Crafting a Secure and Scalable Generative AI Solution with AWS Serverless and Amazon Bedrock

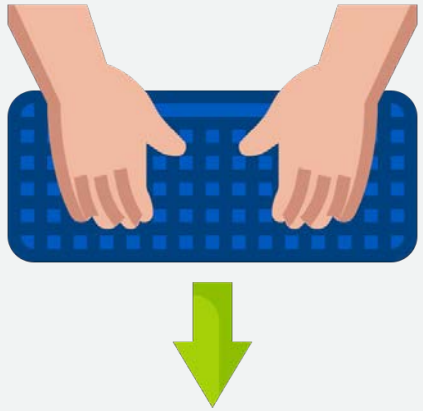
Samuel Baruffi

Principal Solutions Architect
Amazon Web Services (AWS)

Platform Engineering

- Platform Engineering as a Catalyst for Developer Productivity
- Navigating Complexity in Modern Tech Stacks
- Balancing Innovation with Stability
- Ensuring Consistent Security and Compliance
- Scaling Infrastructure Efficiently
- Enhancing Developer Experience

Our Solution



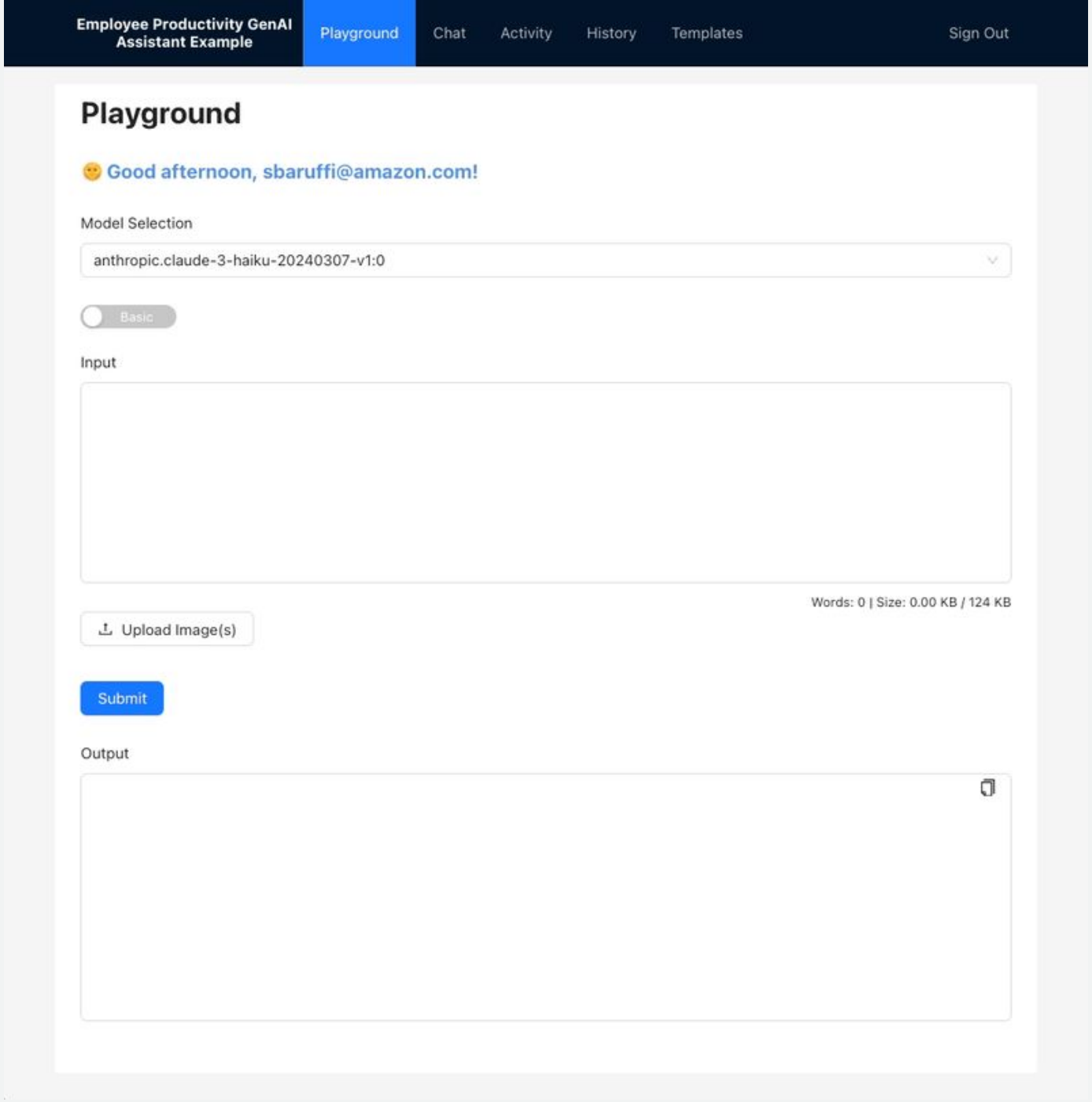
Input (text/images)

