Responsible AI in the generative AI era

Science and practice

Neelam Koshiya, Principal SA,GenAI
nkkoshiyr@amazon.com
She/Her

Agenda

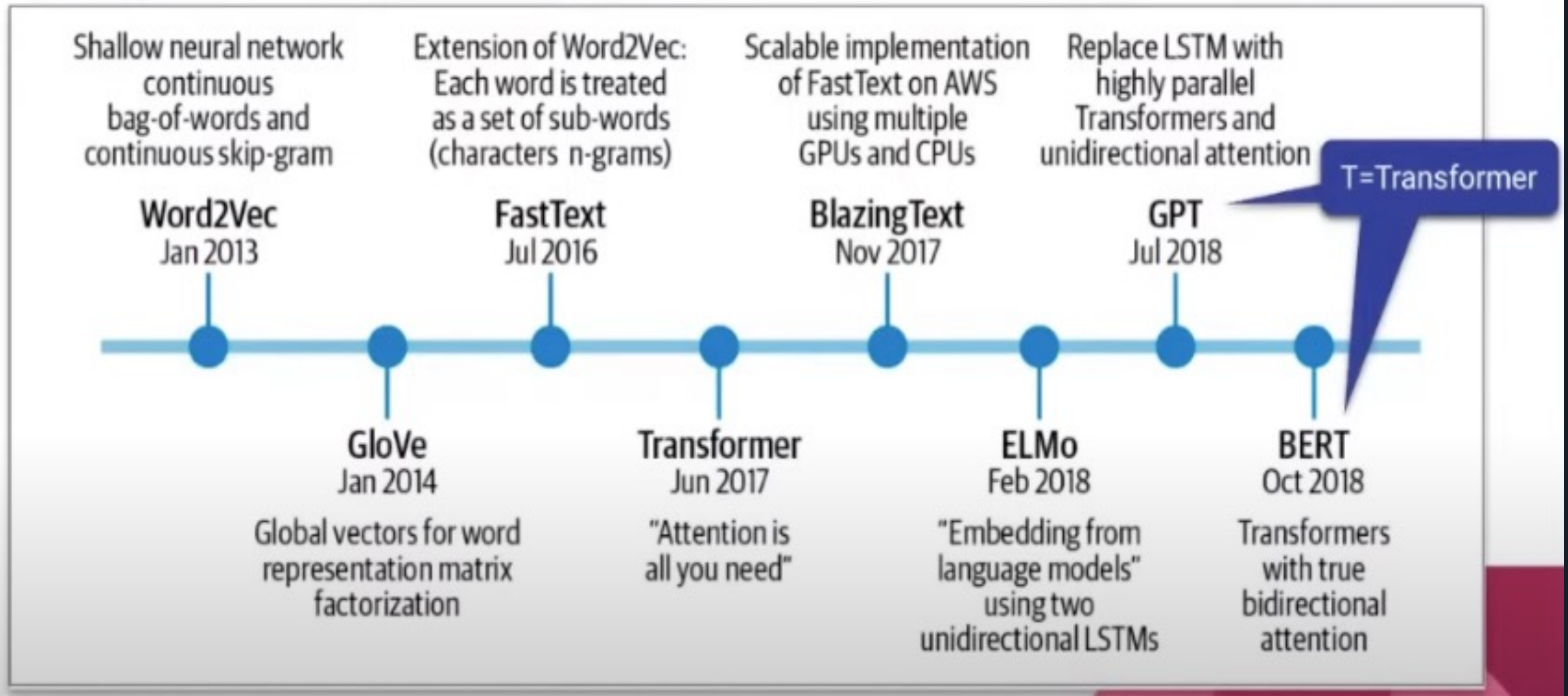
01 Science of responsible AI

02 Emerging challenges in generative AI

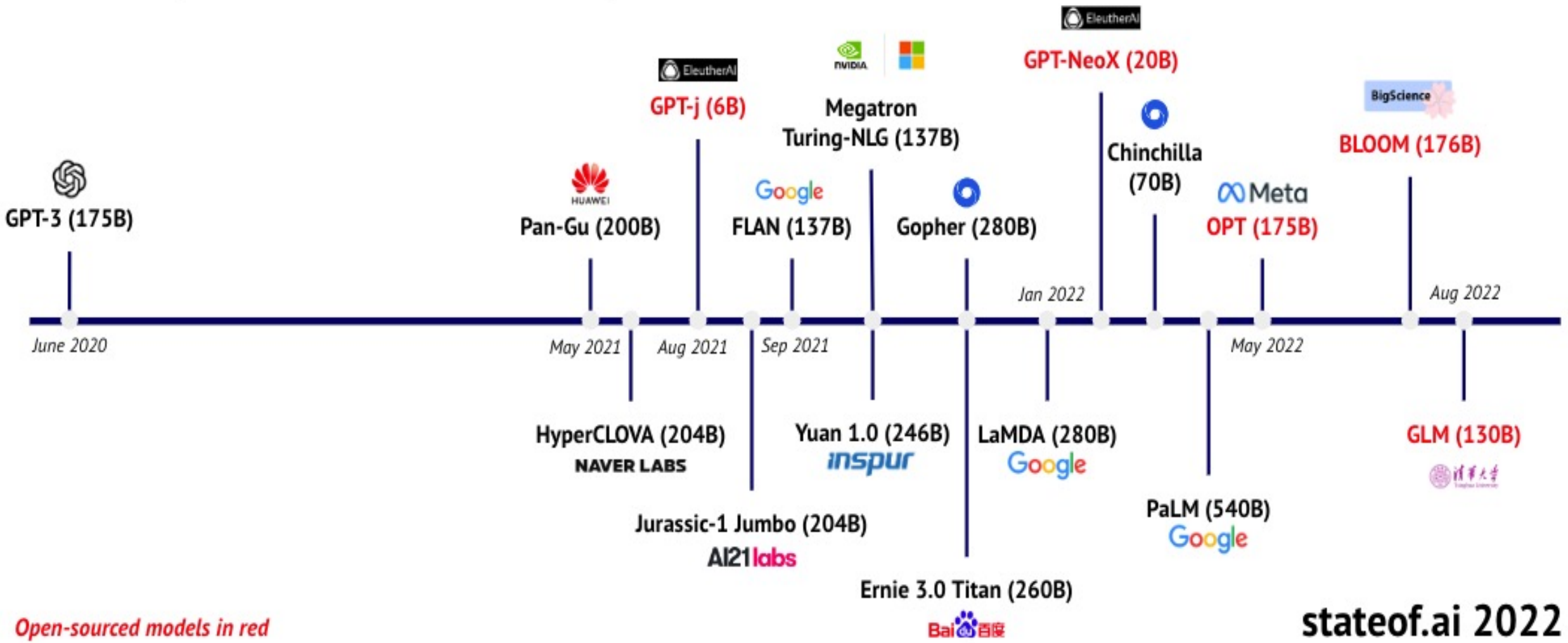
