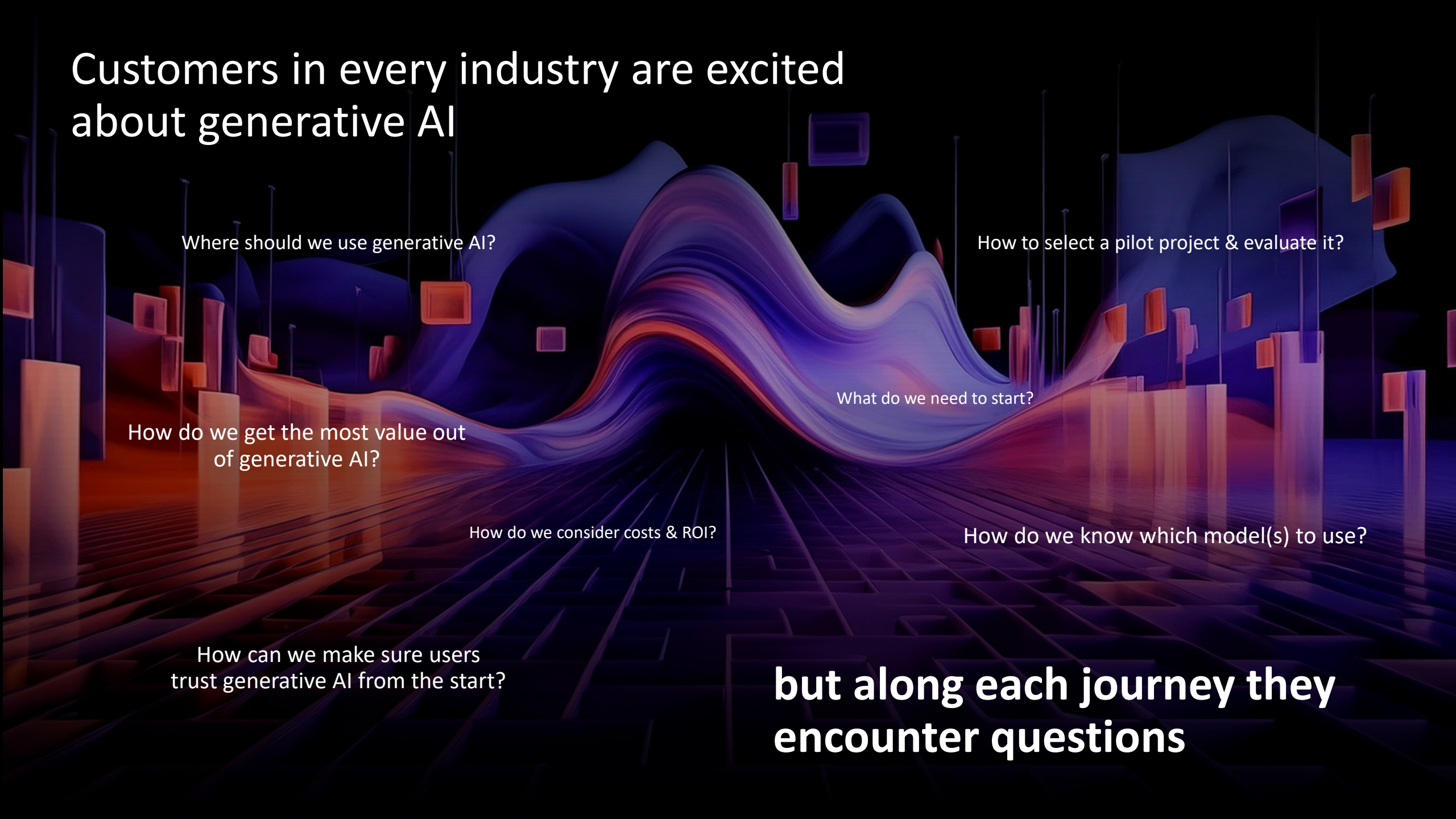
Accelerating Generative AI: Turning Answers into Action

Dilip Kumar (he/him)

VP, Amazon Web Services

Customers in every industry are excited about generative AI

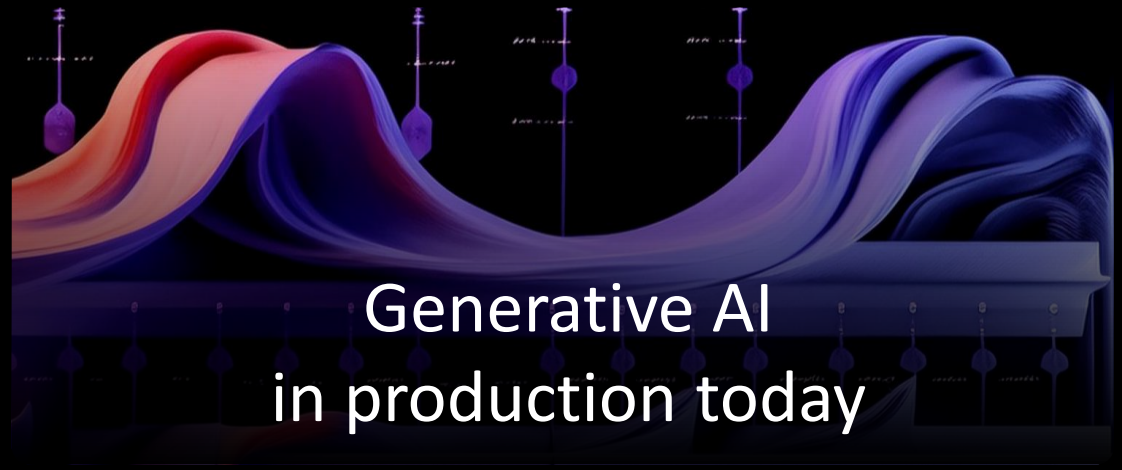


but along each journey they encounter questions

What has AWS learned with AI and ML?