GenAI Processing

Output (polished content)

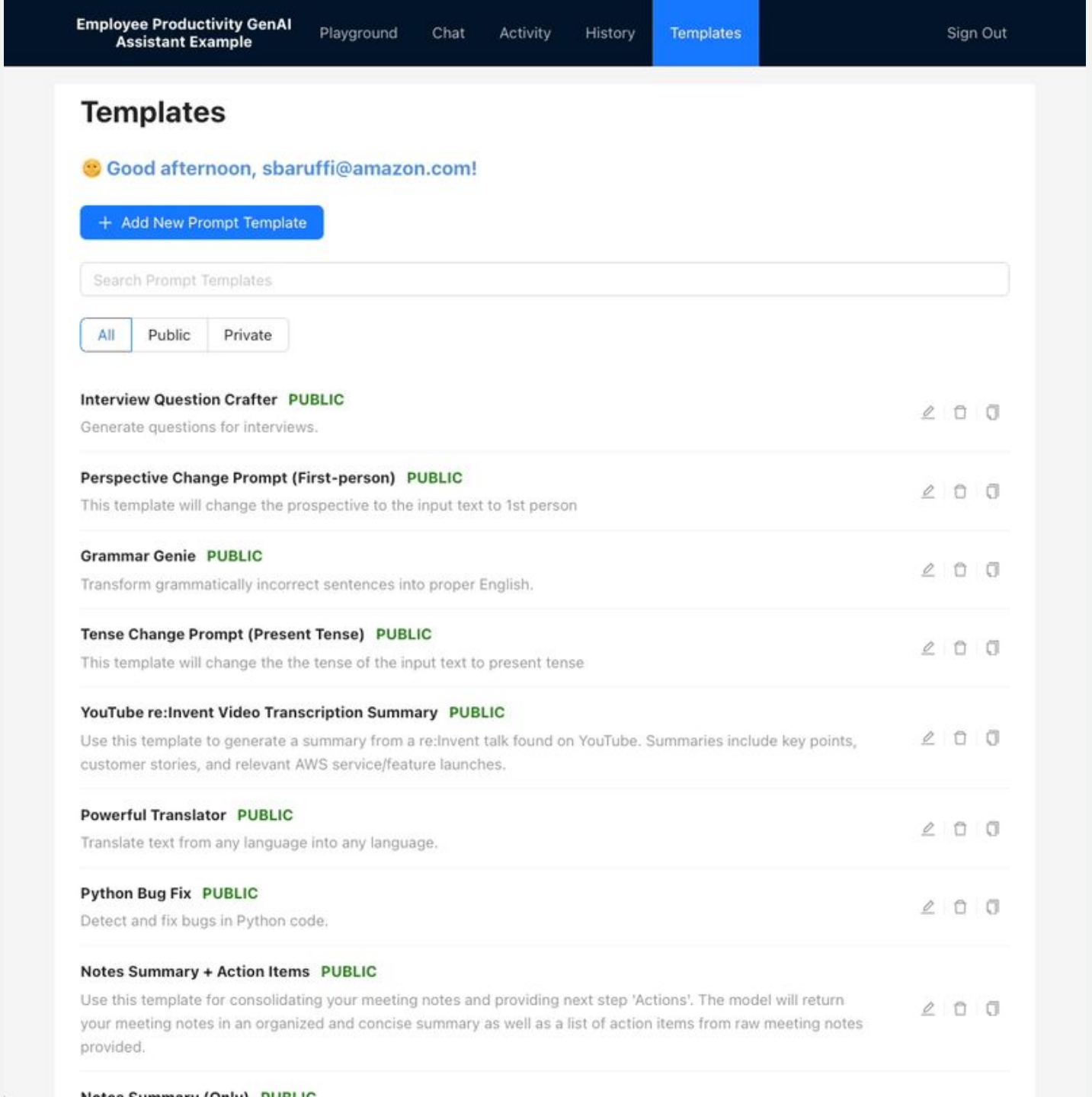
Easy to use

- Real-time interaction with AI models
- Secure with user authentication and authorization
- Streaming responses



