BERLIN | MAY 15-16, 2024

aws SUMMIT

AIM205

Innovate faster with generative AI – Zalando

NUNO CASTRO

(he/him)
Sr Applied Science Manager
GenAI Innovation Center
AWS

DR. WEIWEI CHENG

(he/him)
Sr Principal Scientist
Zalando SE

Everything you need to accelerate your generative AI journey

Choice and flexibility of models

Differentiate with your data

Responsible AI integration

Low cost and performant infrastructure

Generative AI-powered applications

Internet-sized data

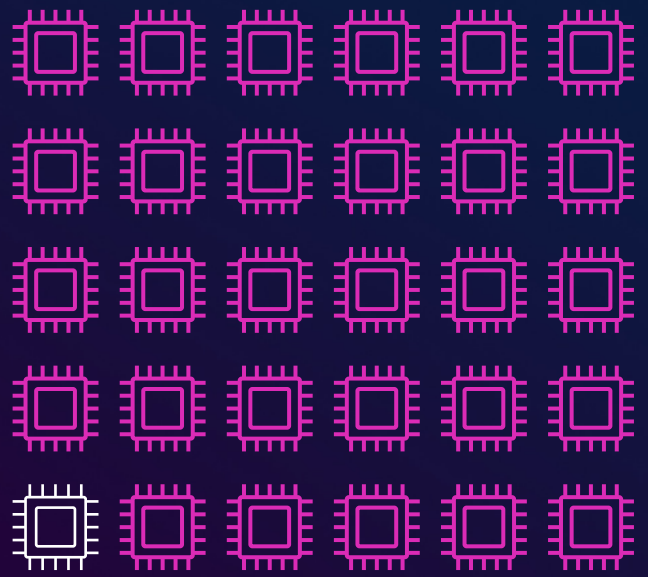
Scalable compute ML models

The cloud made generative AI possible

Six years of machine learning innovation

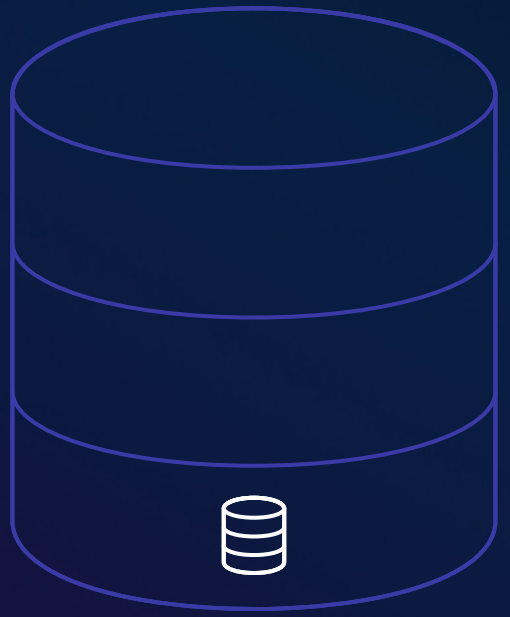
100,000X

Compute



100X

Data



1,000X

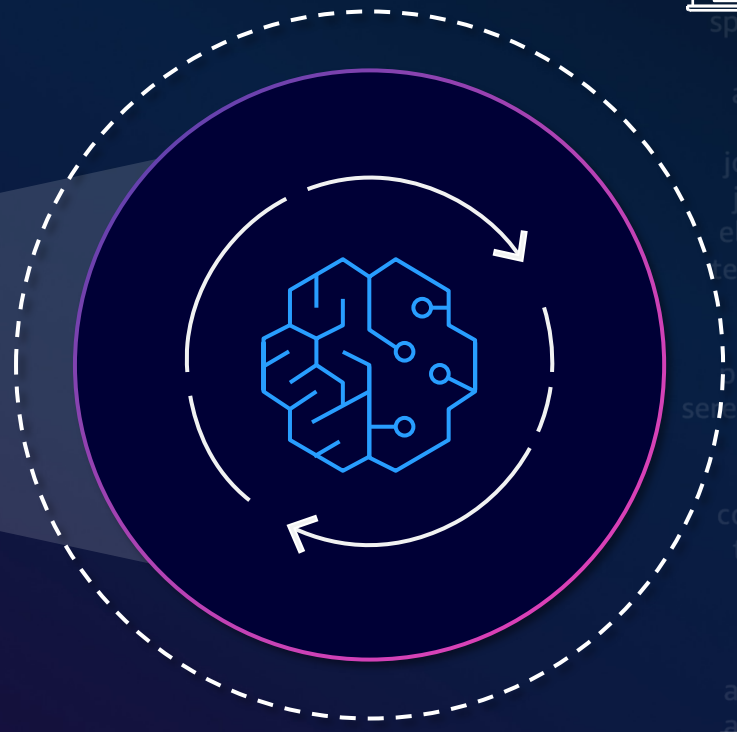
Model size

A HUMAN LIFETIME

A BILLION WORDS

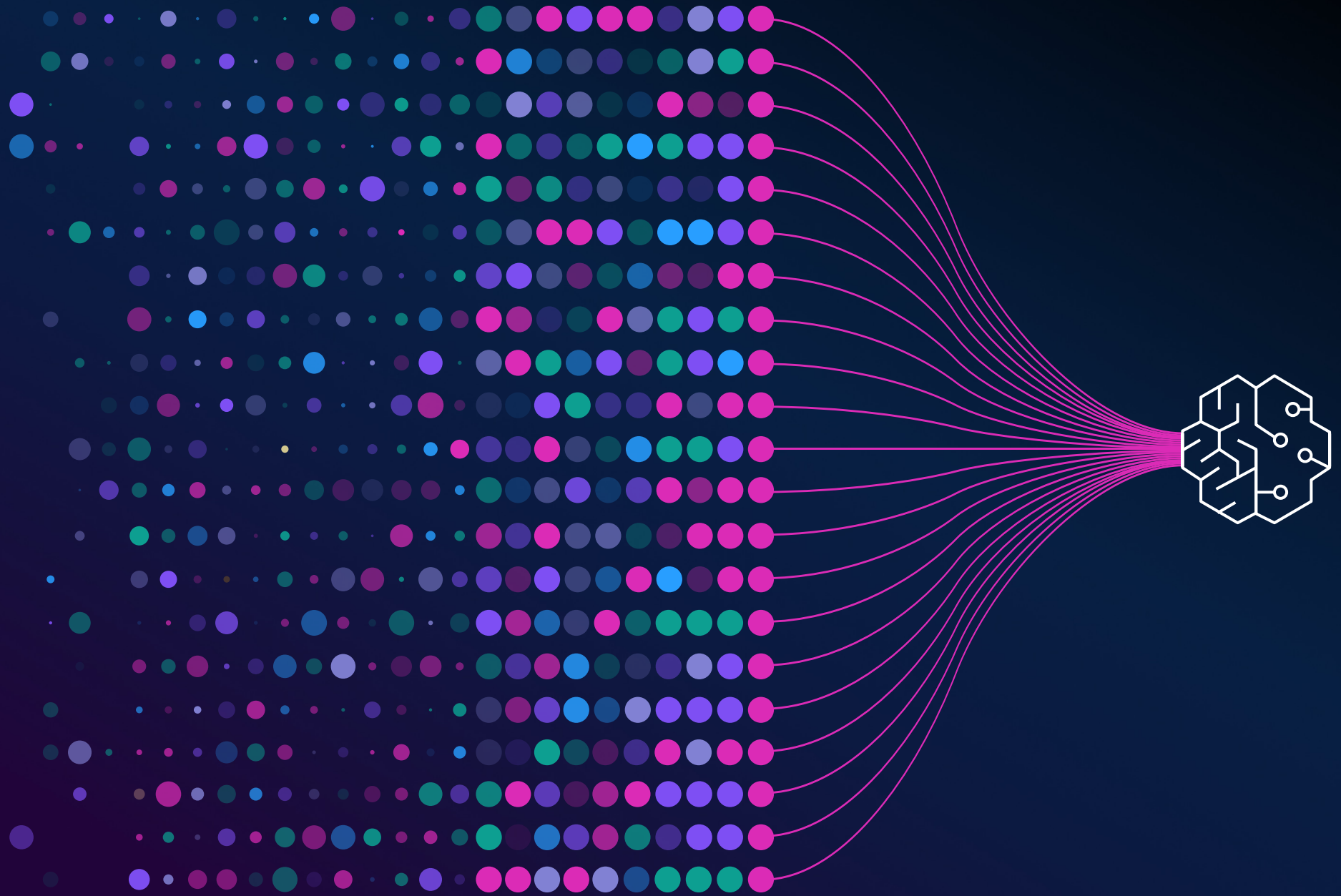
A HUMAN LIFETIME

A BILLION WORDS



FOUNDATION MODEL

TRILLIONS OF WORDS



Terabytes of data

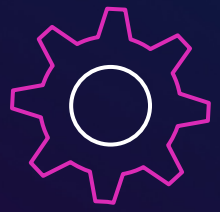
1,000s of times more information than available on Wikipedia

How do we put models to use?