25+

years innovating with AI/ML



Secure by design

Focus on the value for the customer

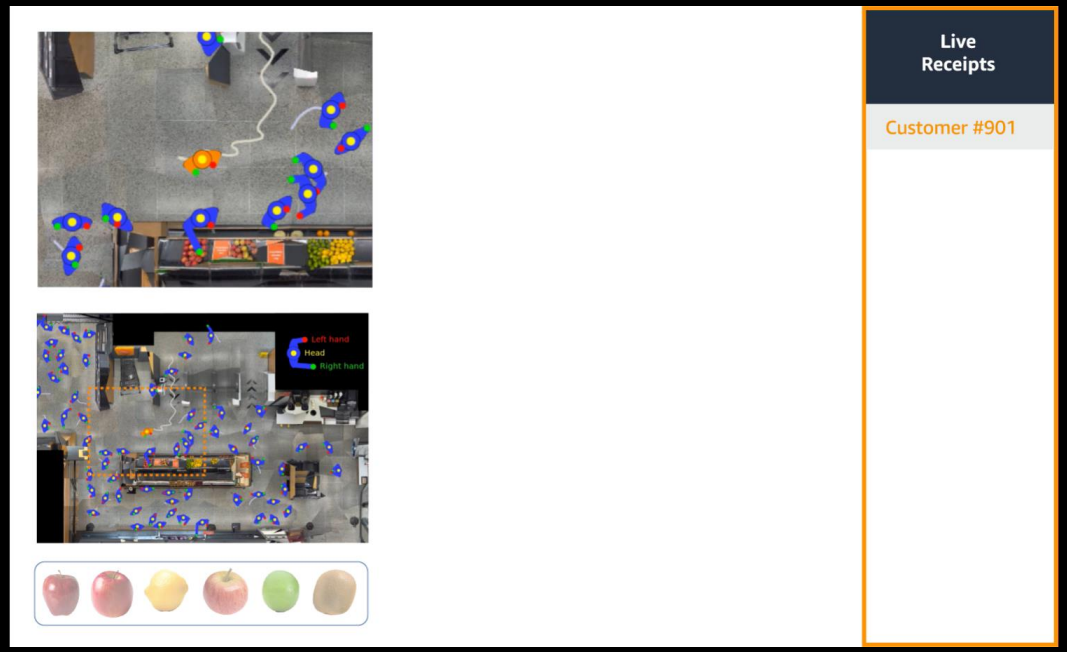
Scale matters

Experience & learnings passed on to customers

Creating new experiences and businesses with AI

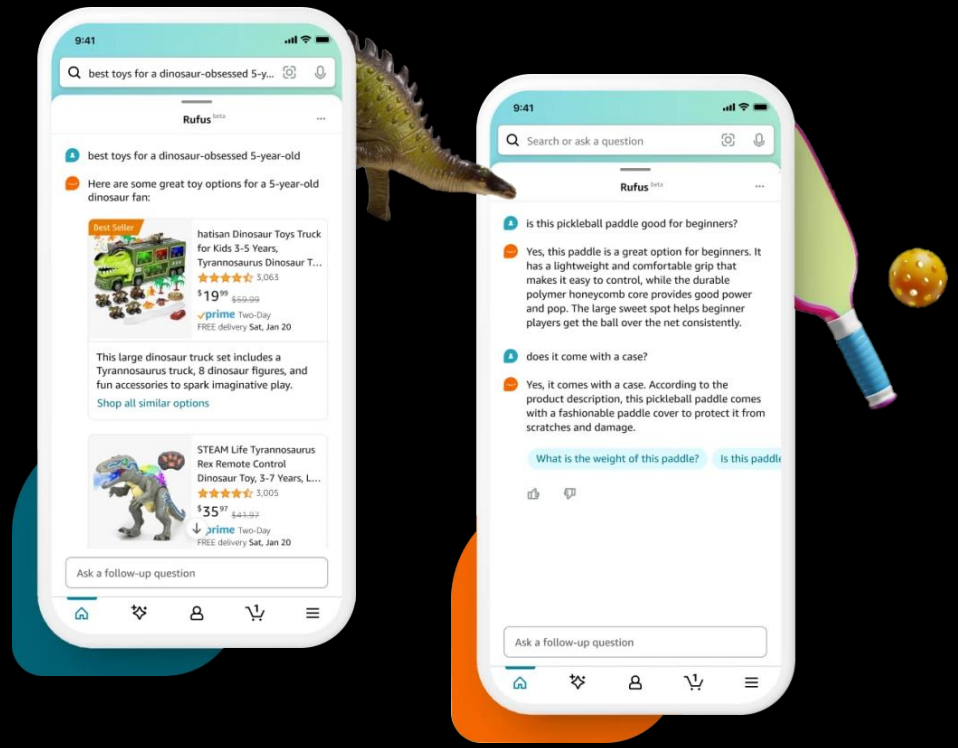
Just Walk Out

A fast and frictionless shopping experience



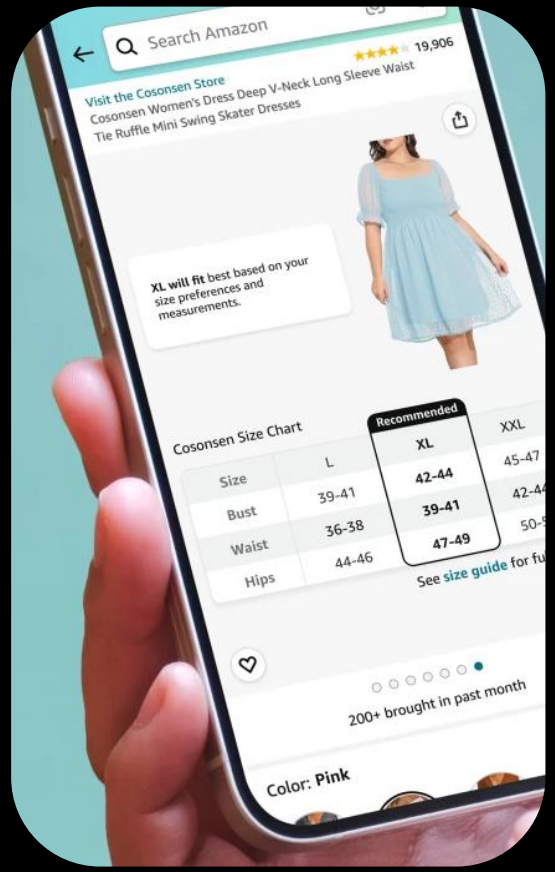
Rufus

Generative AI-powered expert shopping assistance



4

Driving **efficiencies at scale** with AI



4,000 products per minute
sold on Amazon.com

AI throughout the retail value chain

Personalized Recommendations

Fulfillment Centers

Middle and Last Mile Shipping

1.6M packages
every day

5

Move

LEARN.
EXPLORE.
TEST.

