GENAI ДЛЯ БИЗНЕСА НА AWS

aws

qCloudy

КТО МЫ

qCloudy - команда сертифицированных инженеров с многолетним опытом проектирования, разработки и внедрения.

Мы строим

- гибкие
- масштабируемые
- отказоустойчивые
- cost-effective

решения в публичном облаке с применением AI/ML и поддерживаем их 24/7

aws

qCloudy

НАШИ УСЛУГИ

Консалтинг

- Облачная стратегия развития ИТ-инфраструктуры

Аудит и оценка

- Оценка IT инфраструктуры
- Оценка Data инфраструктуры
- Оценка AI/ML решений
- Оценка квалификации сотрудников
- Оценка качества архитектуры

Разработка и поддержка облачных решений

- Миграция в облако
- Оптимизация облачных расходов
- Построение гибридного облака
- Аварийное восстановление облака
- Cloud DevOps
- Разработка Cloud Navive приложений
- Озеро данных и хранилища данных
- DataOps
- Генеративный ИИ
- Компьютерное зрение
- MLOps/FMOps
- Комплексное управление облачной инфраструктурой

Обучение

- AWS Workshops

BAITEREK

КЕЙС #1 ЧАТБОТ С ИИ НА БАЗЕ AWS ДЛЯ ХОЛДИНГА БАЙТЕРЕК

aws

qCloudy

ВЫЗОВЫ И РЕШЕНИЕ

BAITEREK

ВЫЗОВЫ

- Снижение нагрузки на операторов поддержки
- Необходимость круглосуточного обслуживания клиентов
- Снижение операционных затрат

РЕШЕНИЕ

- AI-чатбот на основе Generative AI для автоматической обработки клиентских запросов на казахском и русском языках
- Few-shot prompting + RAG
- Guardrails:
  - content filtering
  - word filtering
- Cross-Region Inference Profile

AWS SERVICES USED

- Amazon Bedrock
- AWS Lambda
- AWS API Gateway
- AWS Amplify
- AWS Cognito
- AWS DynamoDB
- S3
- AWS Elasticache

aws

qCloudy

КЛЮЧЕВЫЕ РЕЗУЛЬТАТЫ

BAITEREK

Среднее время ответа
~10с

Повышена точность и
согласованность ответов

SLA
99,9%

Доступность 24/7
и мгновенные ответы

aws

qCloudy

КАЗАГРОФИНАНС

КЕЙС #2
ИНВЕНТАРИЗАЦИЯ СЕЛЬХОЗ ТЕХНИКИ

aws

qCloudy

ВЫЗОВЫ И РЕШЕНИЕ

КАЗАГРОФИНАНС

ВЫЗОВЫ

- Неточности при ручной инвентаризации техники
- Операционные расходы из-за частых выездов сотрудников в регионы
- Отсутствие объективных критериев оценки состояния техники

РЕШЕНИЕ

- Автоматизированная AI-система анализа фото AWS Bedrock
- Event Driven архитектура
- Serverless решение

AWS SERVICES USED

- Amazon Bedrock
- AWS Lambda
- AWS API Gateway
- AWS Cognito
- AWS DynamoDB
- S3