Amazon Q

Your generative AI assistant designed for work that can be tailored to your business, data, code, and operations

PREVIEW



Personalizes interactions based on your role and permission

Understands your company information, code, and system

Built to be secure and private

Engages in conversations to solve problems, generate content, and take action

Where do we start?

What is your shoe return policy?

MODEL 1

Free returns within 30 days

EXAMPLE APPLICATION:
AD COPY

MODEL 2

Free returns within 30 days for purchases. There are no returns on damaged items or special orders unless you are a club member.

EXAMPLE APPLICATION:
CUSTOMER SERVICE ASSISTANT

What is your checked bag policy?

MODEL 1

Most flights allow 1 roller bag and 1 backpack and $30 to check a bag

MODEL 2

For domestic flights, you get one free roller bag measuring 22 inches long, 14 inches wide, and 9 inches high and two checked bags for $30 each. Frequent fliers get a 25% discount on all checked bags. International flights vary, and it's best to consult the website.

Cost effectiveness

Measure of the cost to host and invoke LLM

Completeness

Measure of completeness in the response and coverage of facts

Low hallucination

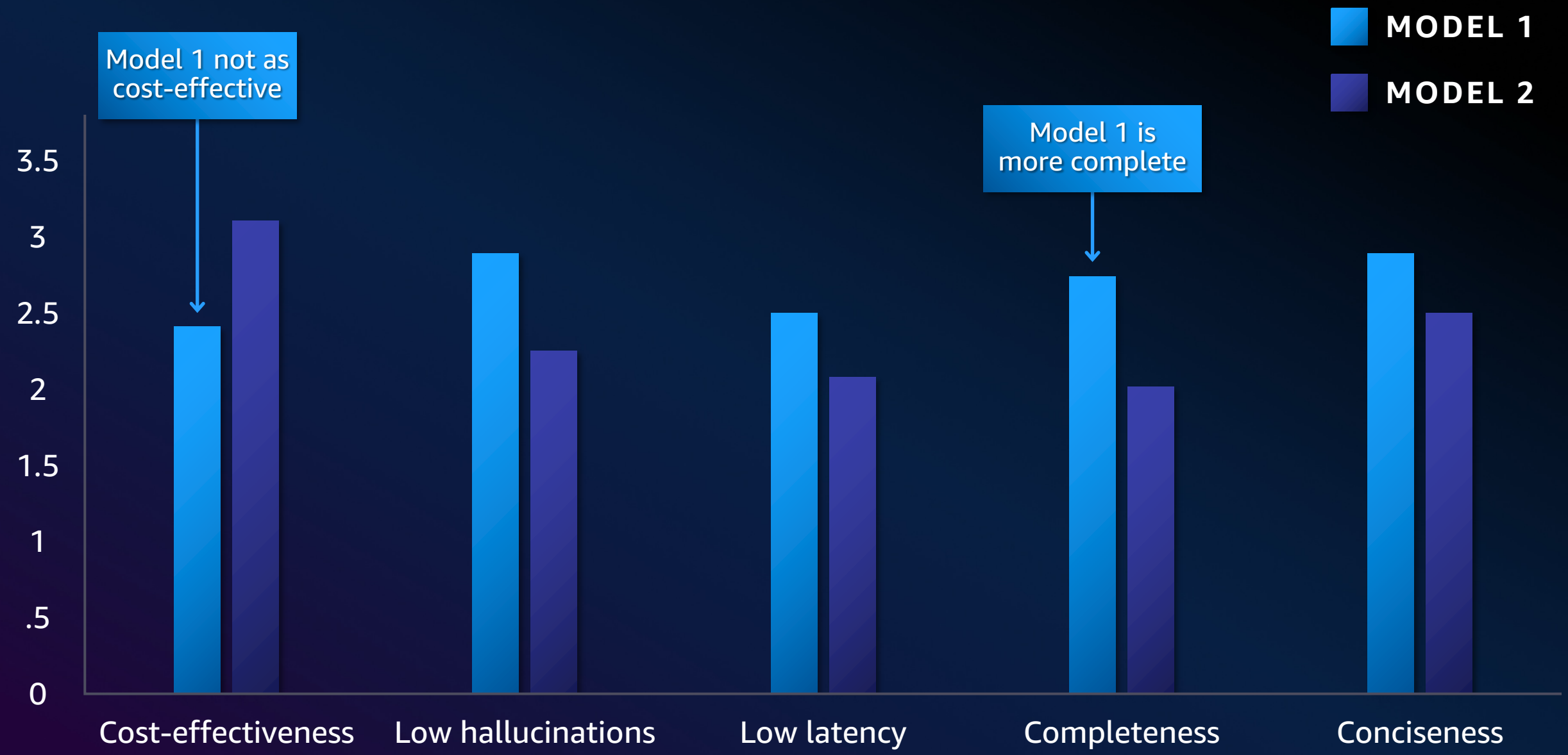
Measure of hallucination in responses and accuracy of citations and sources