Build

CUSTOMIZE.
INTEGRATE.
EVALUATE.

Scale

LAUNCH.
MANAGE.
REPEAT.

Move

LEARN.
EXPLORE.
TEST.

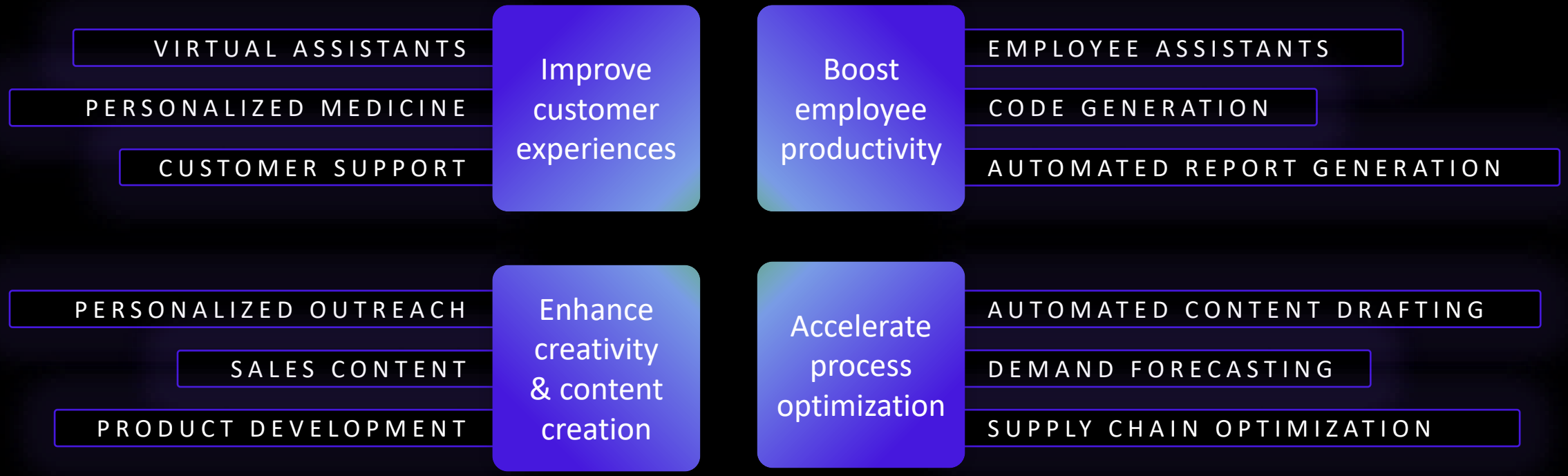
Build

CUSTOMIZE.
INTEGRATE.
EVALUATE.

Scale

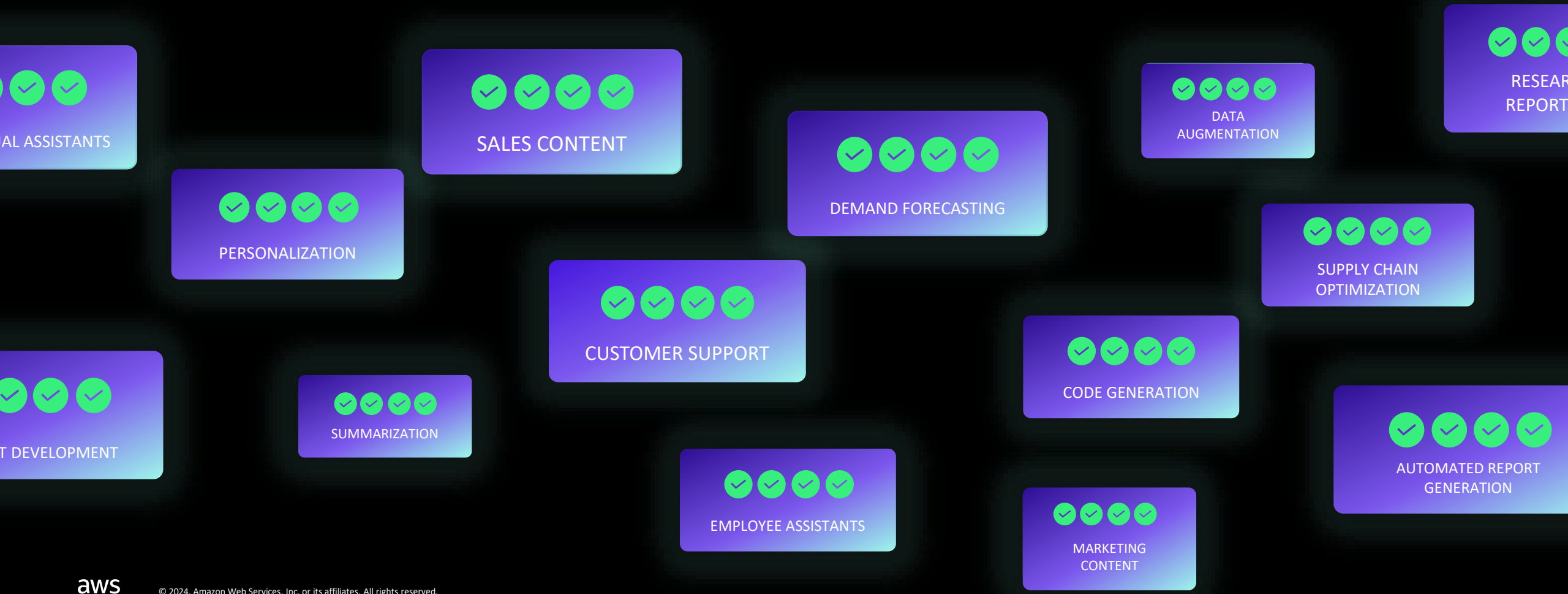
LAUNCH.
MANAGE.
REPEAT.