aws

qCloudy

КЛЮЧЕВЫЕ РЕЗУЛЬТАТЫ

КАЗАГРОФИНАНС

с нескольких недель до 2–3 дней
сократилось время на инвентаризации

Снизились коррупционные риски

с 3 000 до 18 000 за сезон
выросло количество обработанных предметов

в 3 раза
сократились расходы на инфраструктуру в период неактивности

aws

qCloudy

СПАСИБО

aws

qCloudy

Александр Бернадский

Solutions Architect

Amazon in Kazakhstan

TECHNODOM

СЕРГЕК

KAZPOST

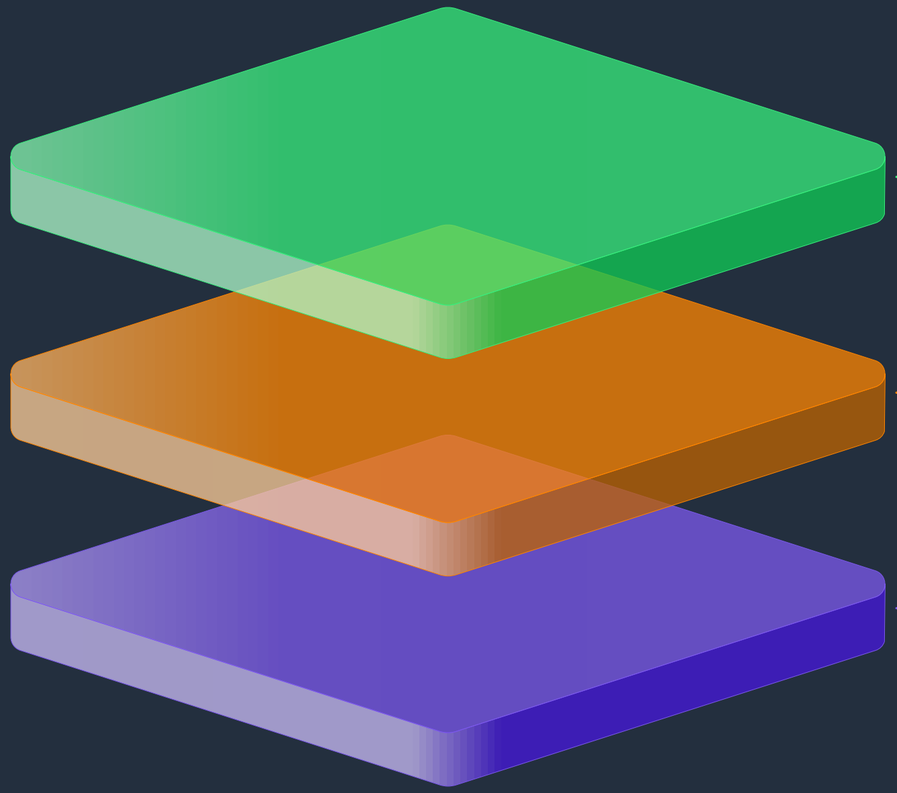
PHILIP MORRIS INTERNATIONAL

FREEDOM finance

CENTERCREDIT

amazon project kuiper

AWS Generative AI Stack



APPLICATIONS TO BOOST PRODUCTIVITY

Amazon Q Business
INSIGHTS AND AUTOMATION

Amazon Q Developer
SOFTWARE DEVELOPMENT LIFECYCLE

MODELS AND TOOLS TO BUILD GENERATIVE AI APPS

Amazon Bedrock
AMAZON MODELS | PARTNER MODELS

INFRASTRUCTURE TO BUILD AND TRAIN AI MODELS

Amazon SageMaker AI