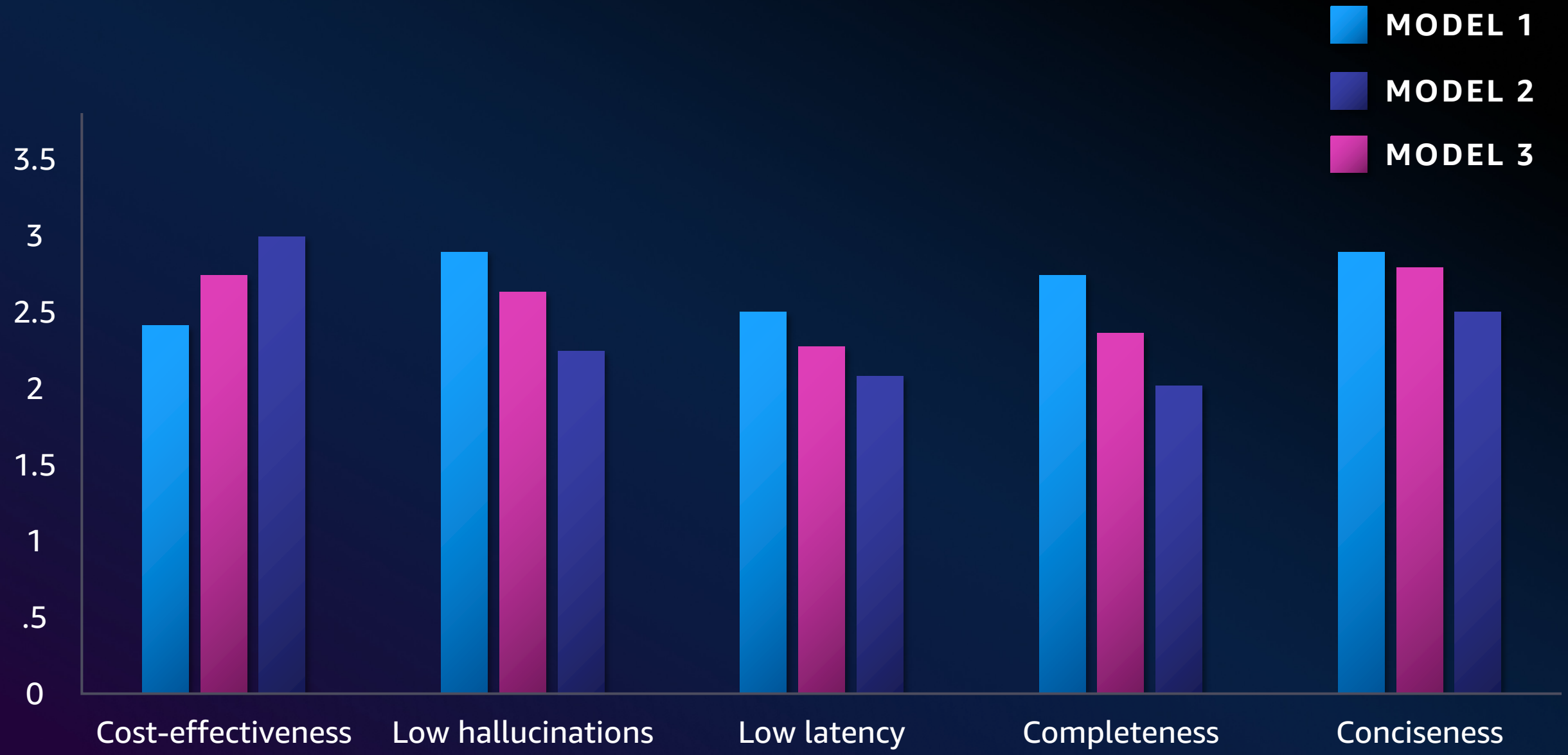
Conciseness

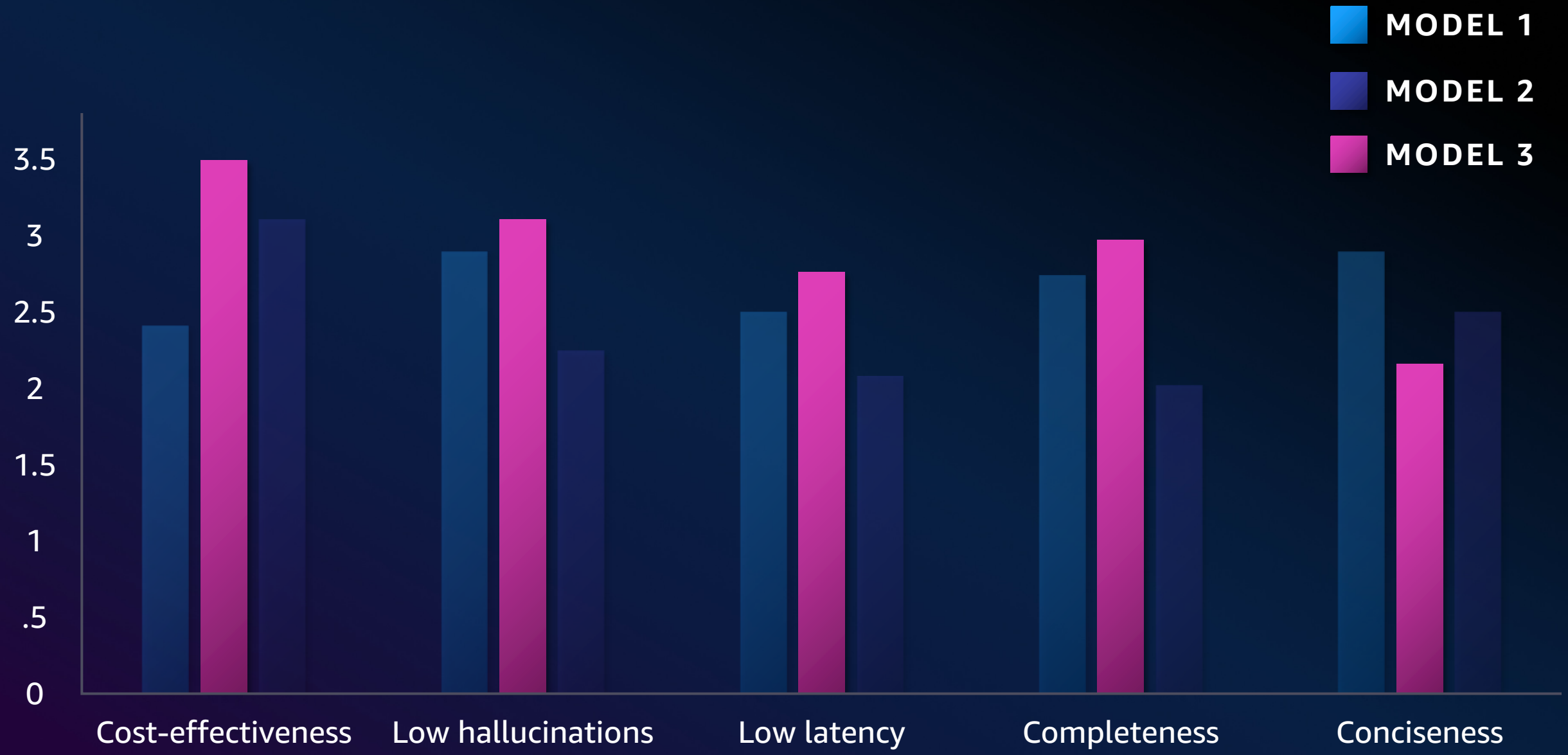
Measure of the conciseness in response

Low latency

Measure of time to first byte and full response







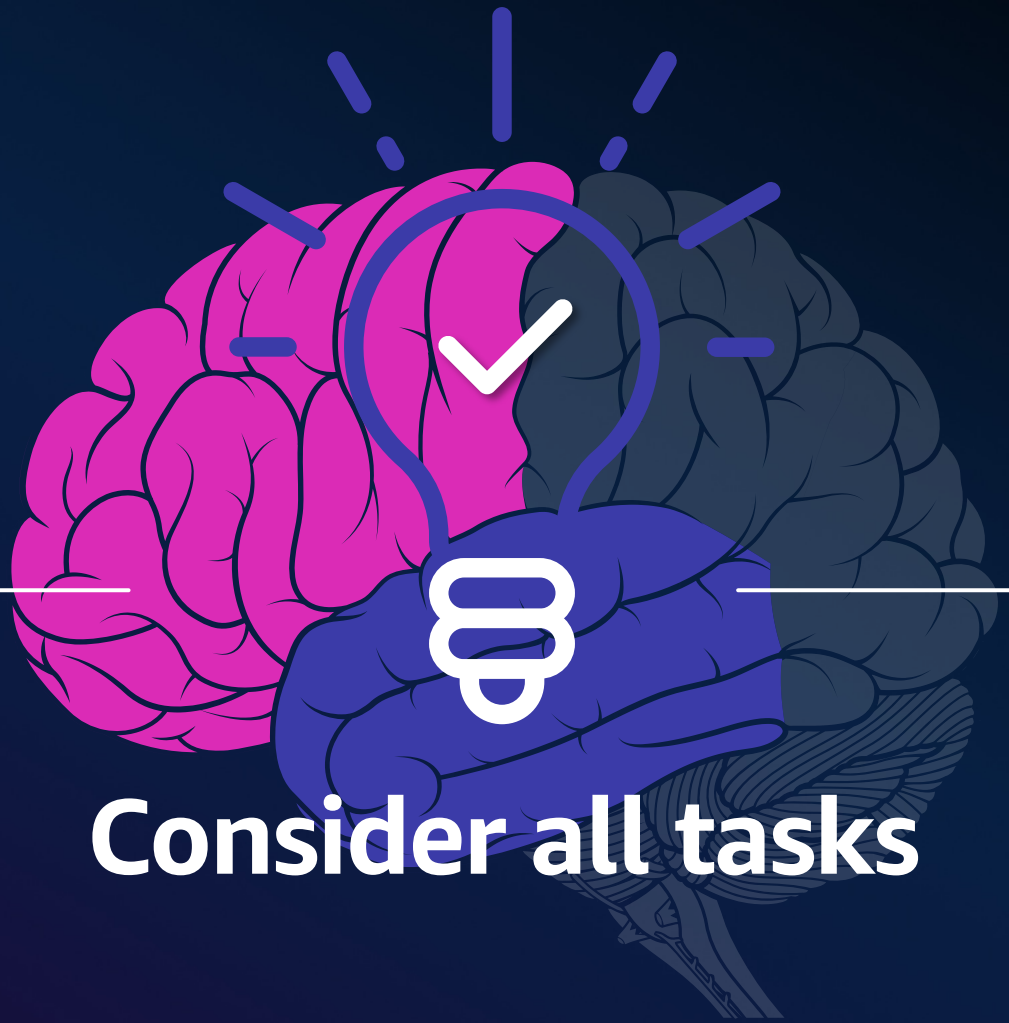
How did we optimize the system?