Where should we use generative AI?



How to select a pilot project & evaluate it?

- Business Impact
- Feasibility
- Risks
- Organizational readiness



It sounds complicated.
What do we need to start?

Leadership buy–in
From the top-down.

Data
Organized. Centralized. Complete and on the cloud.

Tools
To integrate, build, and scale.

Business Value
Measurable and tangible ROI.

Security & Privacy
Built-in. From the start.

Move

LEARN.
EXPLORE.
TEST.

Build

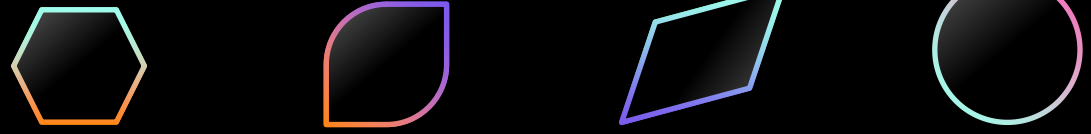
CUSTOMIZE.
INTEGRATE.
EVALUATE.

Scale

LAUNCH.
MANAGE.
REPEAT.

How do we know **which model(s) to use?**

Model specialities



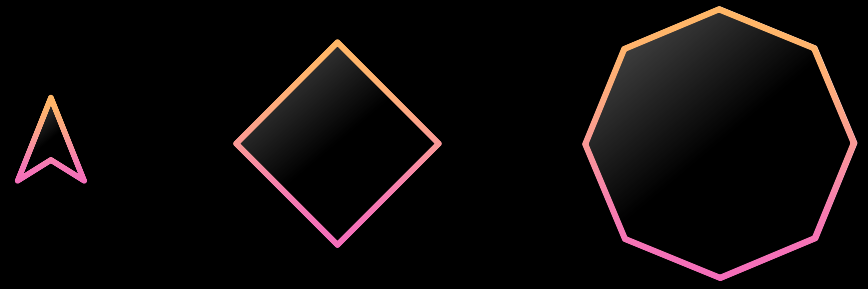
SUMMARIZATION

ANALYSIS

REASONING

CODE INTERPRETATION

Operational characteristics



FAST

LOW COST

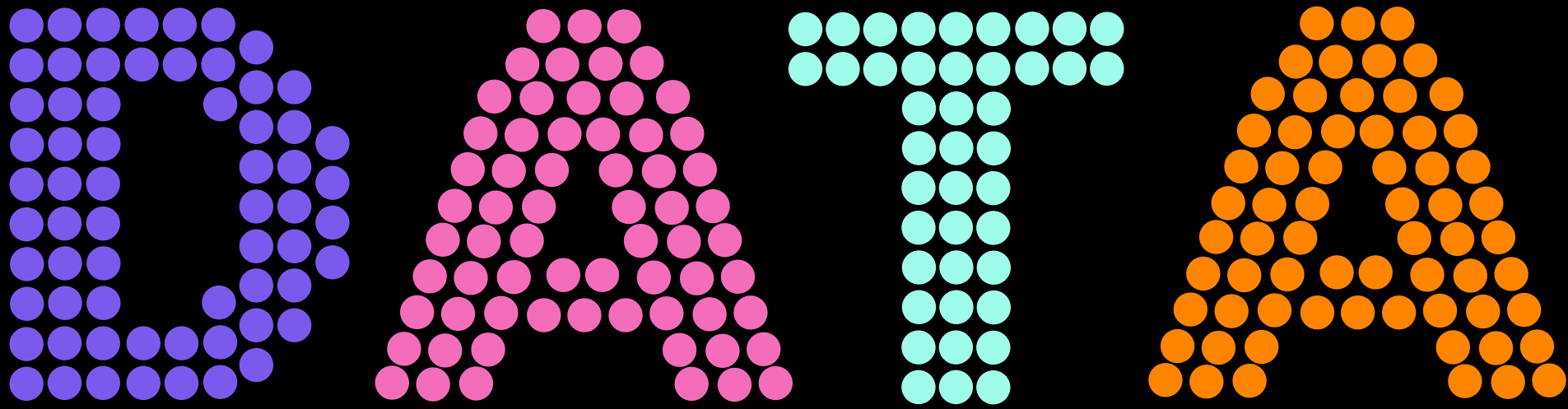
INTELLIGENT

How do we know **which model(s) to use?**

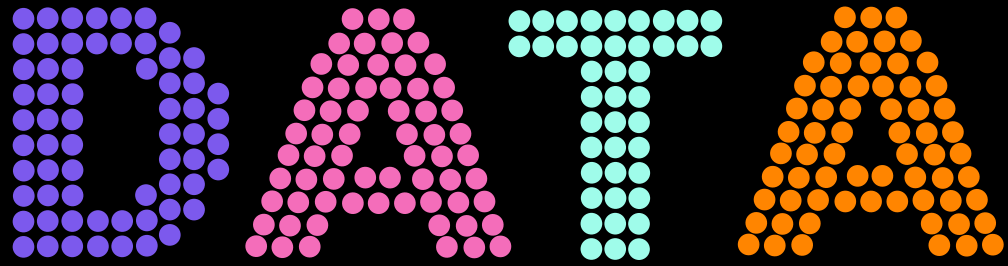
USE CASE **CUSTOMER SUPPORT**	USE CASE NEEDS	MODEL CHOICE
Summarizing & call categorization		
Direct customer chat engagements		
Multi-channel Engagement		

How do we **get the most value** out of generative AI?

How do we get the most value out of generative AI?



What does “good” data for generative AI look like?