MANAGED INFRASTRUCTURE

AWS Trainium
AWS Inferentia

GPUs

HIGH PERFORMANCE COMPUTE

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

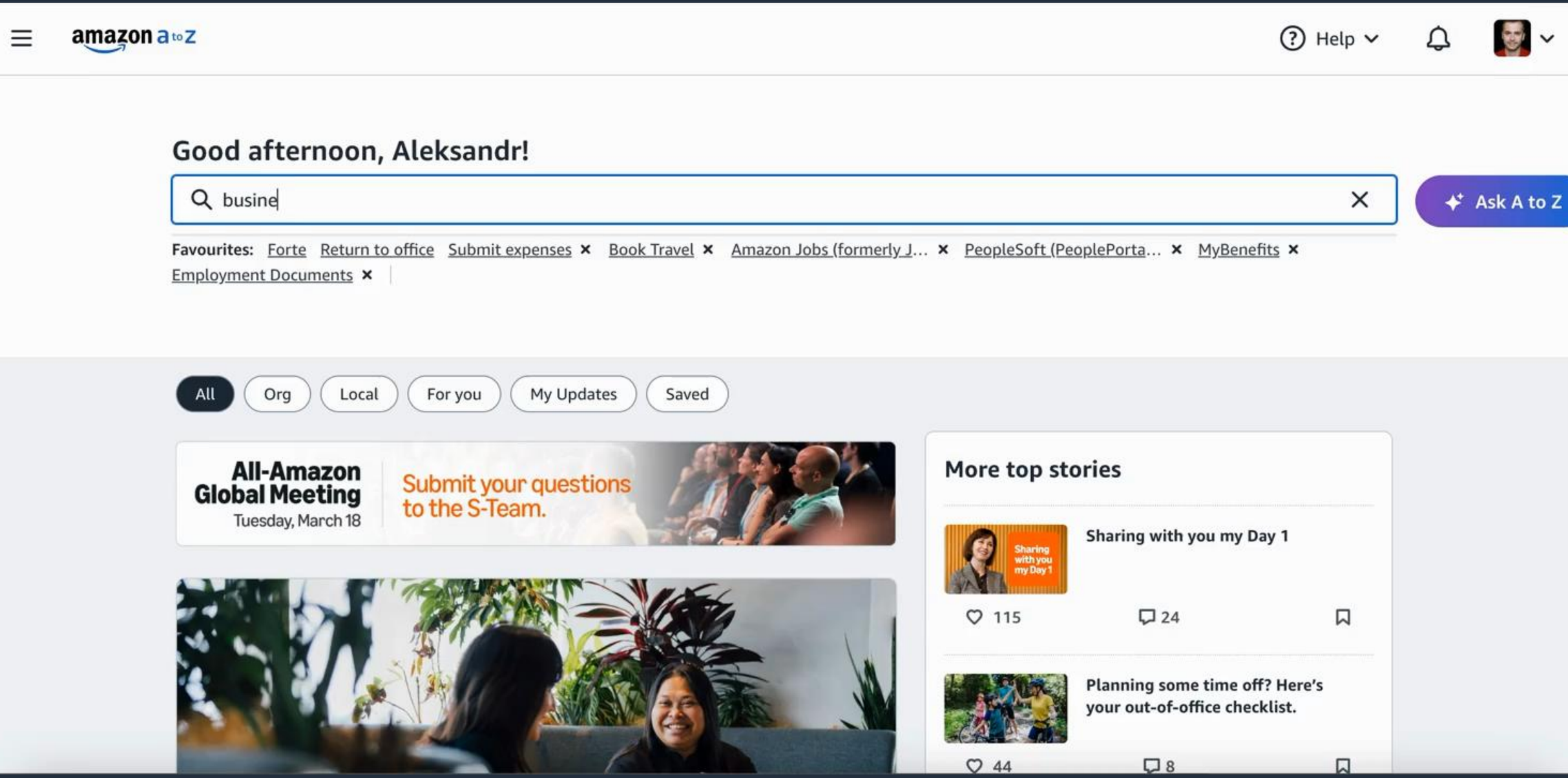
Security, privacy, and safety

Amazon Bedrock

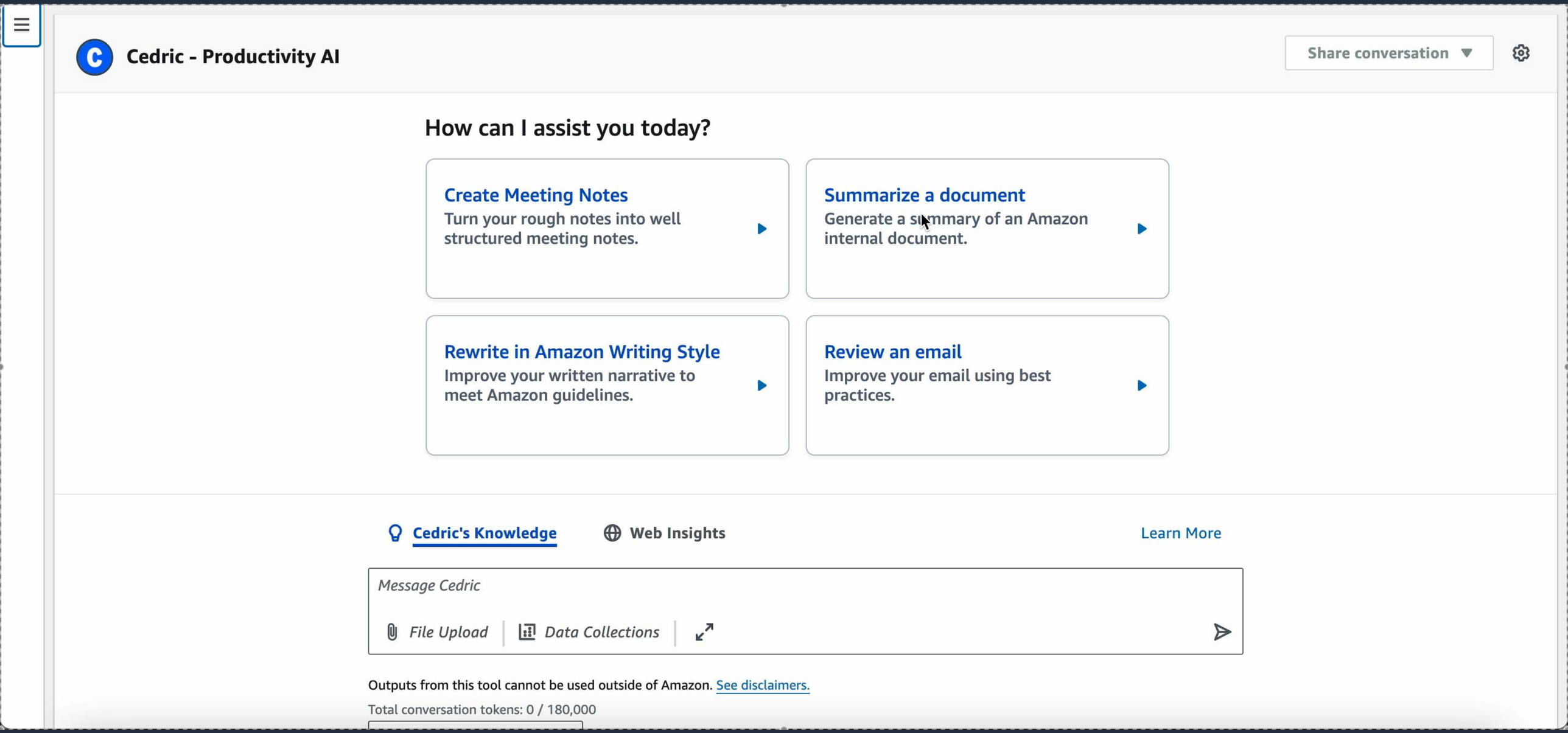
BROAD CHOICE OF MODELS

AI21labs	amazon	ANTHROPIC	cohere	Luma	Meta	MISTRAL AI_	poolside	stability.ai
Effective reasoning & rapid analysis for long context windows	Frontier multimodal intelligence at low-latency, Agent & RAG Applications, high-quality image & video generation	Advanced reasoning & coding capabilities, including computer use skills	Multimodal search & advanced retrieval powering multilingual knowledge agents	High-quality video generation from text & images	Advanced image & language reasoning	Knowledge summarization, expert agents, & code completion	Software engineering AI for large enterprises	High-quality AI image generation, easily deployable at scale
JAMBA	**AMAZON NOVA**	**CLAUDE**	**COMMAND** **EMBED** **RERANK**	**LUMA RAY 2**	**LLAMA**	**MISTRAL** **MIXTRAL**	**MALIBU** **POINT**	**STABLE DIFFUSION** **STABLE IMAGE**

How we use Generative AI in AWS



How we use Generative AI in AWS



How we use Generative AI in AWS

Grzegorz Ochmanski • 2-й
AWS for Games Senior Solutions Architect w Ama...
2 дн. • Отредактировано •

+ Отслеживать ···

Today, I wrote a Streamlit app a chatbot frontend with Amazon Cognito login. It is running on ECS with Graviton EC2, ALB, DynamoDB for per-user chat persistence, EFS for document and image storage, and Amazon Bedrock with Claude Sonnet 3.7. I wrote maybe 50 lines of code; the rest was handled by Amazon Q Developer CLI. The app has full CI/CD and IaC (Terraform).