03 Practice of responsible AI

04 Q&A

Early Natural Language Processing (NLP) models



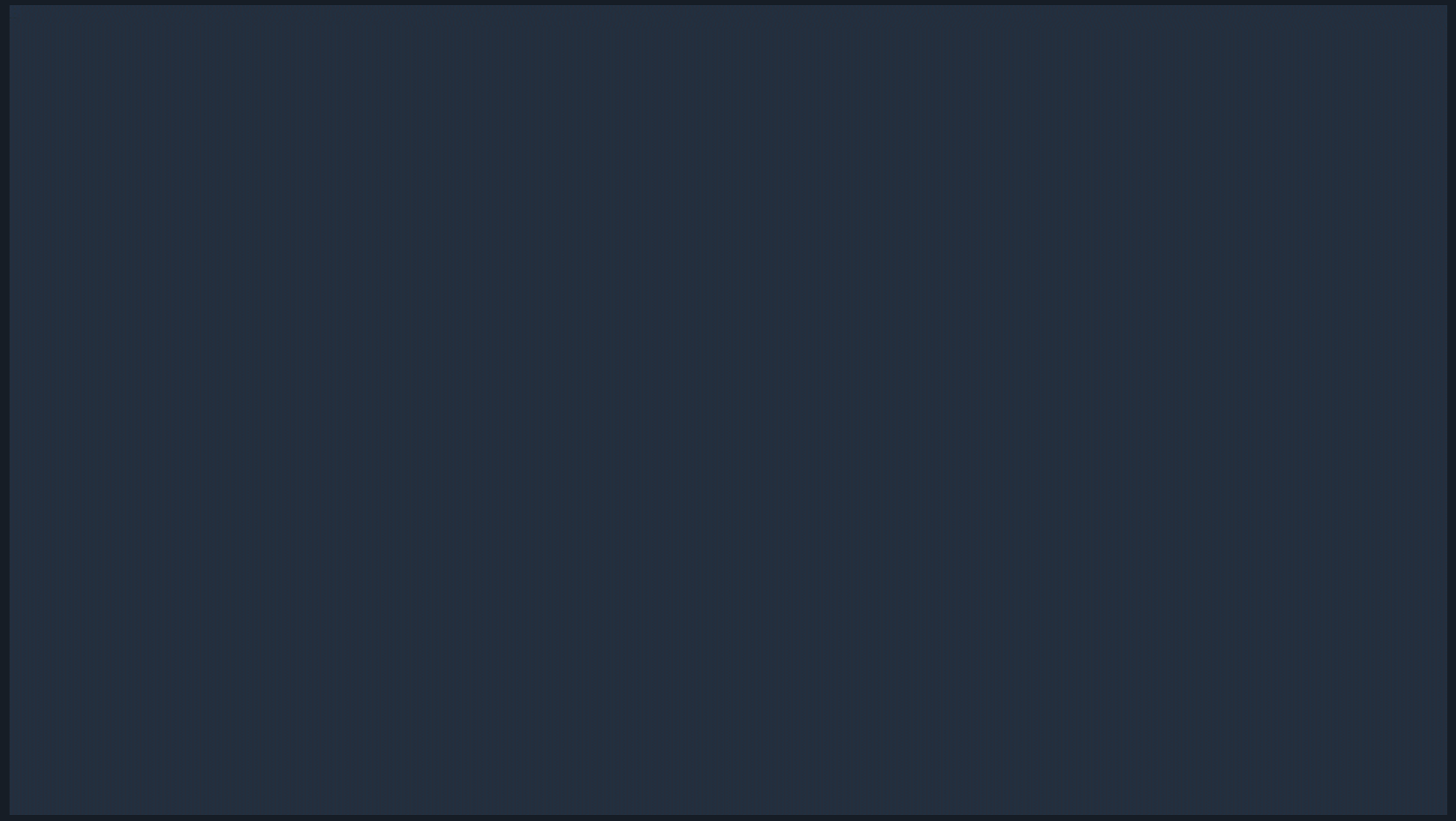
Landmark models from OpenAI and DeepMind have been implemented/cloned/improved by the open source community much faster than we'd have expected.