Timely
Up-to-date reflecting the current situation

Complete
Contains all necessary information

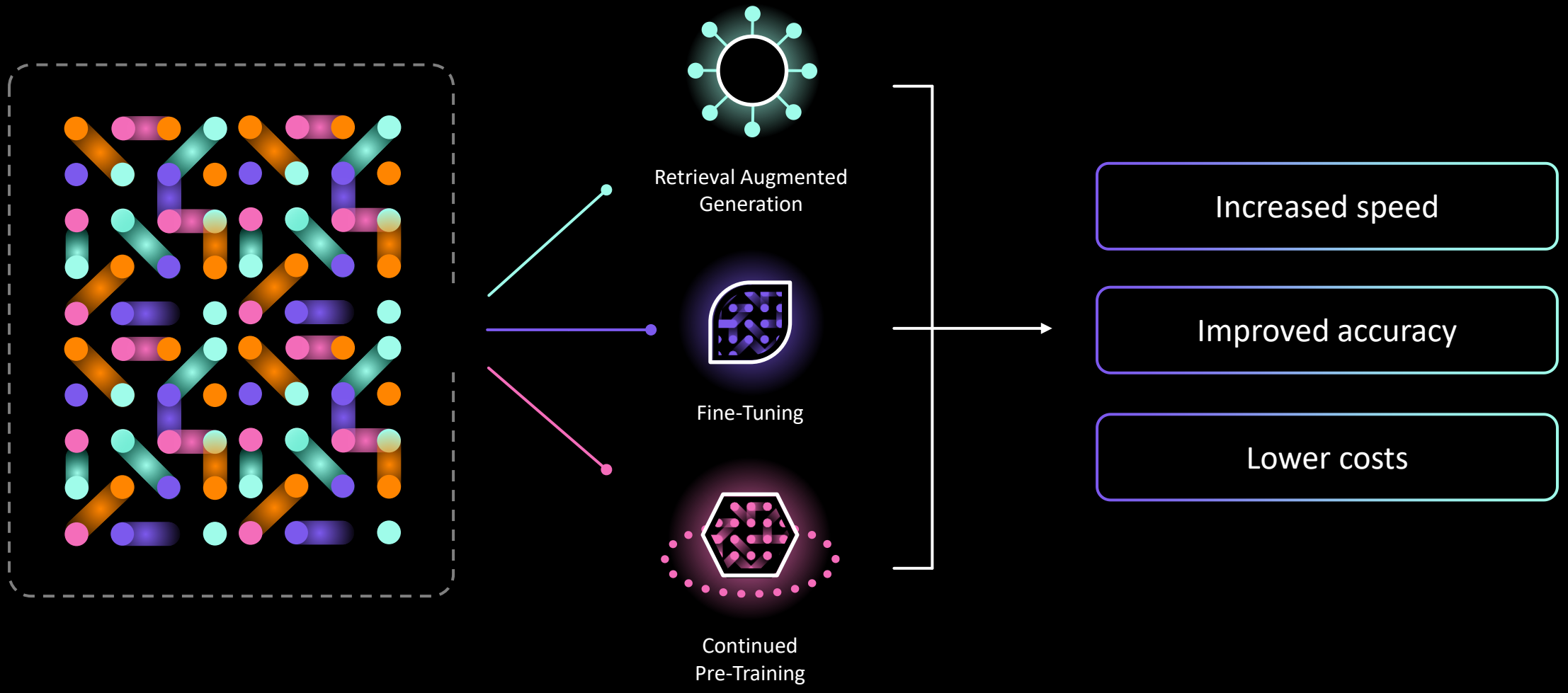
Unique
Each record is distinct with no duplicates

Accurate
Free from errors, from a verified source

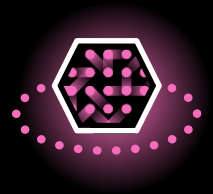
Governed
Well-managed, adhering to established rules

Traceable
Clear lineage across all processing steps

How can we **customize** with data?