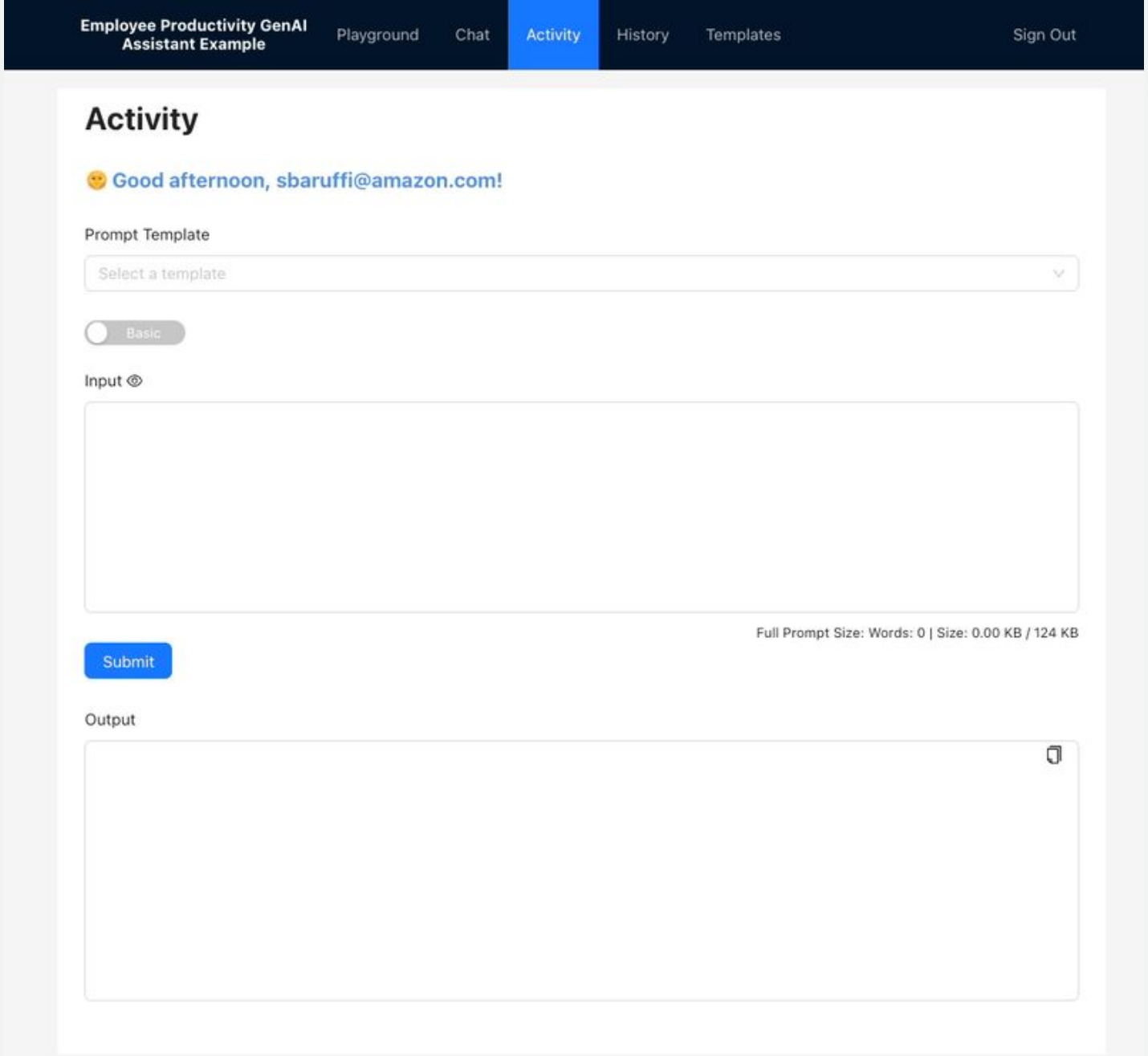
Templates

- Template creation (prompt reusability)
- Selection of models
- Private or public templates
- Easy searchability