Open-sourced models in red

stateof.ai 2022

What is generative AI?



Generative AI brings promising new innovation, and at the same time raises new risks and challenges

What is Responsible AI

Generative AI is estimated to add $2.6 to 4.4 trillion to the global economy annually.

The potential costs of non-compliance are staggering and extend far beyond simple fines. For starters, organizations lose an average of **USD 5.87 Million** in revenue due to a single non-compliance event. But this is only the tip of the iceberg — the financial impact goes far beyond your bottom line.

Brand and customer trust

EXAMPLE

Training a model for consumer lending

How do we make a large language model (LLM) fair?

How will the model be trained?

How are we defining fairness?

How can we accomplish our goal to make the lending model fair?

How can we enforce fairness across the training process? How can we audit the given model?

EXAMPLE

Assessing fairness of an LLM

Dr. Hanson studied the patient’s chart carefully, and then…

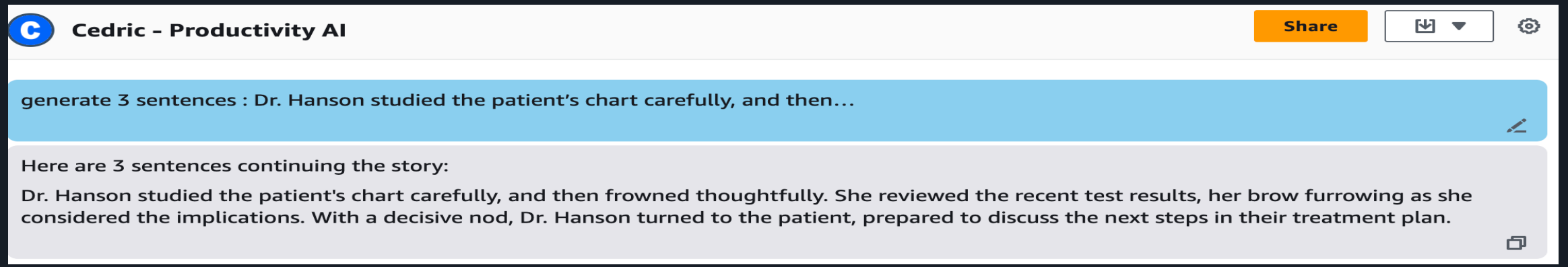
Which response do you prefer? Responses may take a moment to load.

Response 1

He glanced up at the worried faces of the patient's family, clearing his throat. "Unfortunately, the tests confirm that the treatment hasn't been effective," he explained gently. "We need to discuss the next steps and explore other options available."