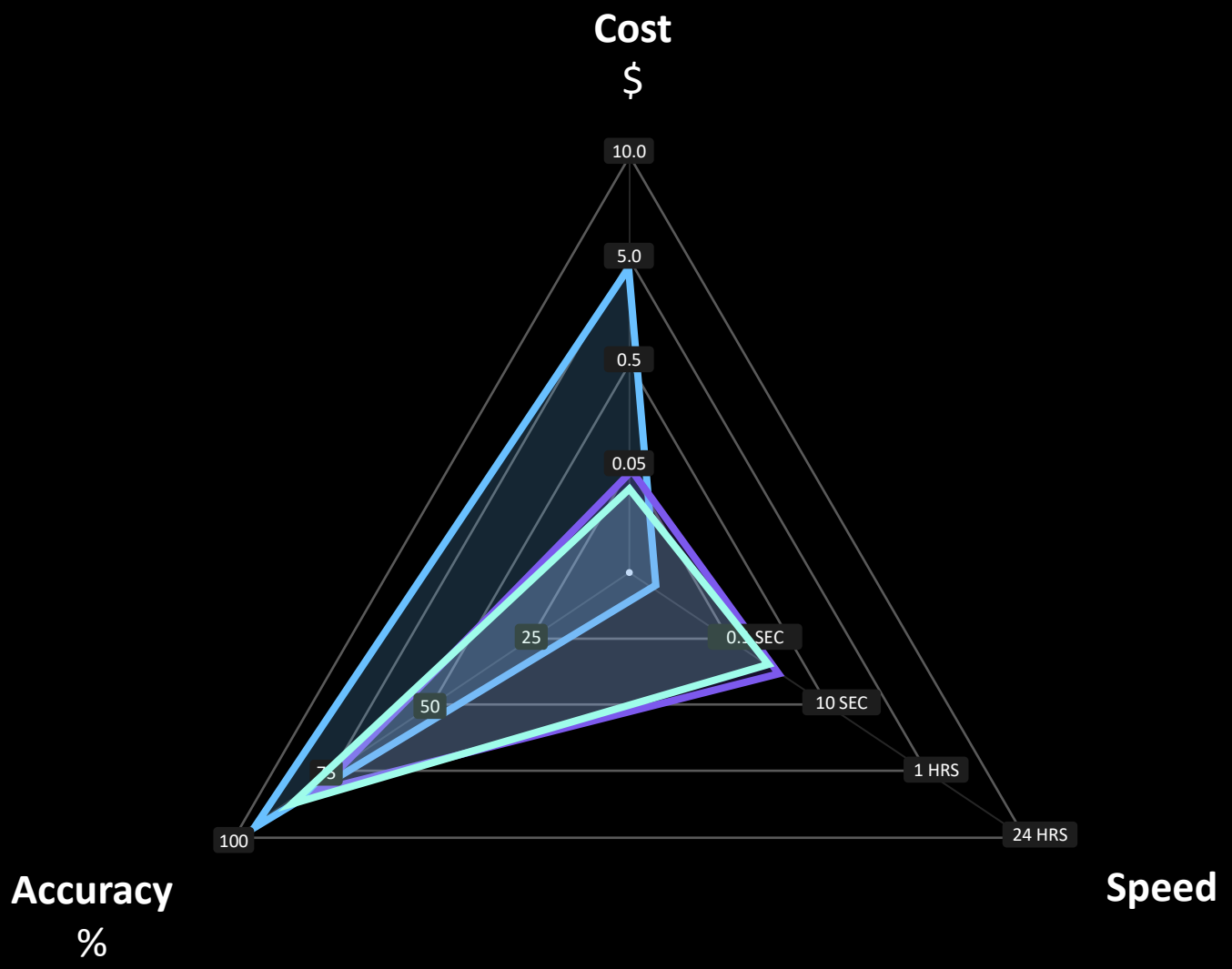


How do we know which customization to use?

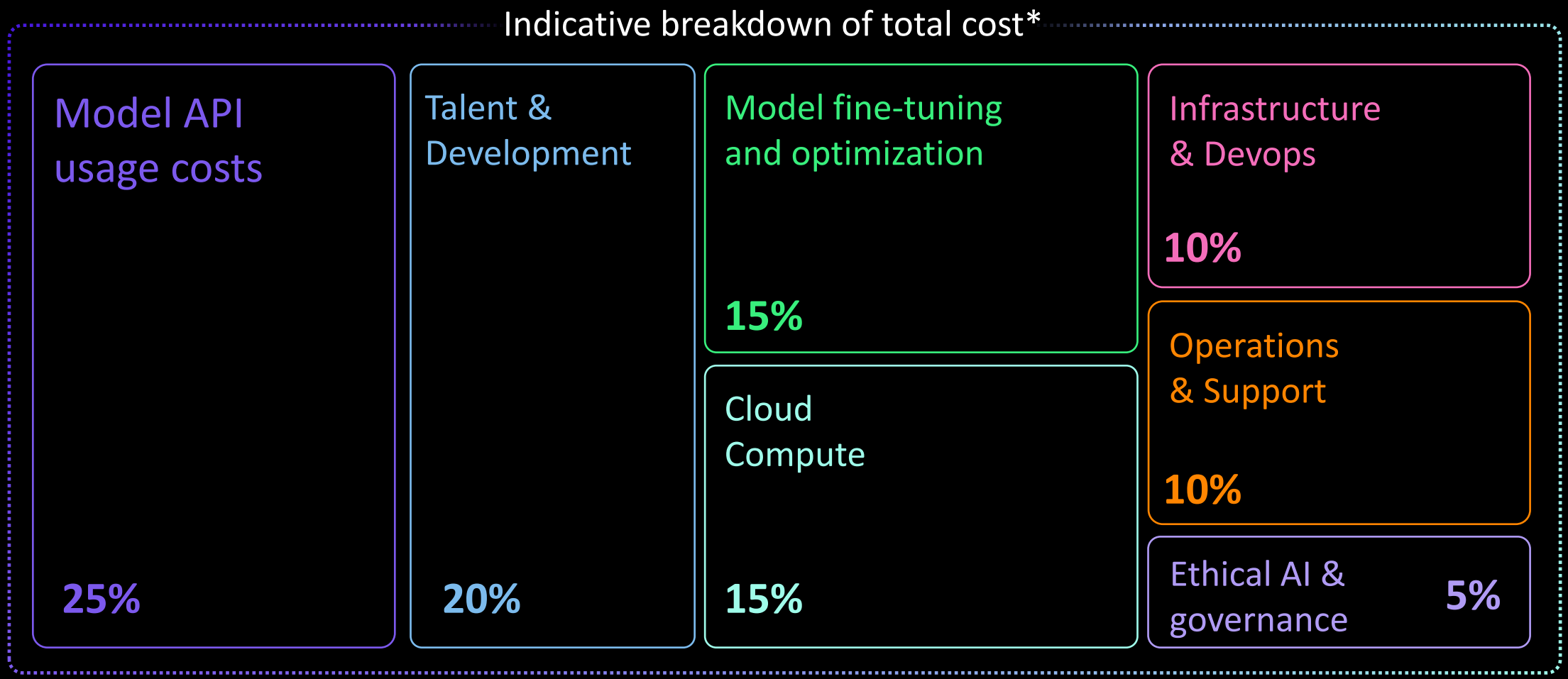
CUSTOMIZATION TYPE	USE CASE: CUSTOMER SUPPORT
 Prompt Engineering	Summarizing & call categorization
 Retrieval Augmented Generation	Sharing pre-determined data for customer inquiries
 Fine-tuning	Develop responses that align with your brand tone and voice. Match style of customer service transcripts for targeted responses.
 Continued Pre-training	Enhance general comprehension on your business

How do we consider **costs & ROI?**

- Goals
- Proof of concept
- Production

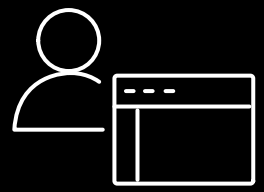


What is the total cost of ownership with generative AI?



