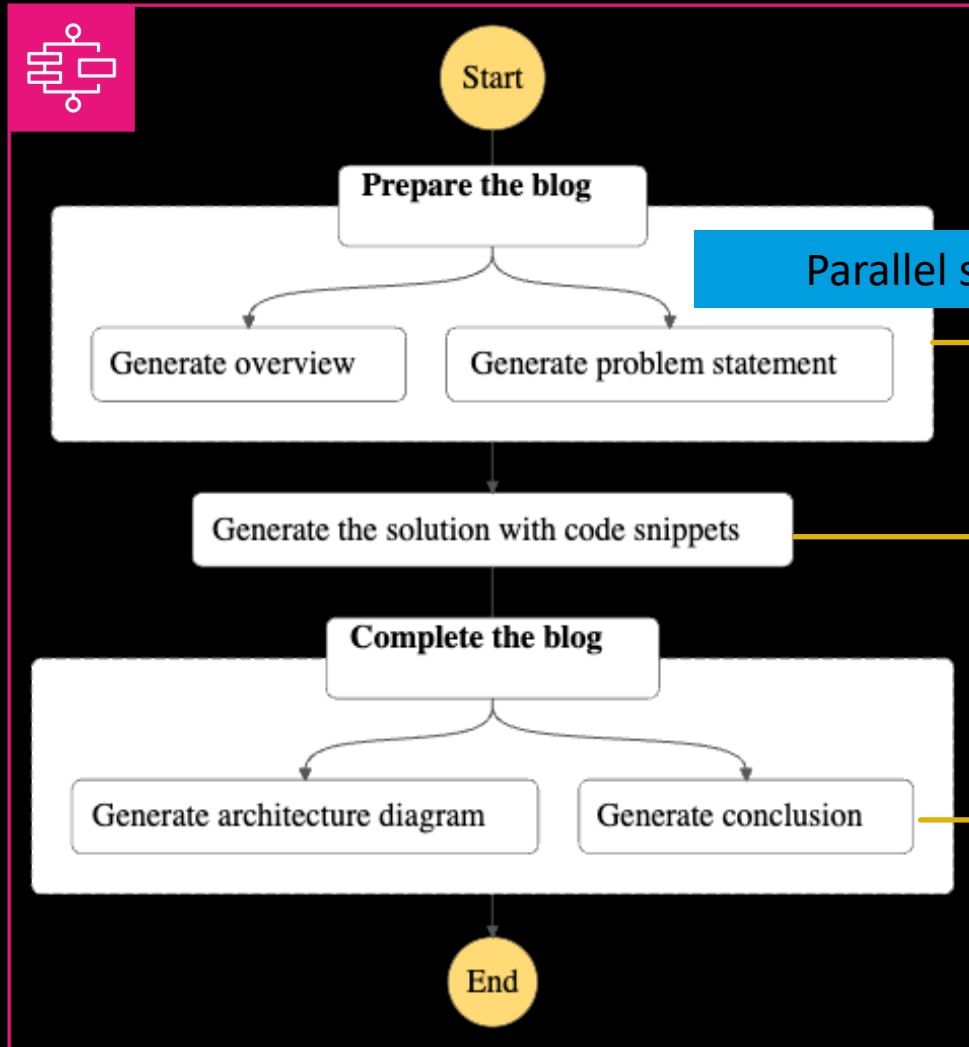
Decompose prompts into smaller and more manageable prompts and chain them

Why?

- Each change could potentially **impact** the rest of the prompt
- Can be **slower** and **expensive**
- Use purpose built models



Writing a blog with prompt chaining



Benefits

- Use purpose-built models
- More performant with smaller prompt and parallel processing
- Explainability