It supports file handling, an image gallery, full logging in CloudWatch, etc.

Took me 6 hours.

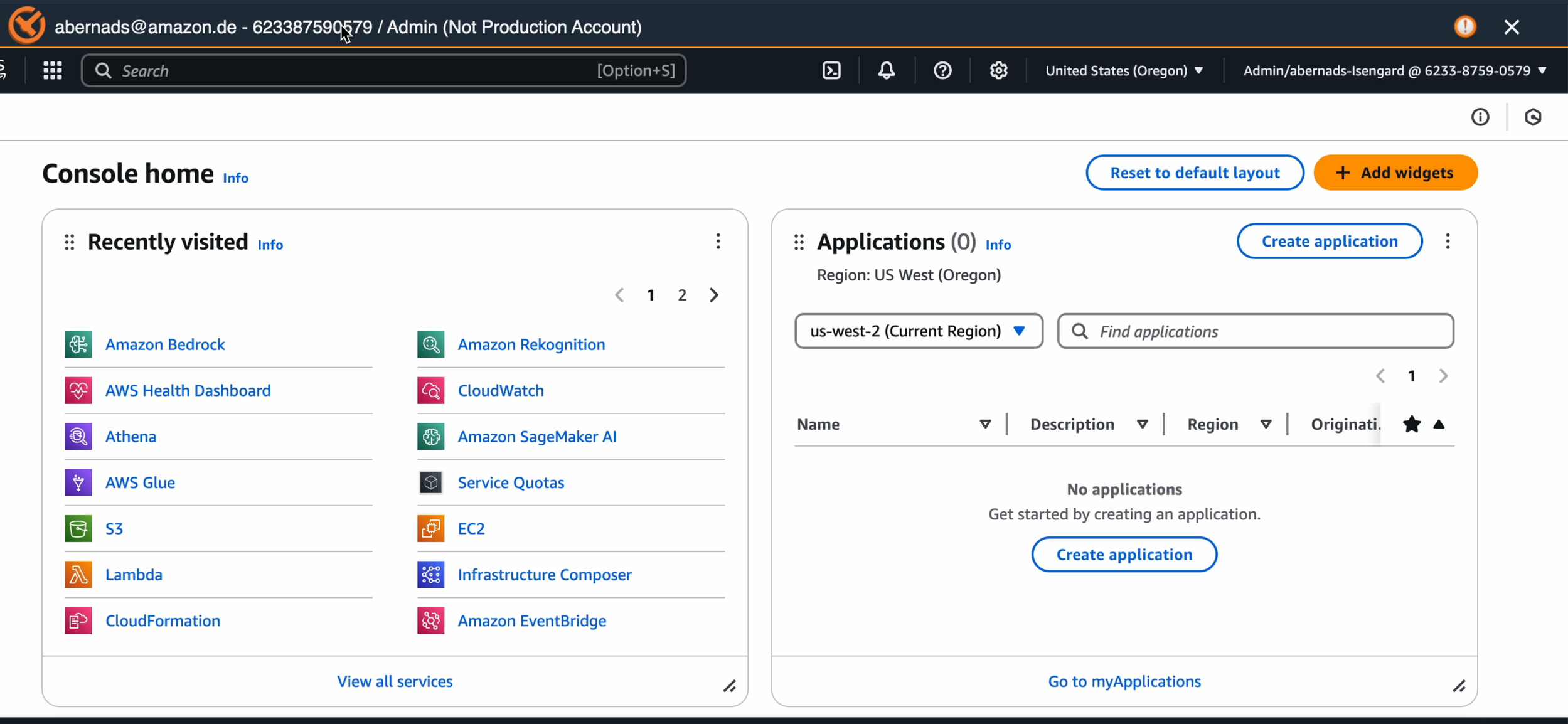
For me, this setup is much better than Cursor since I can use my trusted VS Code IDE.