*This example is for illustrative purposes only

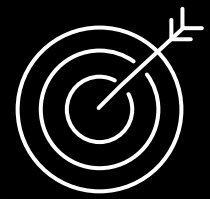
How can we make sure users **trust generative AI from the start?**



Include a “human in the loop”

Educate builders and users

Establish strong guardrails



Prioritize accuracy

Commit to ongoing testing & assessment

Move

LEARN.
EXPLORE.
TEST.

Build

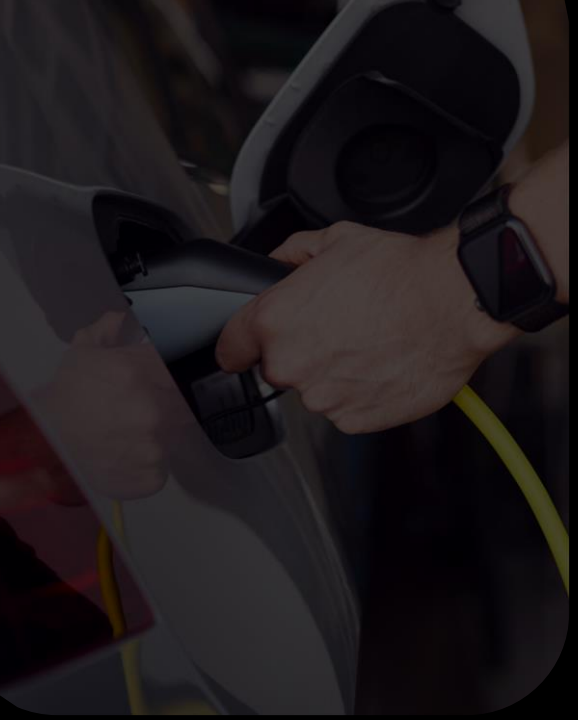
CUSTOMIZE.
INTEGRATE.
EVALUATE.

Scale

LAUNCH.
MANAGE.
REPEAT.

When should we move from **pilot to full production?**

- Data is secure & private
- Proven return-on-investment
- Stakeholder alignment
- Compliance & governance in place
- Responsible AI guardrails