Reusability

- Use created templates
- Copy responses
- Ability to change advanced settings

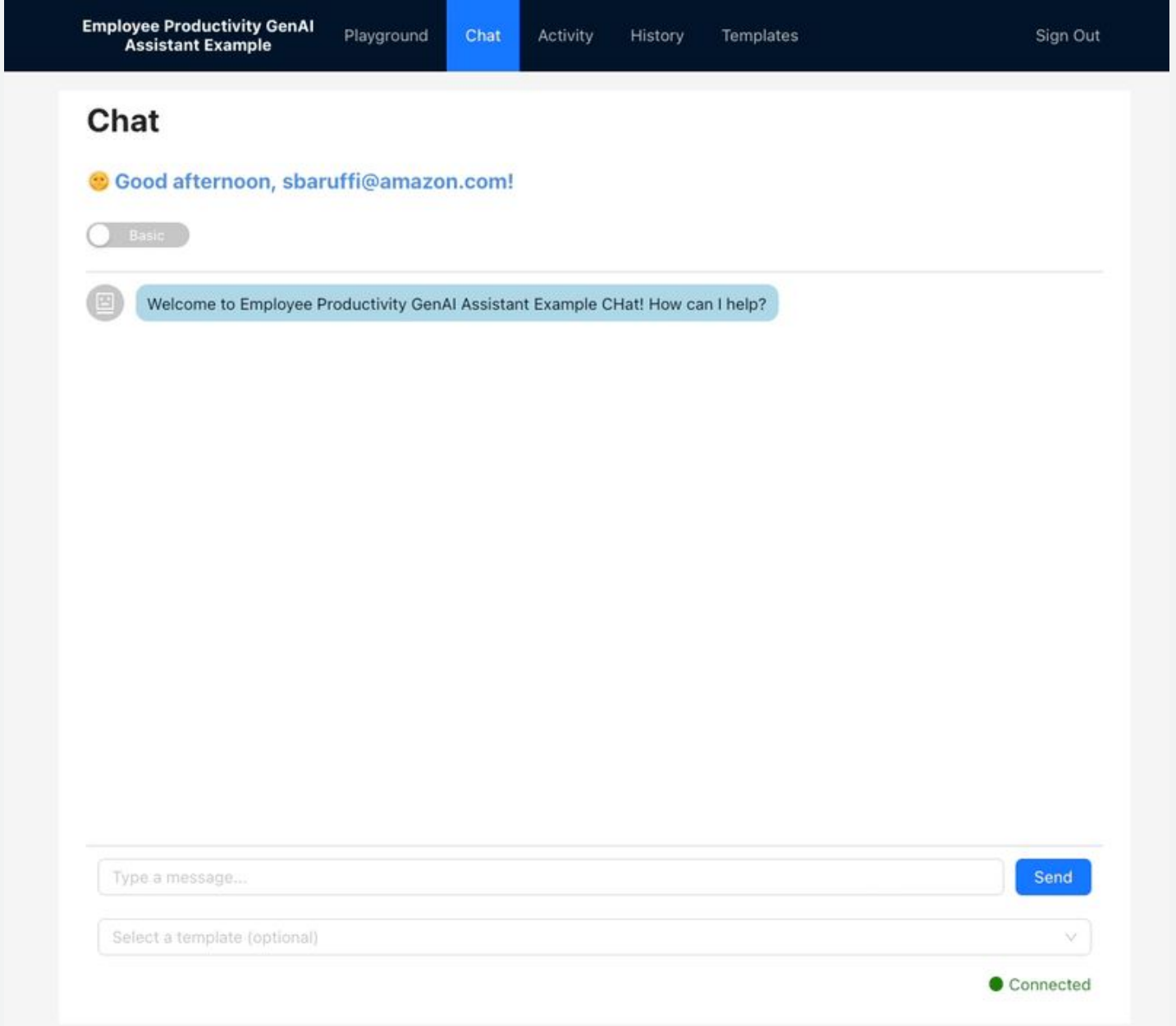


Multi Modal

- Provide images and text as inputs
- Supported on all Claude 3 models
- Vision capabilities

Chat

- Ability to chat using templates
- History context
- Streaming responses
- Choose advance settings



Boost Platform Engineering with GenAI

- **Employee Productivity GenAI Assistant**
  - Automate with AI
  - AWS Serverless & Bedrock

- **Benefits:**
  - Boost Productivity: Automated docs, templates, and more
  - Streamlined Workflows: Reusable templates, real-time interaction
  - Scalable & Secure: AWS-powered
  - Open source and simple deployment

AWS Services Overview: Simplified Explanations

Amazon API Gateway Rest API

Amazon Simple Storage Service (Amazon S3)

AWS WAF

Amazon API Gateway WebSocket

Amazon Cognito

Amazon CloudFront

AWS Lambda

Amazon DynamoDB

Amazon Bedrock

Amazon **Bedrock**

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

Choice of leading FMs through a single API

Model customization

Retrieval Augmented Generation (RAG)