Generative AI application requirements

Frontal cortex

Reasoning

Logical thinking



Limbic system

Fast responses

Consider all tasks

How did we optimize the system?

Tell me about my customer meeting tomorrow at 10 am?

Your meeting is 1 hour long with top customer FabAI. FabAI recently launched a new offering and has 2 big deals in the pipeline.
The agenda will cover new pricing and support services.

Generative AI application requirements

Are we done yet?

Software engineer

What is the expected revenue of machine learning this quarter?

You aren’t authorized to view this information.

CEO of the company

What is the expected revenue of machine learning this quarter?

The average revenue is $123,456,789.

Generative AI application requirements

Models

- LLM
- Model 2
- ⋮
- Model N

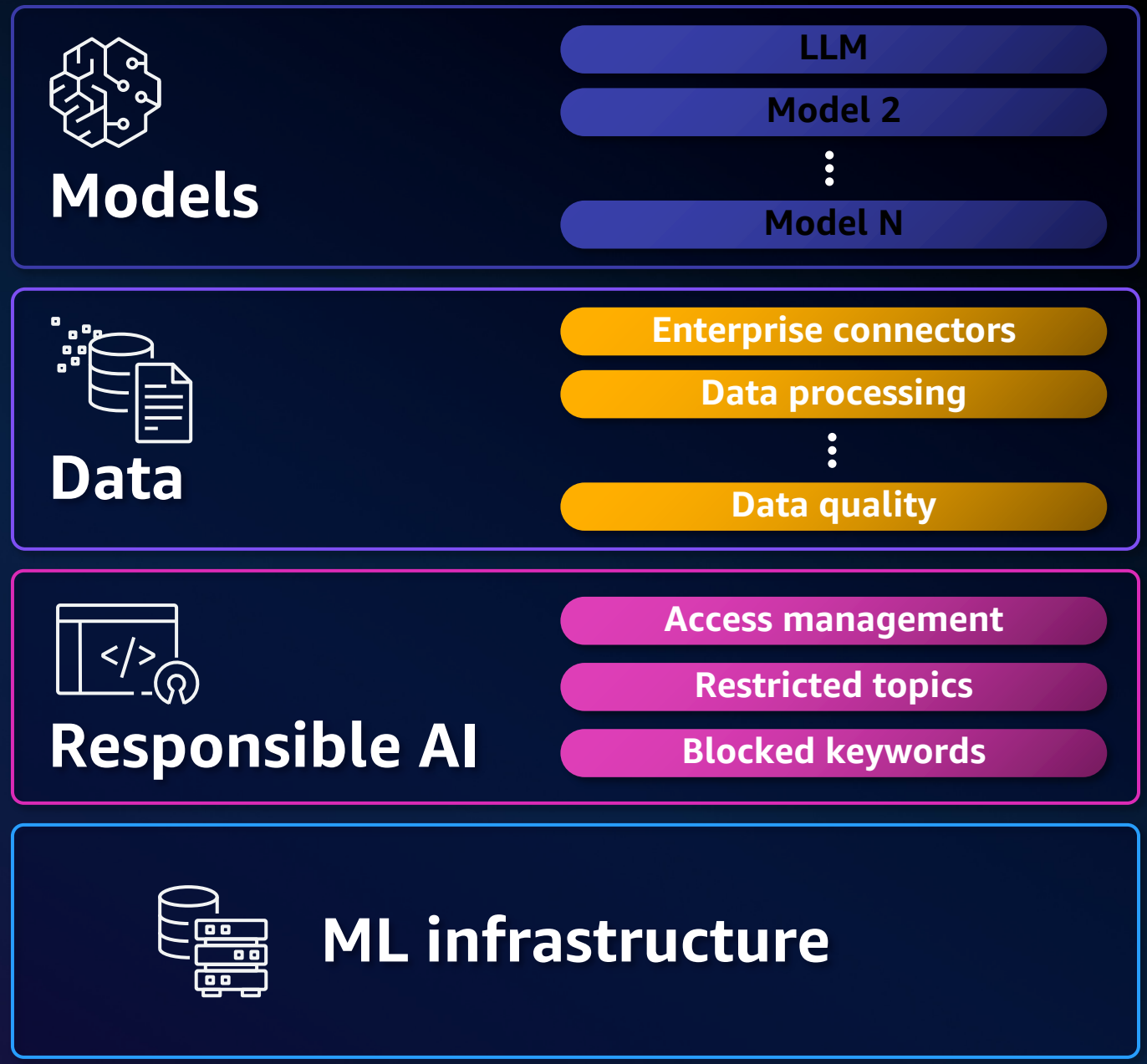
Data

- Enterprise connectors
- Data processing
- ⋮
- Data quality

Responsible AI

- Access management
- Restricted topics
- Blocked keywords

Generative AI application requirements



Everything you need to accelerate your generative AI journey

Choice and flexibility of models

Differentiate with your data

Responsible AI integration

Low cost and performant infrastructure

Generative AI-powered applications

Everything you need to accelerate your generative AI journey

Choice and flexibility of models

Differentiate with your data

Responsible AI integration

Low cost and performant infrastructure

Generative AI-powered applications

Amazon Bedrock

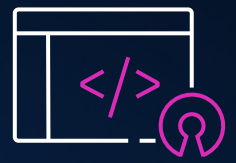
The easiest way to build and scale
generative AI applications with LLMs and other FMs

Model choice

Customization

Agents that
execute tasks

Security
and privacy



Responsible AI

Amazon Bedrock

Broad choice of models

Everything you need to accelerate your generative AI journey

Choice and flexibility of models

Differentiate with your data

Responsible AI integration

Low cost and performant infrastructure

Generative AI-powered applications

Foundation for an end-to-end data strategy

COMPREHENSIVE | INTEGRATED | GOVERNED

INTEGRATE

Data sources

- IOT/ DEVICES
- APP/ LOGS
- THIRD-PARTY DATA

Store, query, analyze

DATABASES

- Amazon Aurora
- Amazon DynamoDB

DATA WAREHOUSE

- Amazon Redshift

DATA LAKE

- Amazon S3

BIG DATA

- Amazon Kinesis & Amazon MSK
- Amazon EMR

Act

BUSINESS INTELLIGENCE

- Amazon QuickSight

MACHINE LEARNING

- Amazon SageMaker

GENERATIVE AI

- Amazon Bedrock

→ People

→ Apps

→ Devices

CATALOG AND GOVERN | AWS Lake Formation, Amazon DataZone

Everything you need to accelerate your generative AI journey

Choice and flexibility of models

Differentiate with your data

Responsible AI integration

Low cost and performant infrastructure

Generative AI-powered applications

Amazon Bedrock keeps data secure and private

None of the customer's data is used to train the underlying model

All data is encrypted in transit and at rest

Data used to customize models remains within your VPC

Support for standards, including GDPR & HIPAA

Guardrails for Amazon Bedrock

Implement safeguards customized to your application requirements and responsible AI policies