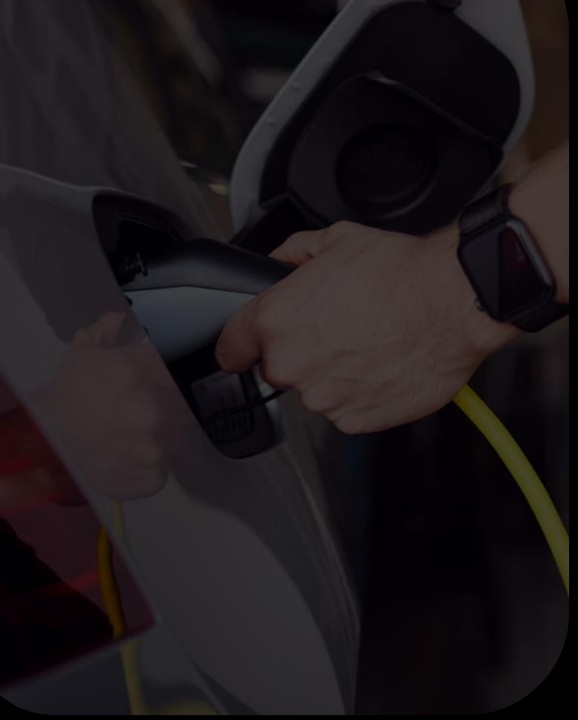
THOMSON REUTERS

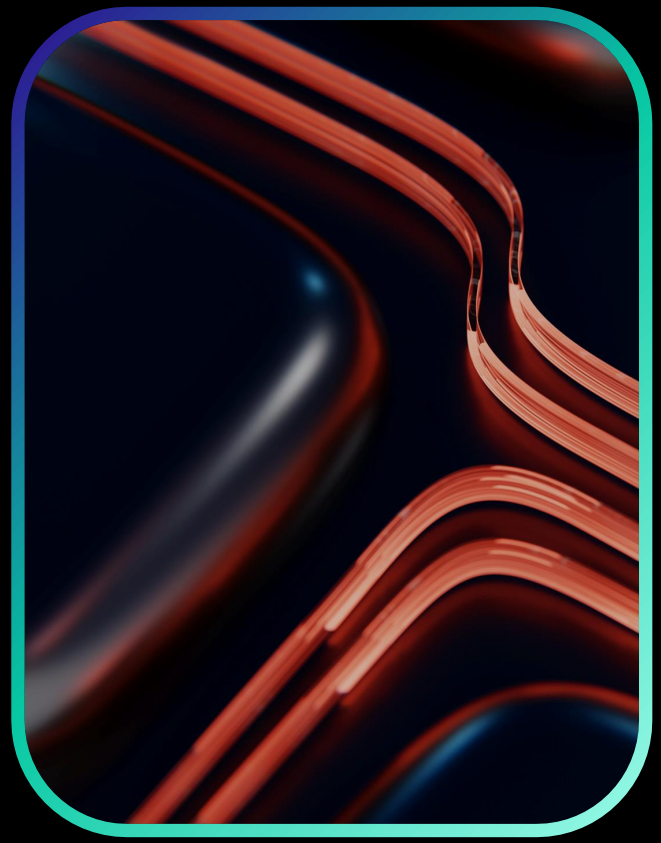
THOMSON REUTERS

ENABLING INNOVATION

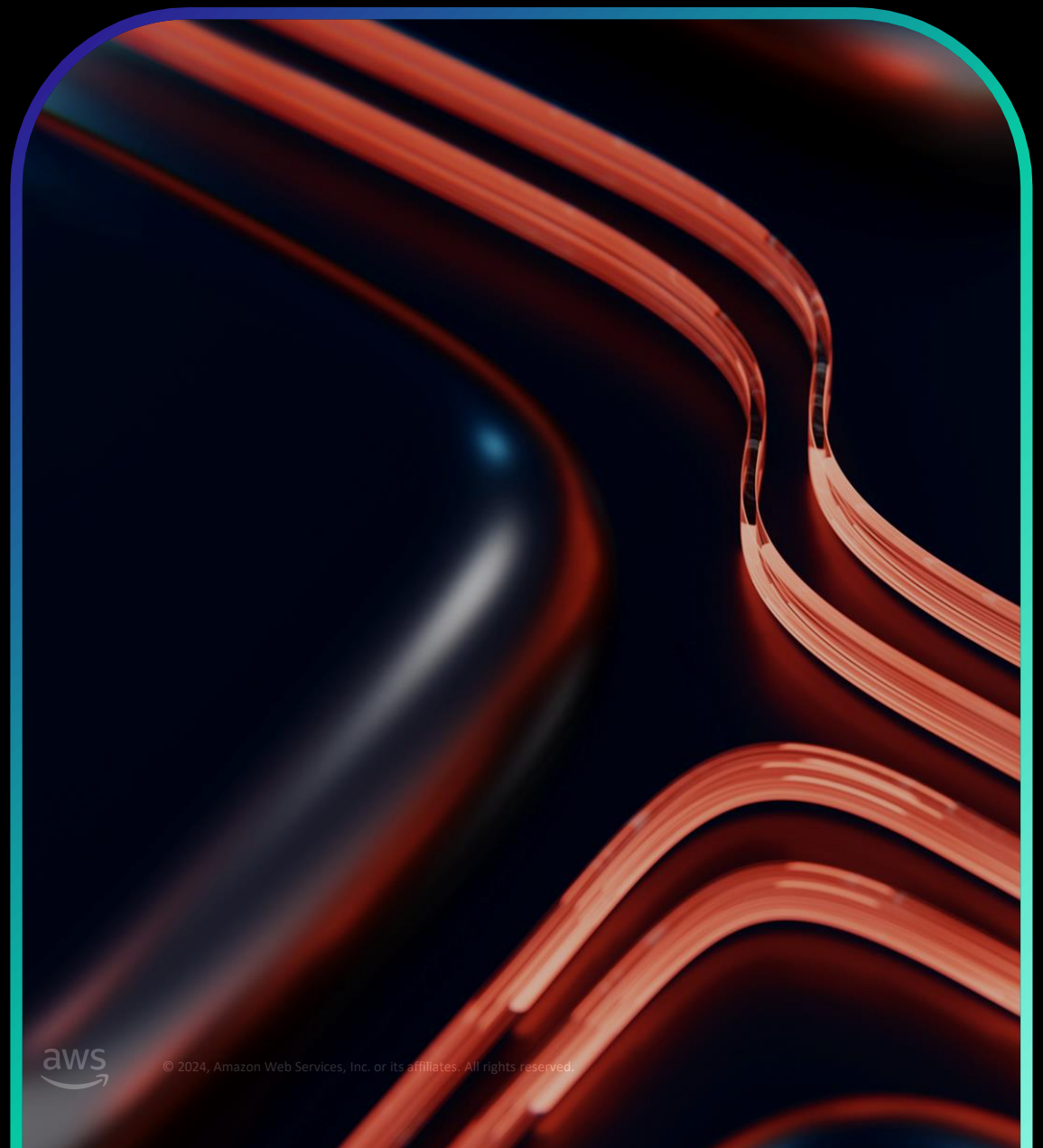
AMAZON BEDROCK

- Thomson Reuters launched its own LLM playground in under 6 weeks
- Leveraged Amazon Bedrock and Amazon SageMaker Jumpstart to build and enhance Open Arena’s capabilities
- Continuously improving Open Arena to keep pace with rapidly evolving generative AI landscape





Showpad

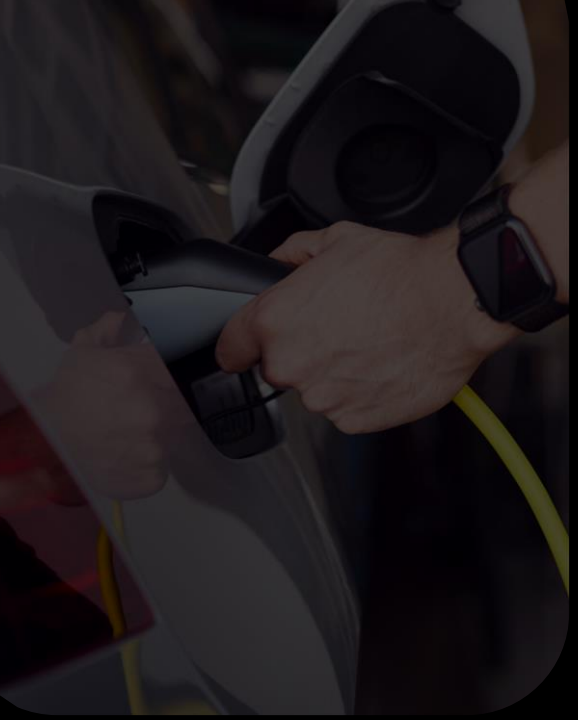


Showpad

BOOSTING PRODUCTIVITY

AMAZON BEDROCK

- Showpad leveraged Amazon Bedrock to securely train and run Anthropic's LLMs
- Scaled from early AI experimentation to 12 new AI features in just 1 year on Amazon Bedrock
- Boosted customer productivity with asset summaries, search, and message composing



DEMOCRATIZE DATA

AMAZON Q

- Empowered supply chain teams to derive actionable insights from complex data more efficiently with Amazon Q in QuickSight
- Build, discover, and share meaningful insights in seconds through natural language interactions

What does **the future** hold for generative AI?

Agents

Complex tasks.
End-to-end work flow automation.

Multi-modal

Any data types.
Audio, video, text, tabular

Multiple Models

Complex solutions.
Generative AI + Machine Learning.

AI Policies & Standards

Preparing for the future
Safety and responsibility at the forefront.

Move

LEARN.
EXPLORE.
TEST.

Build

CUSTOMIZE.
INTEGRATE.
EVALUATE.

Scale

LAUNCH.
MANAGE.
REPEAT.

Thank you!