HTTP API



Image generation model



Coding model



Smaller model

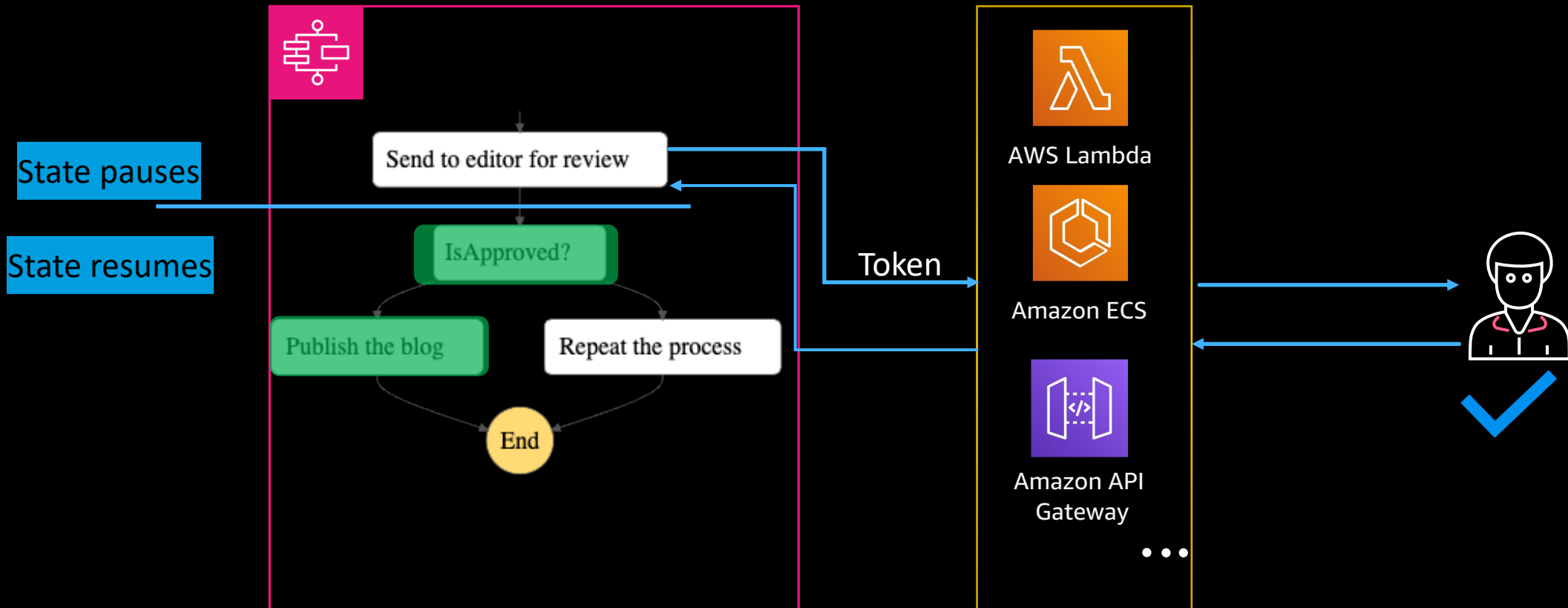
Observability

Human approval

- Safety and accuracy is critical
- Automated validation fails
- Data preparation as well as inferencing

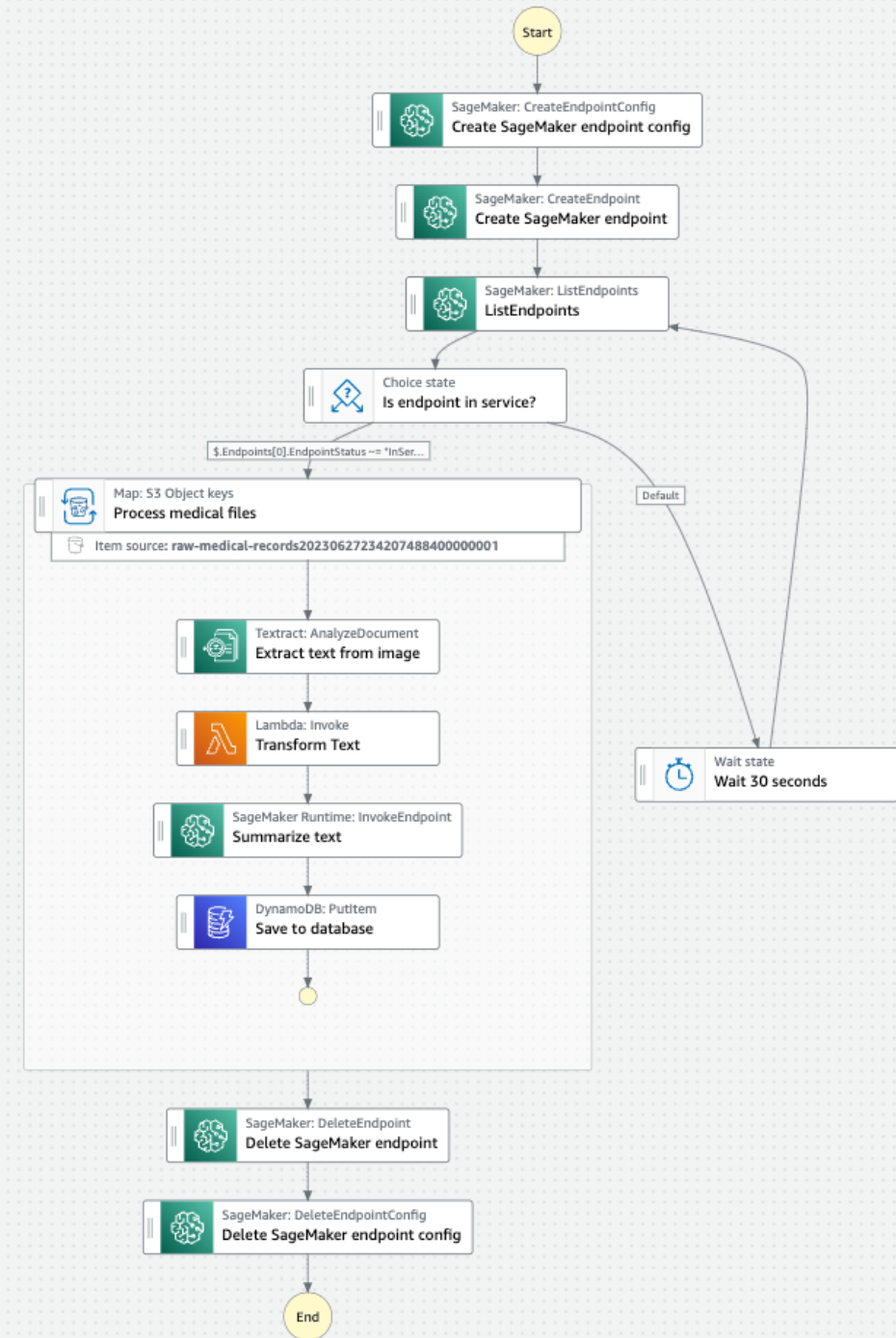


Human-in-the-loop pattern



Ephemeral computing pattern

- Provisions the endpoint for inference
- Run the inference
- Delete the endpoint



Thank you!