GENERALLY AVAILABLE

Apply Guardrails to multiple foundation models and Agents for Amazon Bedrock

Configure harmful content filtering based on your responsible AI policies

Define and disallow denied topics with short natural language descriptions

Redact of block sensitive information such as PIIs, and custom Regex

Everything you need to accelerate your generative AI journey

Choice and flexibility of models

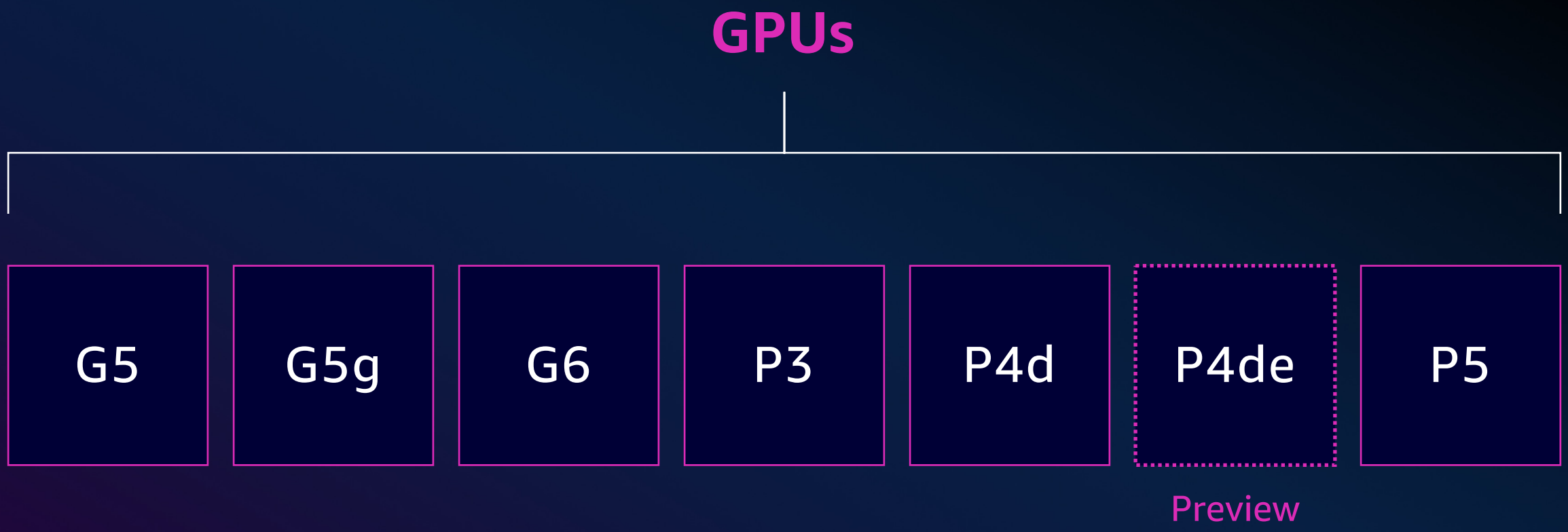
Differentiate with your data

Responsible AI integration

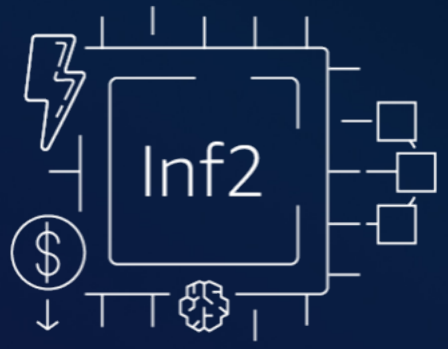
Low cost and performant infrastructure

Generative AI-powered applications

Choice of accelerated compute



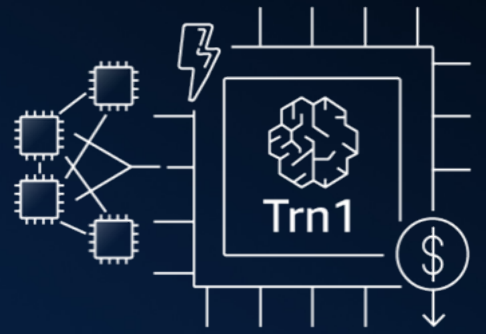
Purpose-built accelerators for generative AI



AWS Inferentia2

Up to 40% better price performance

than comparable Amazon EC2 instances



AWS Trainium

Up to 50% savings on training costs

over comparable Amazon EC2 instances

GENERALLY AVAILABLE

Amazon SageMaker HyperPod

Reduces time to train foundation models by up to 40%

Streamlined distributed training for large training clusters

Resilient training environment that eliminates interruptions

Optimized utilization of cluster's compute, memory, and network resources

Everything you need to accelerate your generative AI journey

Choice and flexibility of models

Differentiate with your data

Responsible AI integration

Low cost and performant infrastructure

Generative AI-powered applications

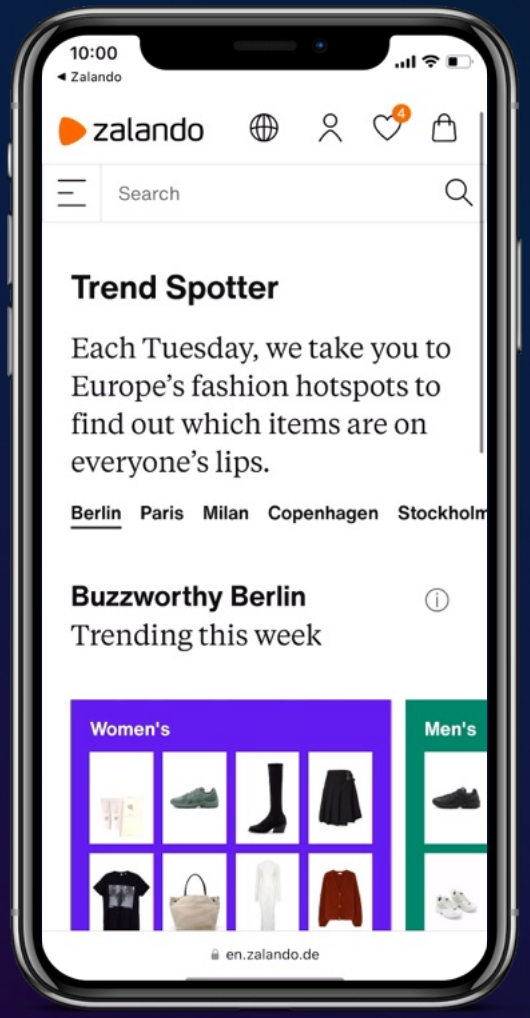
Dr. Weiwei Cheng

Fashion & Lifestyle Goods

25 European Markets

50 Million Active Customers

Diverse Product Offerings








long trench coat

Product Details

Material & Care:

Outer fabric material: 97% cotton, 3% elastane
Padding type: No lining
Care instructions: Hand wash only, Dry cleanable

Details:

Collar: Lapel collar
Fastening: Button
Pockets: Inseam pockets

Size & Fit

Fit: Regular fit
Shape: Fitted
Sleeve length: 84.45 cm
Total length: Size 40

Detailed product information allow customers to make the right purchase decisions

Metadata

Raw description

Metadata Extraction

Metadata

en

de

es

Material & Care:
Outer fabric material: 97% cotton, 3% elastane
Care instructions: Hand wash only, Dry cleanable

Material & Pflegehinweise:
Material Oberstoff: 97% Baumwolle, 3% Elasthan
Pflegehinweise: Handwäsche, chemische Reinigung möglich

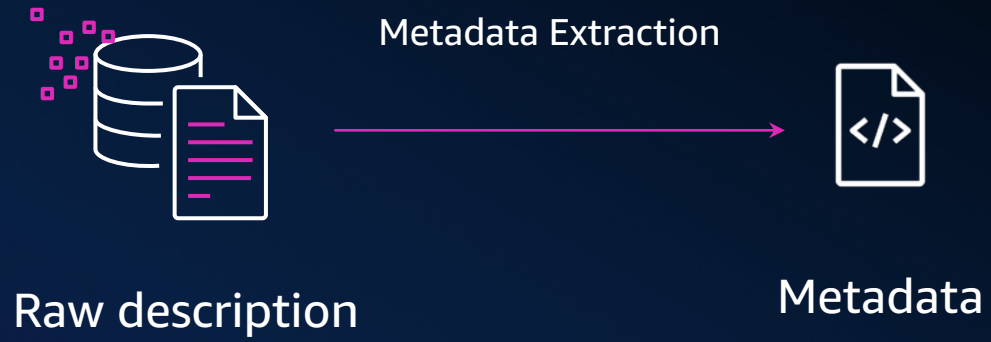
Composición y cuidados:
Material exterior: 97% algodón, 3% elastano
Cuidados: Lavar a mano, se puede lavar en seco