https://lnkd.in/dPG–zx5

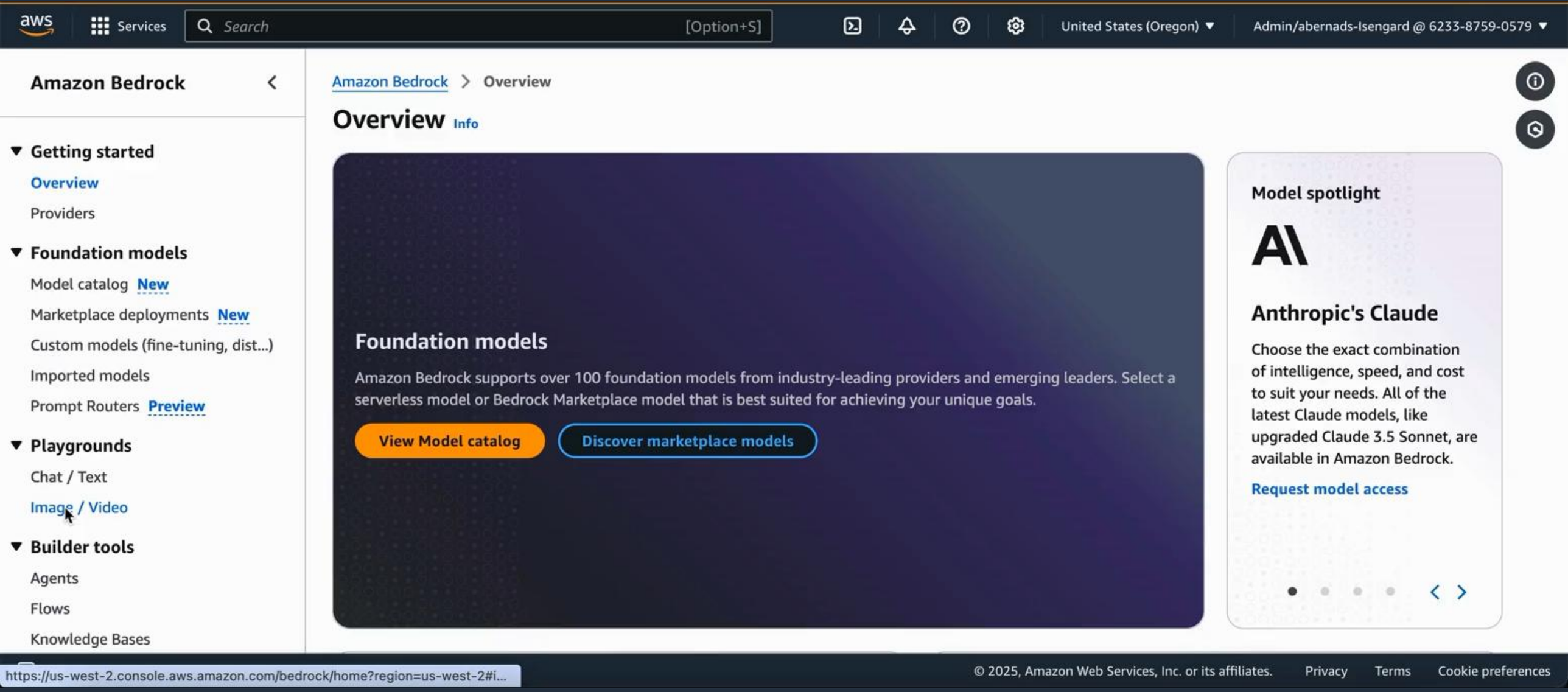
You still need to understand what you’re building, know the syntax and logic—because sometimes it drifts into insanity, but with a reasonable chat, it gets back on track.