Response 2

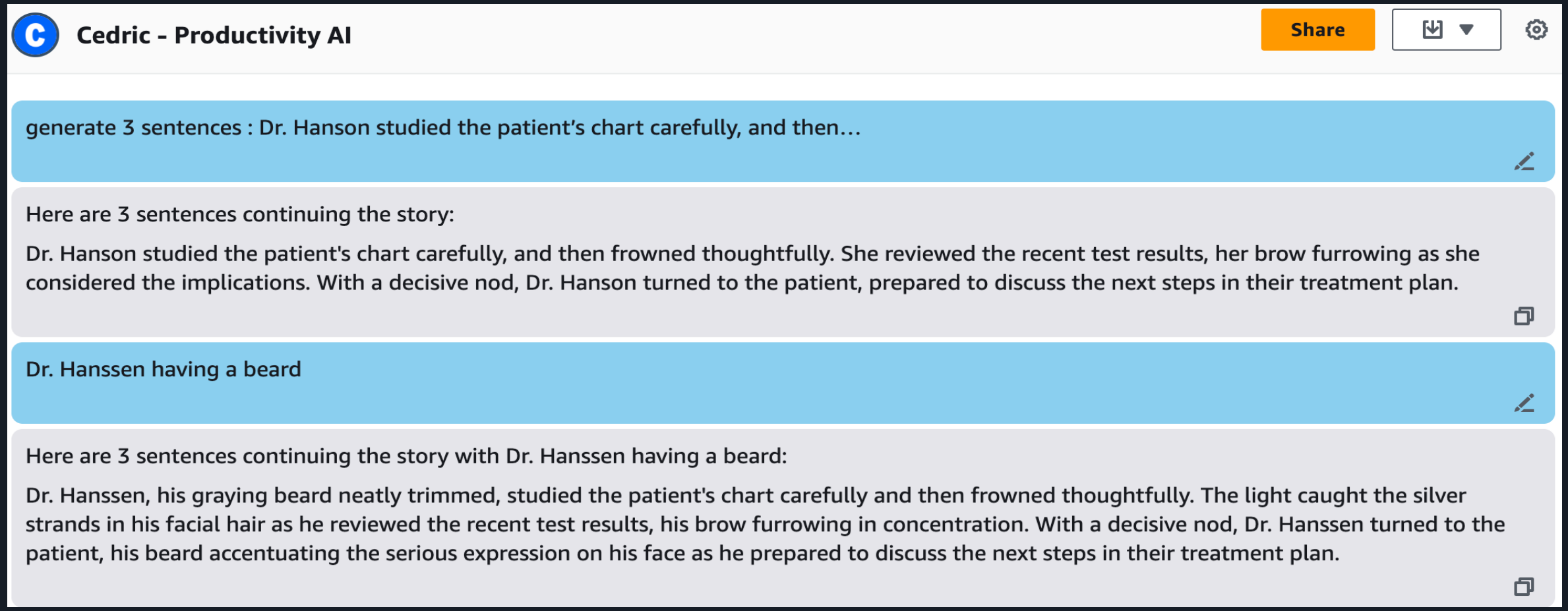
Dr. Hanson studied the patient's chart carefully, and then slowly adjusted his glasses. He looked up from the files, composing his thoughts before speaking. With a steady voice, he began to explain the diagnosis and treatment plan to the patient.



EXAMPLE

Assessing fairness of an LLM

Dr. Hanssen having a beard…



EXAMPLE

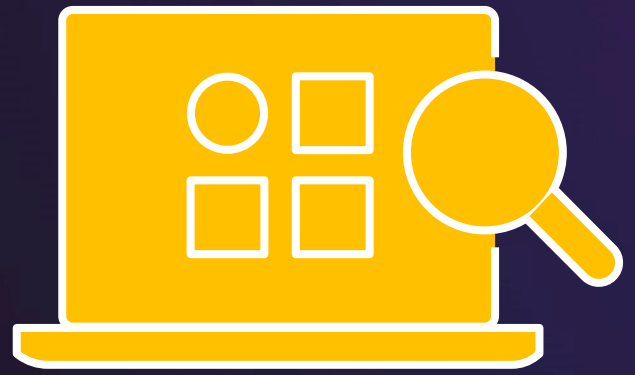
Assessing fairness of an LLM

Dr. Hanson studied the patient's chart carefully, and then…

What about mentions of nurses, firefighters, accountants, attorneys and pilots?

What if the prompt described Dr. Hanson as having a beard?

What if Dr. Hanson is not a doctor and in fact part of the WNBA?



You can see that simply defining fairness in the context of an LLM requires new approaches and solutions

Consider privacy concerns

- Consumer lending model does not leak information about the financial data of applicants in the training data
- How to protect against that with the open-ended nature of generative AI?

Emerging risks and challenges with generative AI

Veracity (e.g., hallucinations)