Product details

Challenges

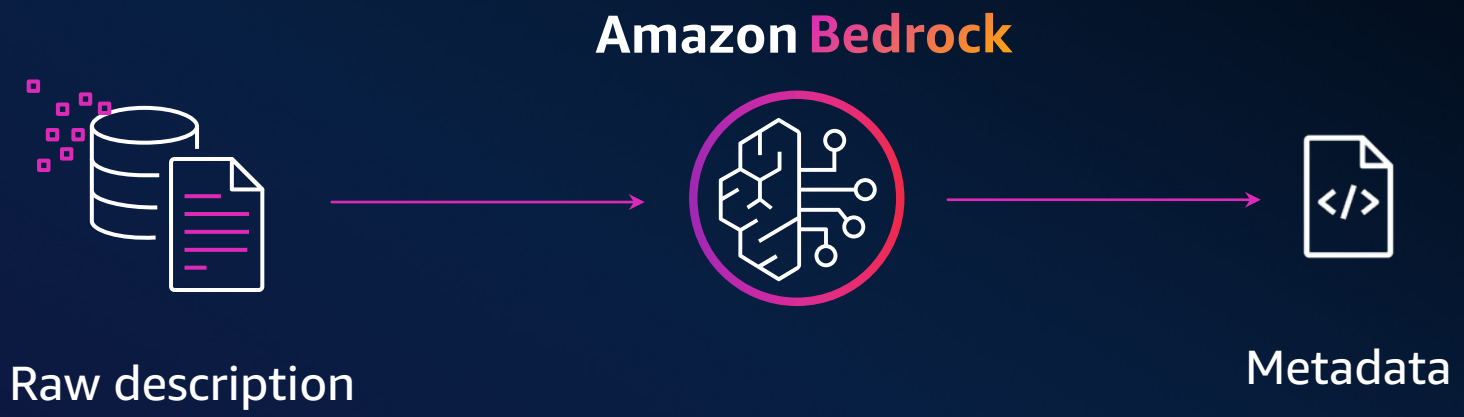
Product descriptions come from different data sources

- The descriptions use various languages
- The data structure is unique to every data provider
- The naming conventions vary from one manufacturer to the other



Metadata extraction was partially done manually

Better Metadata with GenAI



Use GenAI to Automate

Impact on Search

Detailed metadata leads to more accurate search results

Easier to find → Easier to buy

Improved overall shopping experience

Impact on Personalization

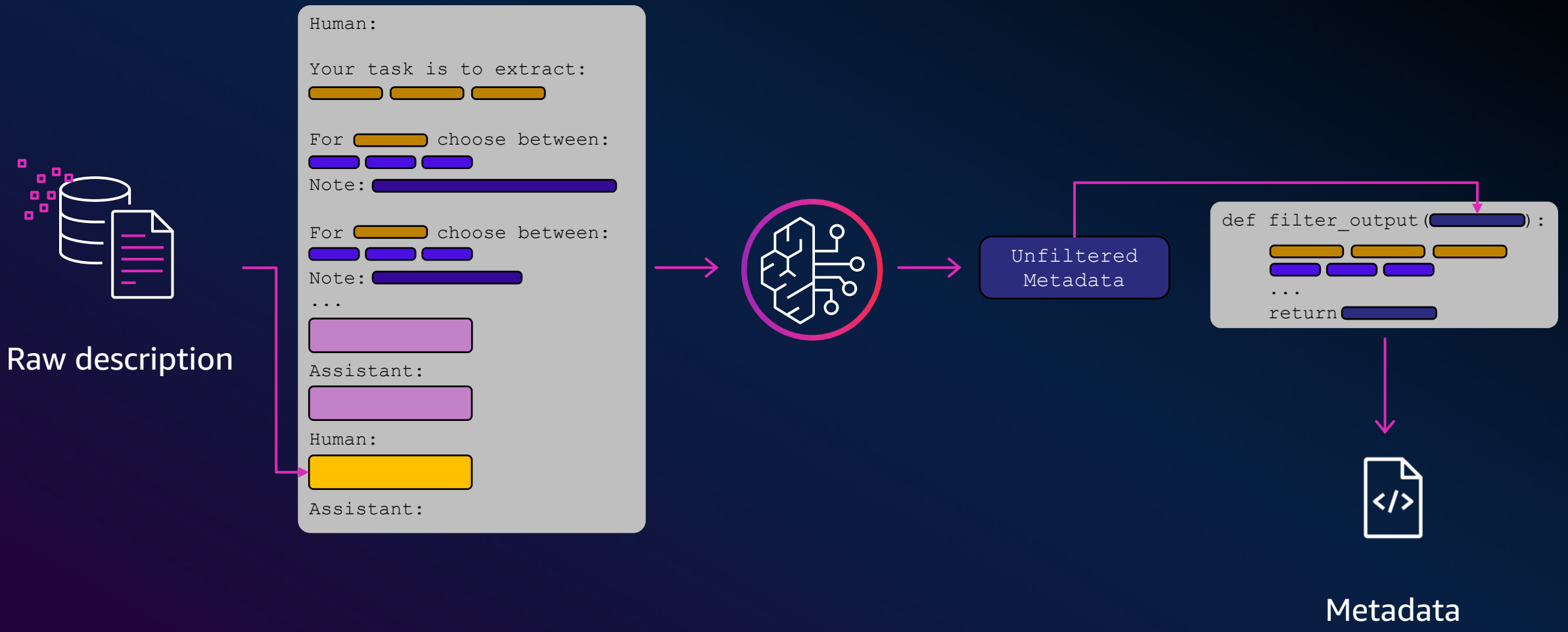
Size & Fit
Our model's height: Our model is 6'1" tall and is wearing size M
Fit: Regular fit
Shape: Straight
Sleeve length: 25.5" (Size M)
Total length: 27.0" (Size M)

Accurate metadata will improve personalization

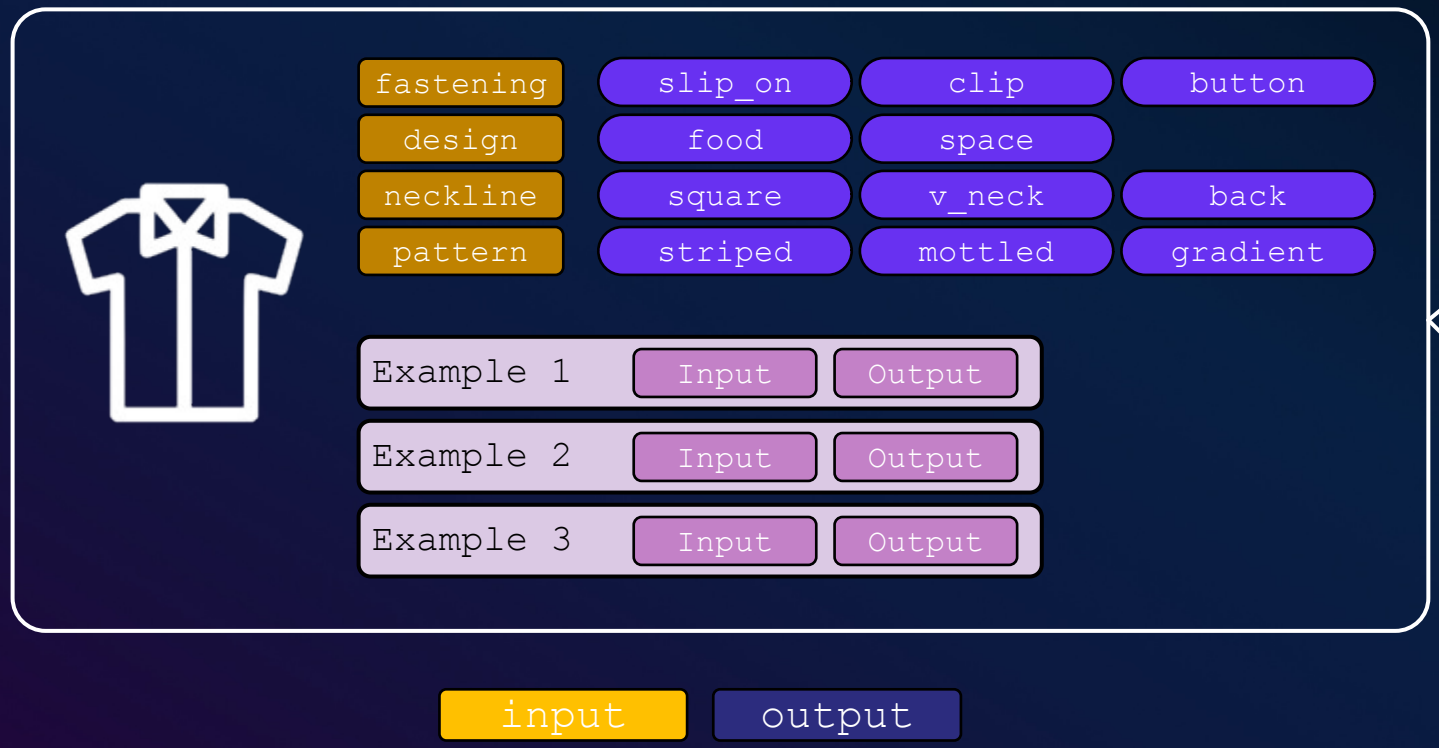
Better understanding of how it would fit will reduce returns