Agents that execute multistep tasks

Security, privacy, and safety

Anthropic Claude Models on Bedrock

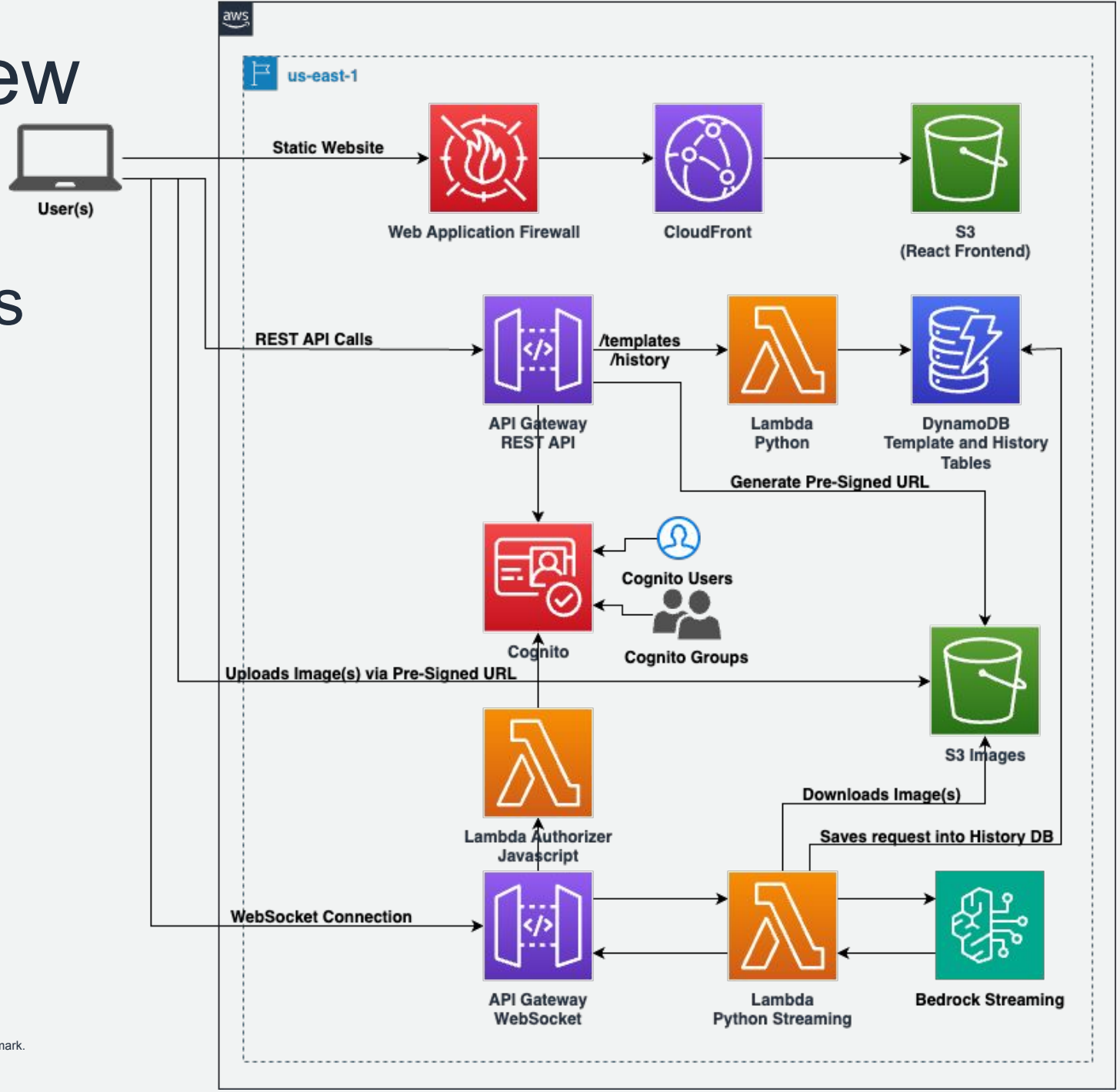
CHOOSE THE EXACT COMBINATION OF INTELLIGENCE, SPEED, AND COST TO SUIT YOUR NEEDS

	Claude 3.5 Sonnet	Claude 3 Haiku	Claude 3 Sonnet	Claude 3 Opus
Use case	Most intelligent, built for high-volume use cases	Fastest performance at the lowest cost	Balance between intelligence, speed, and cost	Second-most intelligent overall; most intelligent in Claude 3 family
Context	200K	200K	200K	200K
Vision	✓	✓	✓	✓

*Per 1K tokens

Architecture Overview

- Fully Managed Services
- Real-Time Processing
- Security Built-In
- Generative AI models
- Nice UI built on React
- Using AWS serverless services