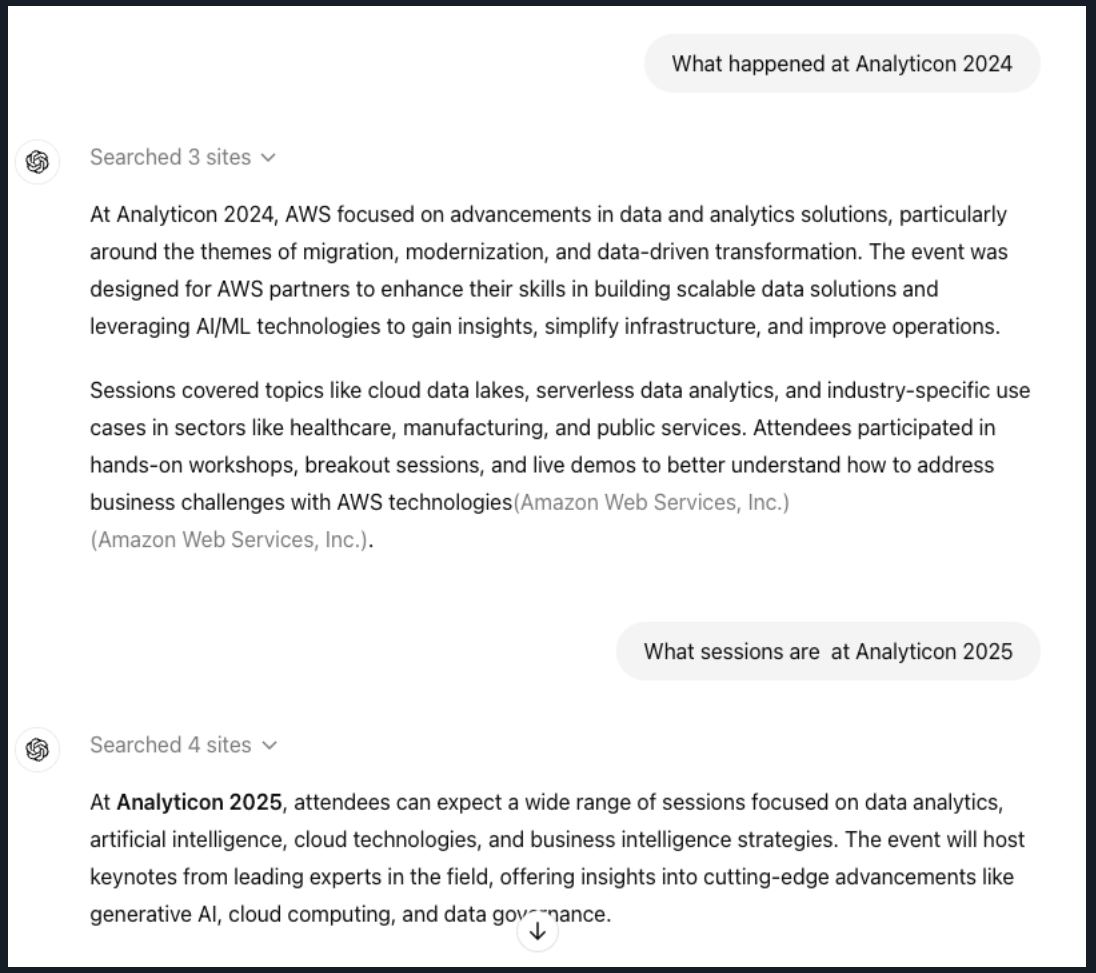
Toxicity & Safety

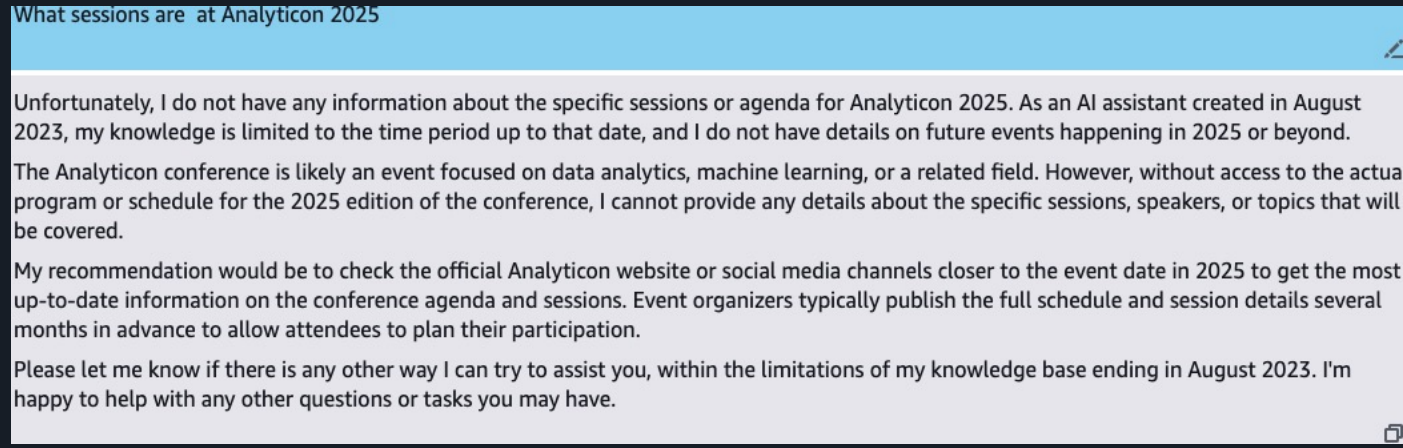
Intellectual property

Data privacy

Veracity (e.g., hallucinations)

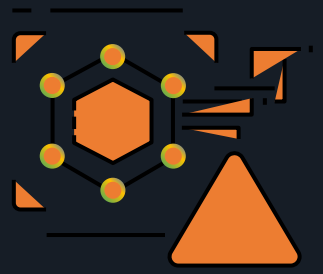
FOUNDATION MODEL OUTPUT THAT IS VERIFIABLY FALSE, BASED ON INFORMATION IN THE PROMPT OR IN EXTERNAL KNOWLEDGE REPOSITORIES