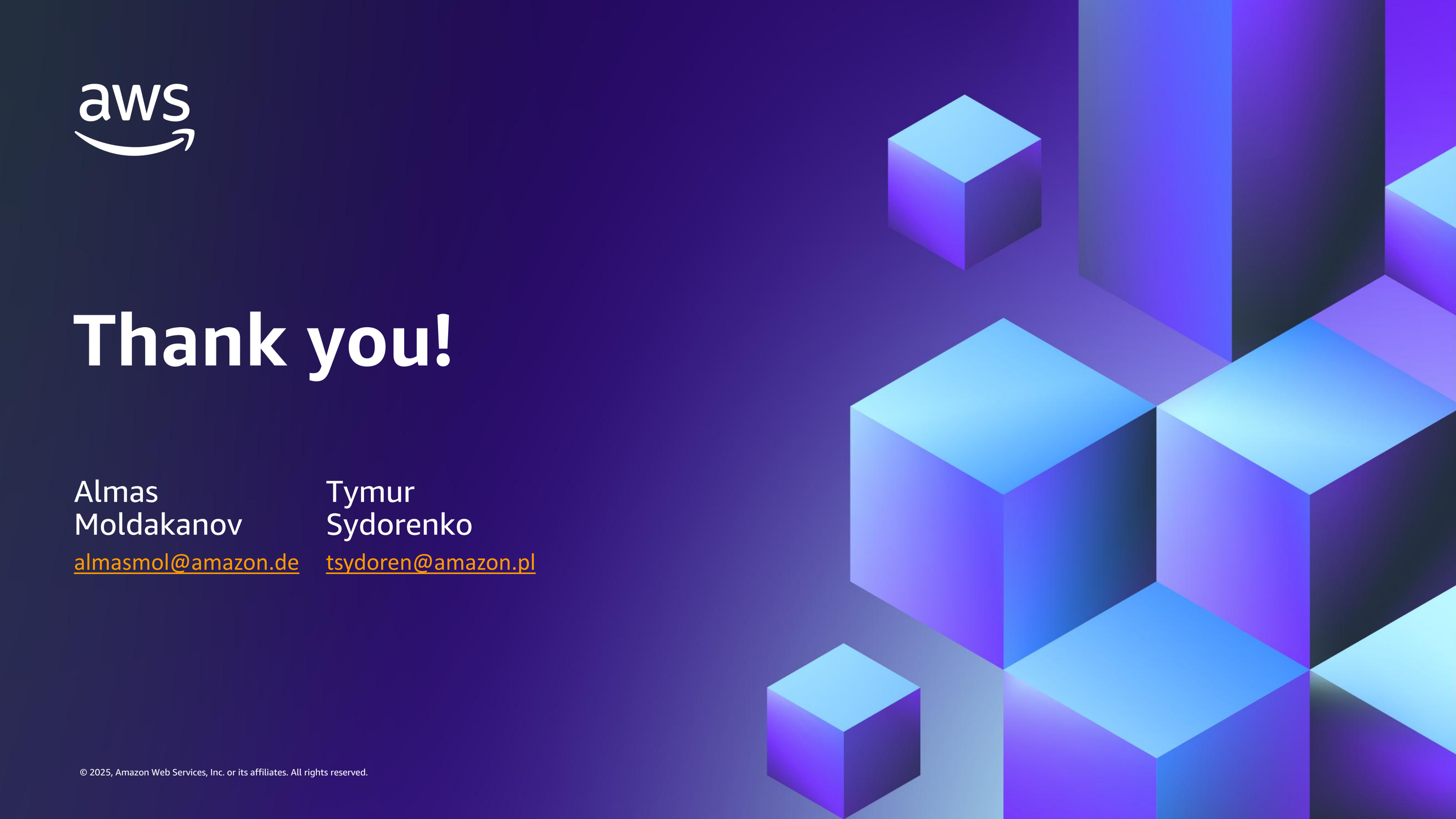
I give it my seal of approval!

TRY IT YOURSELF



TRY IT YOURSELF





Thank you!

Almas Moldakanov
almasmol@amazon.de

Tymur Sydorenko
tsydoren@amazon.pl