Cost Efficiency

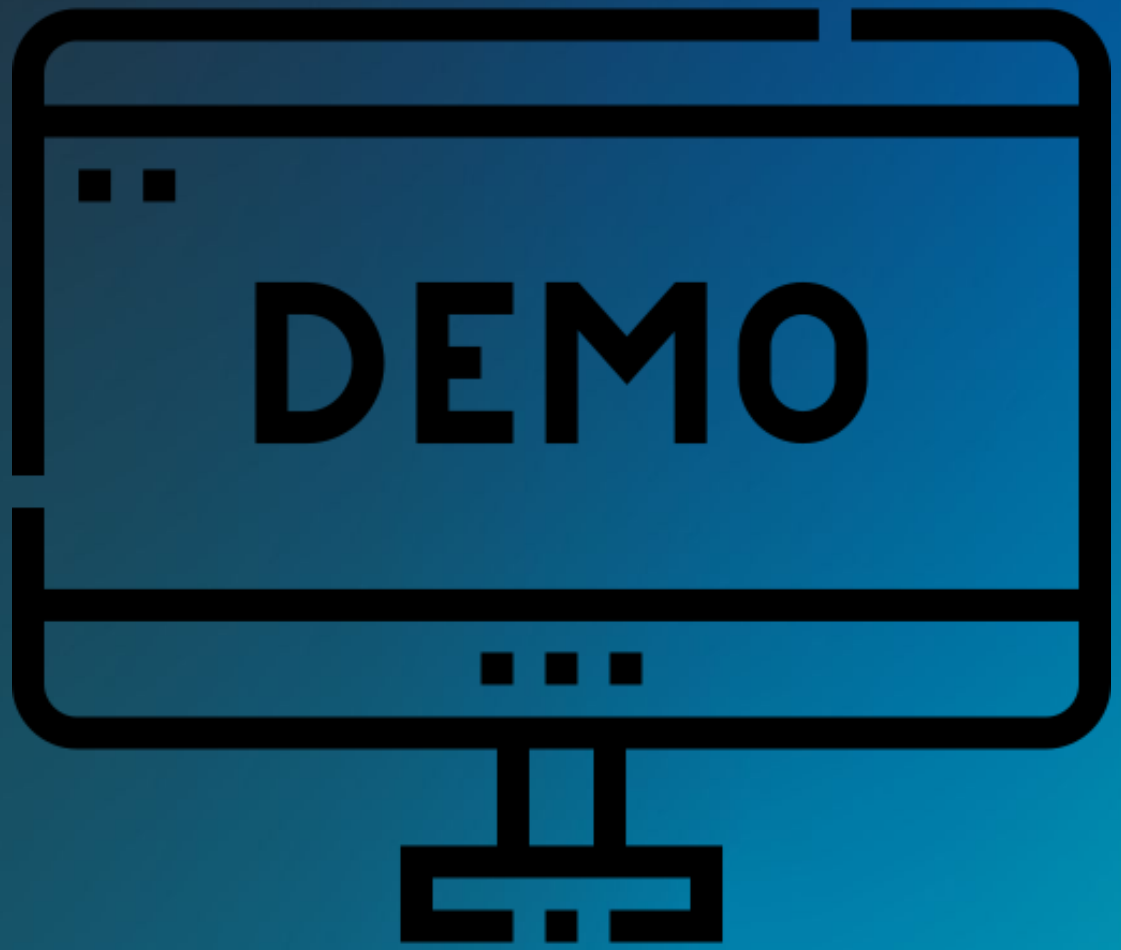
Our cost efficiency is based on the following usage assumptions:

- Users: **50 users**
- Usage: Each user utilizes the tool **5 times a day**
- Tokens: Average of **500 input tokens** and **200 output tokens** per request

Model	Bedrock Cost	Other AWS Services	**Total Monthly Cost**
Claude 3 Haiku	$2.81	$16.51	**$19.32**
Claude 3 Sonnet	$33.75	$16.51	**$50.26**
Claude 3.5 Sonnet	$33.75	$16.51	**$50.26**
Claude 3 Opus	$168.75	$16.51	**$185.26**

Get Started Now!

- Open source repository:
  - https://github.com/aws-samples/improve-employee-productivity-using-genai
- Simple and easy deployment
- Pay as You Go, secured and deployed in your AWS account (you have full control)



Thank you!