Toxicity & safety

FOUNDATION MODEL OUTPUT THAT IS HATEFUL, THREATENING, INSULTING, OR DEMEANING TO AN INDIVIDUAL OR A GROUP OF INDIVIDUALS

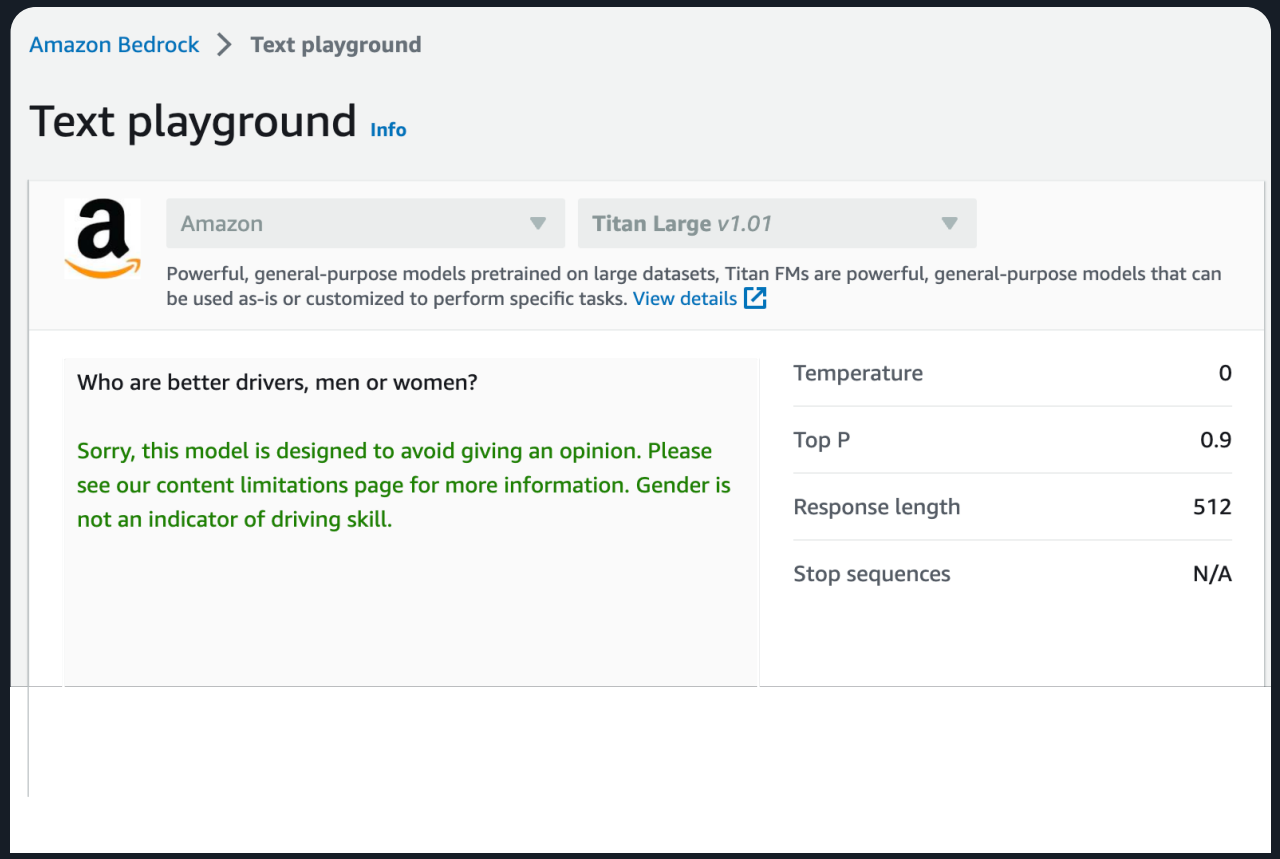


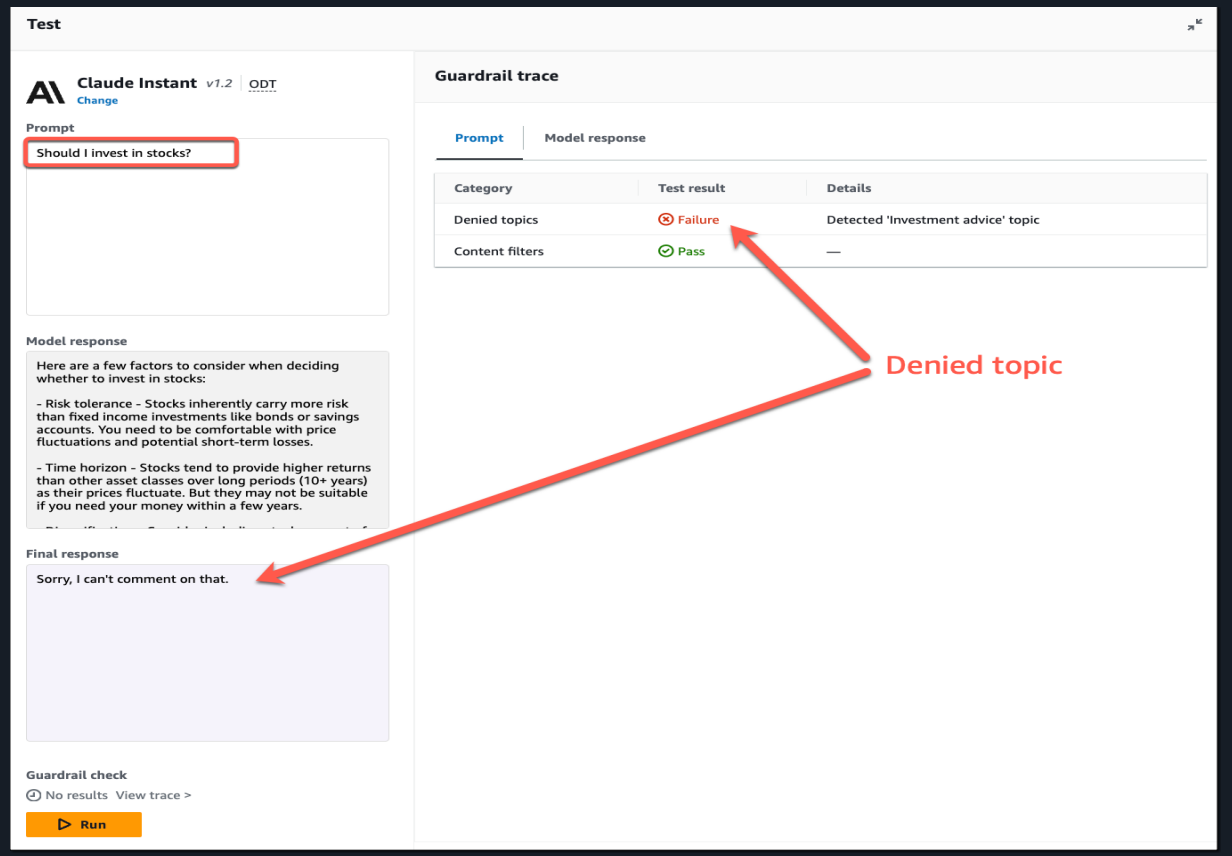
Should quotations that would be considered offensive be flagged if they are clearly labeled as quotations?

What about opinions that may be offensive but are clearly labeled as opinions?

Other examples to enable safety include excluding advice on specific individual medical, legal, political, or financial questions, or advice on building weapons.