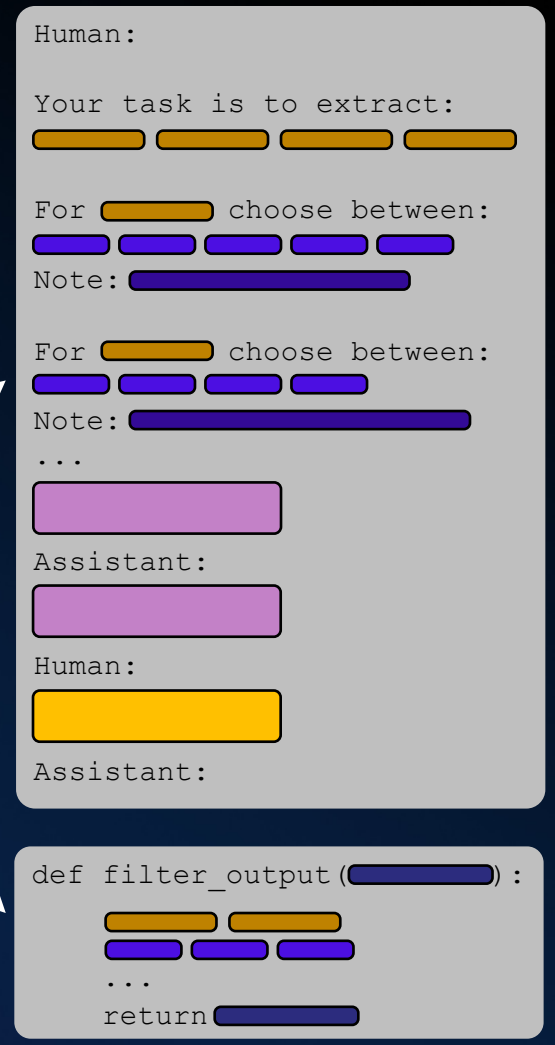
Less repackaging, less pollution, faster delivery, lower costs

How it works



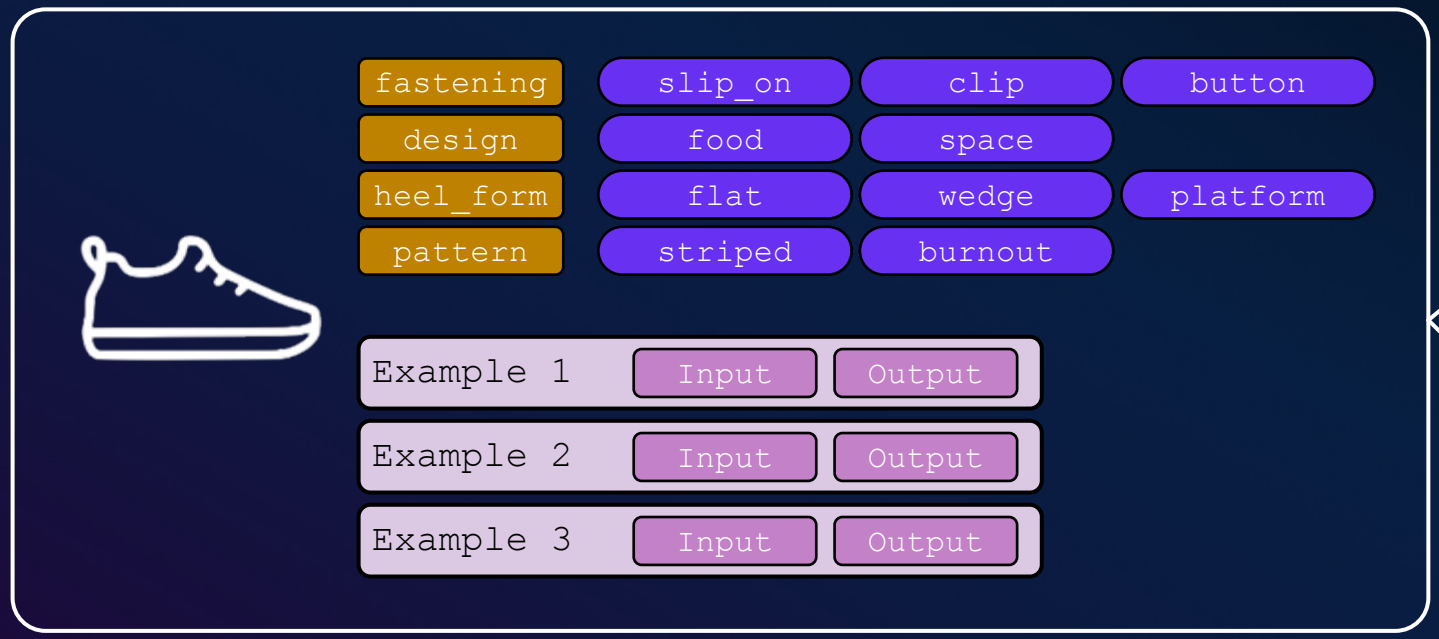
Prompts and filters

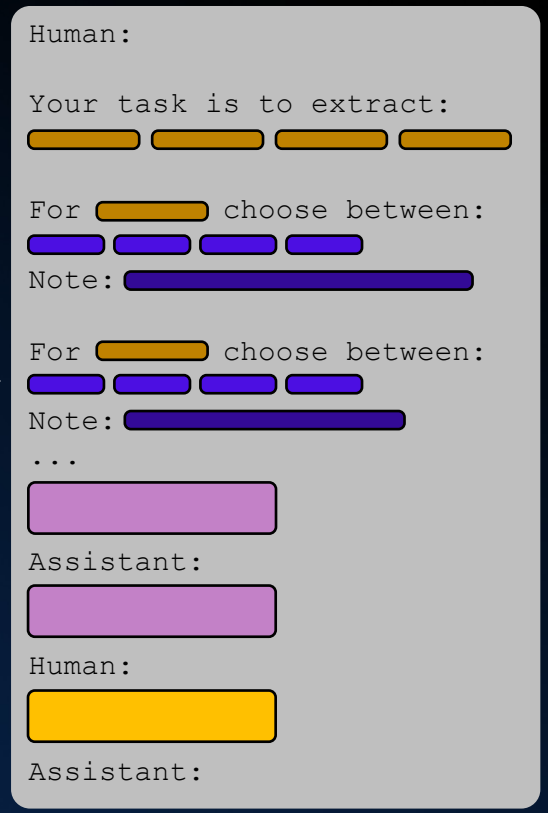


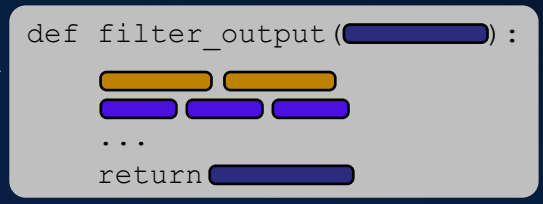


Given a product description and the related taxonomy, we create a custom prompt and a filter function