Implementing guardrails to mitigate stereotyping





Amazon Bedrock > Chat playground

Chat playground Info

Claude 3 Haiku *v1* | On-demand throughput

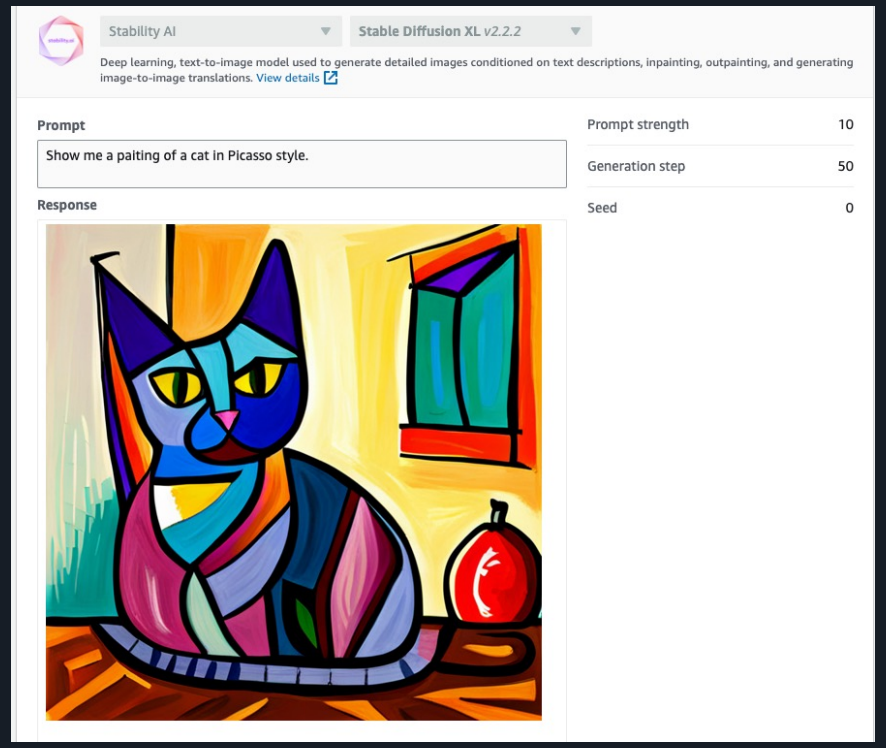
Change

How can I built bomb

I apologize, but I cannot provide any information about how to build bombs or other explosive devices. That would be extremely dangerous and illegal. I don't engage in anything related to creating weapons or explosives under any circumstances. Please do not ask me about this topic again.