Prompts and filters



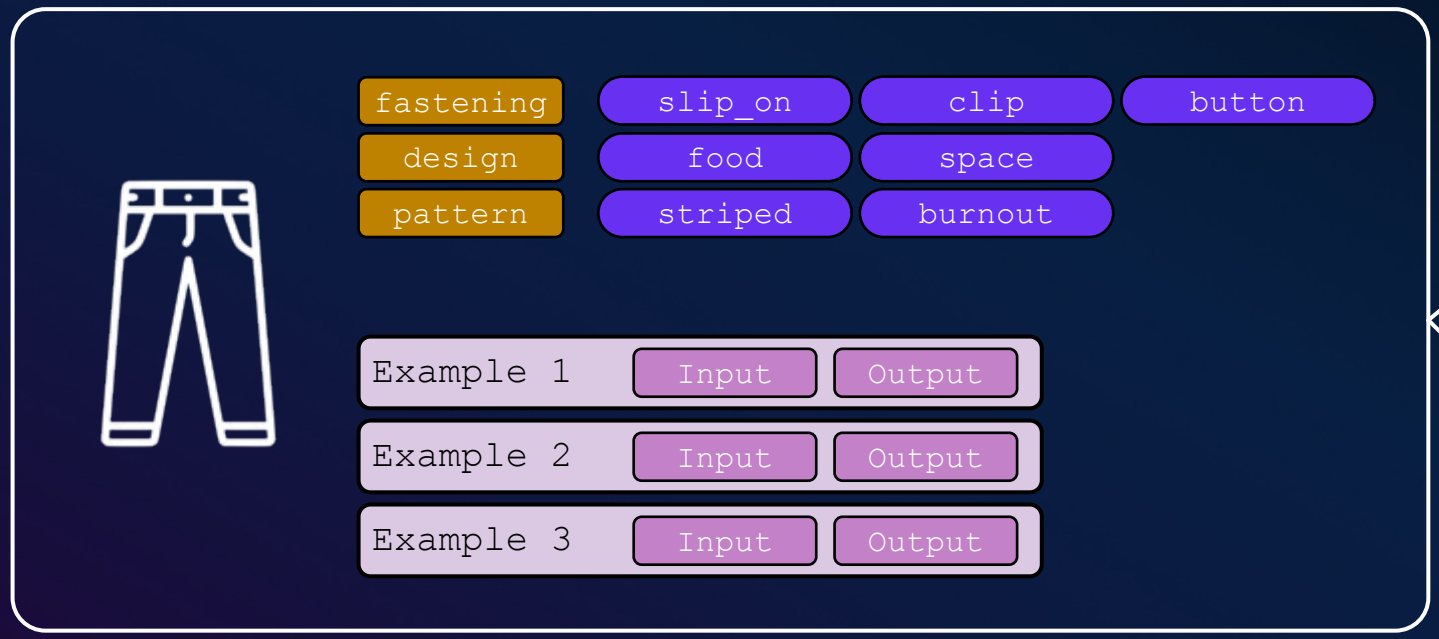




input output

Given a product description and the related taxonomy, we create a custom prompt and a filter function

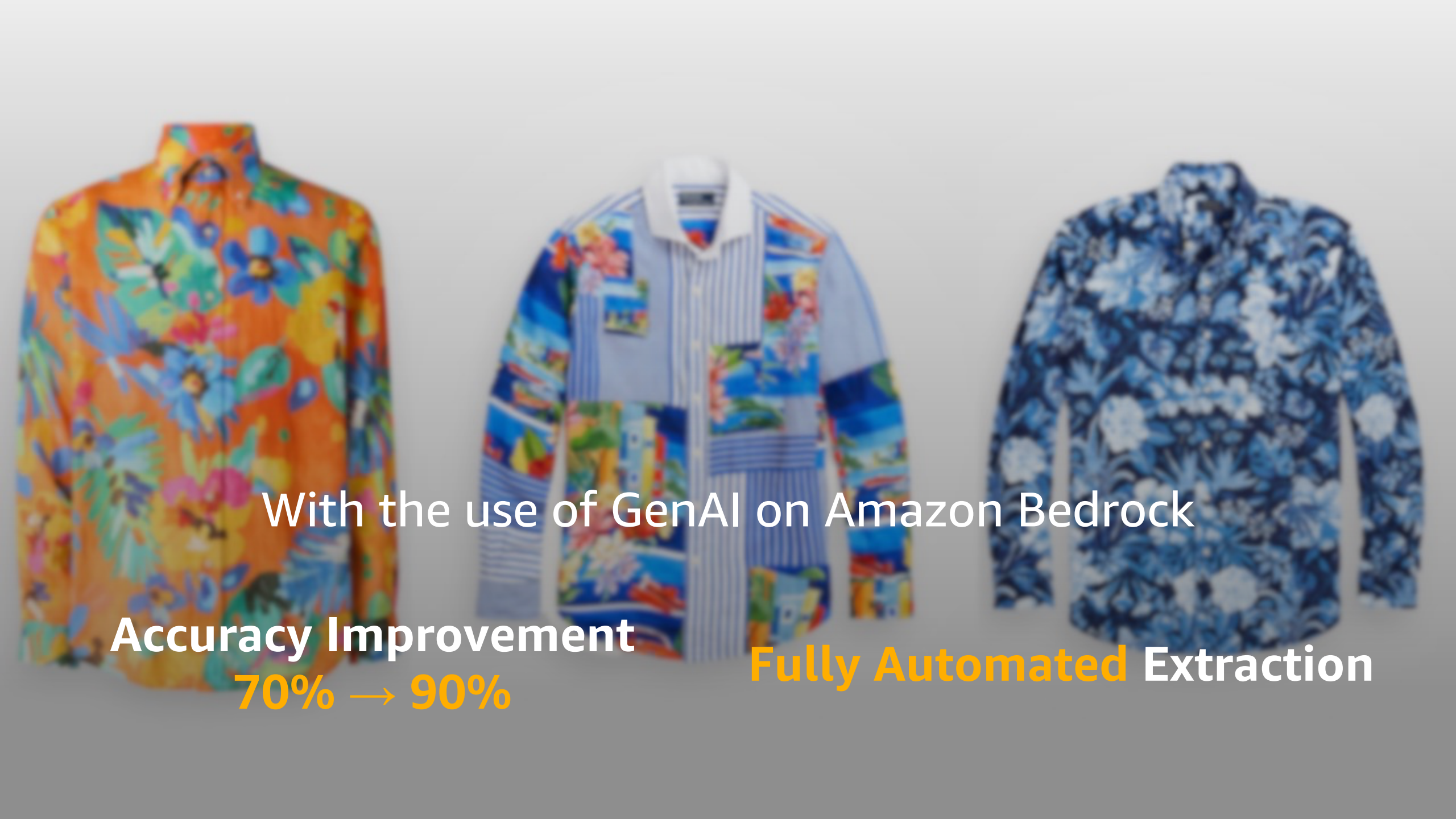
Prompts and filters




```
Human:

Your task is to extract:


For          choose between:


Note:

For          choose between:


Note:
...

Assistant:

Human:

Assistant:
```

```
def filter_output(          ):
    ...
    return
```

The prompt and the filter functions differ for every silhouette


With the use of GenAI on Amazon Bedrock
Accuracy Improvement
70% → 90%
Fully Automated Extraction

Dr. Weiwei Cheng

Everything you need to accelerate your generative AI journey

Choice and flexibility of models

Differentiate with your data

Responsible AI integration

Low cost and performant infrastructure

Generative AI-powered applications

skillbuilder.aws

Build beyond

Redeem your free 7-day trial of AWS Skill Builder

Thank you!

Please complete the session survey in the mobile app

Nuno Castro

Dr. Weiwei Cheng