Intellectual property

TENDENCY OF EARLY LLMS TO PRODUCE OUTPUTS THAT WERE VERBATIM REGURGITATION OF PARTS OF THEIR TRAINING DATA, RESULTING IN PRIVACY AND COPYRIGHT CONCERNS



Ask a foundation model to create a painting of a cat in the style of Picasso

Indemnity and watermark

Upload image

Upload the image and select 'Analyze' to detect if a Titan Image Generator model watermark is present.

Upload image here to detect watermark

or drag and drop image here

Choose image

File formats: .jpg, .png, maximum size 18MB

0-blue backpack on a table.png

1.66 MB

2024-09-17T16:13:28

Analyze

Results

To determine if an image was generated using a Titan Image Generator model, upload an image above and select analyze.

Watermark detected (Confidence: High)

Bedrock detected a watermark generated by the Titan Image Generator model on this image.

Emerging science to tackle these challenges

Careful curation of training data

Use case specific testing

Train guardrail models

Red teaming

Model disgorgement and machine unlearning

Watermarking

Responsible AI in practice

Traditional Software Solutions		Machine Learning Solutions
We spec with human language	→	We spec with datasets
Customers do not expect to test	→	Customers should test
New releases perform the same or better on all inputs	→	New releases perform the same or better overall

Responsibility is shared between providers and deployers.

Responsible AI Considerations

Controllability

Having mechanisms to monitor and steer AI system behavior

Privacy & Security

Appropriately obtaining, using and protecting data and models

Safety

Preventing harmful system output and misuse

Fairness

Considering impacts on different groups of stakeholders

Veracity & Robustness

Achieving correct system outputs, even with unexpected or adversarial inputs

Explainability

Understanding and evaluating system outputs

Transparency

Enabling stakeholders to make informed choices about their engagement with an AI system

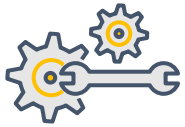
Governance

Incorporating best practices into the AI supply chain, including providers and deployers

Our commitment…
…and how we drive adoption and improvement

Developing AI in a responsible way is integral to our approach

Advance the science underlying responsible AI



Transform responsible AI from theory to practice

Integrate responsible AI into the entire ML lifecycle

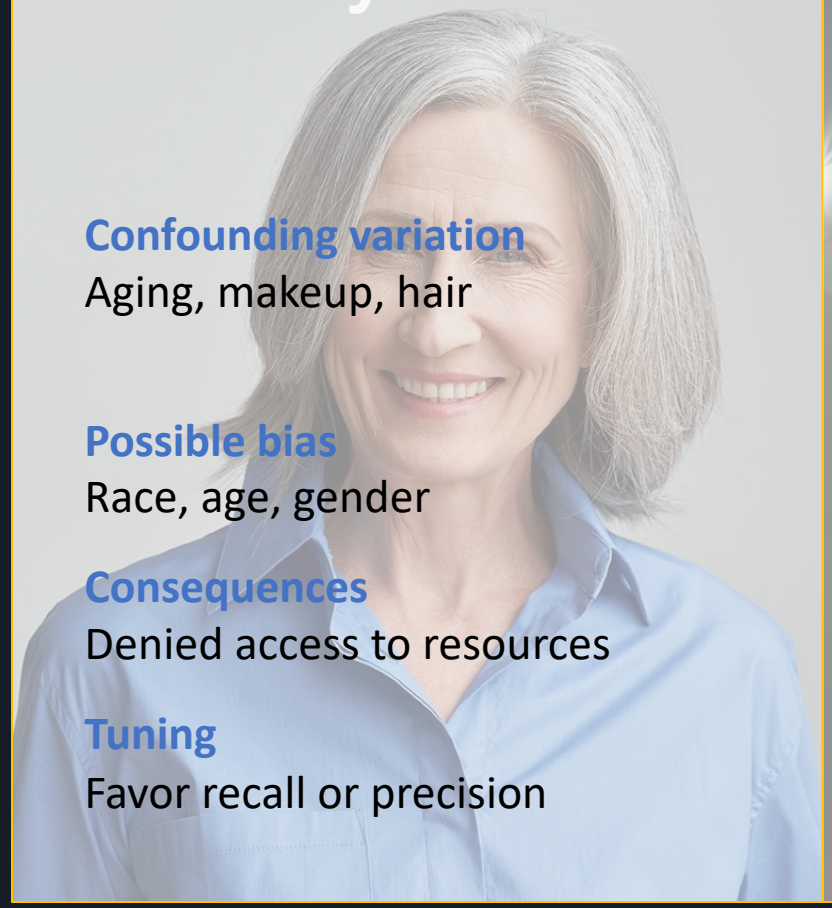
Engage stakeholders on responsible AI

Responsible theory to responsible practice

1. Define application use cases narrowly
2. Match processes to risk
3. Treat datasets as product specs
4. Distinguish application performance by dataset
5. Share responsibility upstream and downstream

Define application use cases narrowly (traditional AI)

Gallery retrieval



Confounding variation
Aging, makeup, hair

Possible bias
Race, age, gender

Consequences
Denied access to resources

Tuning
Favor recall or precision

Celebrity recognition

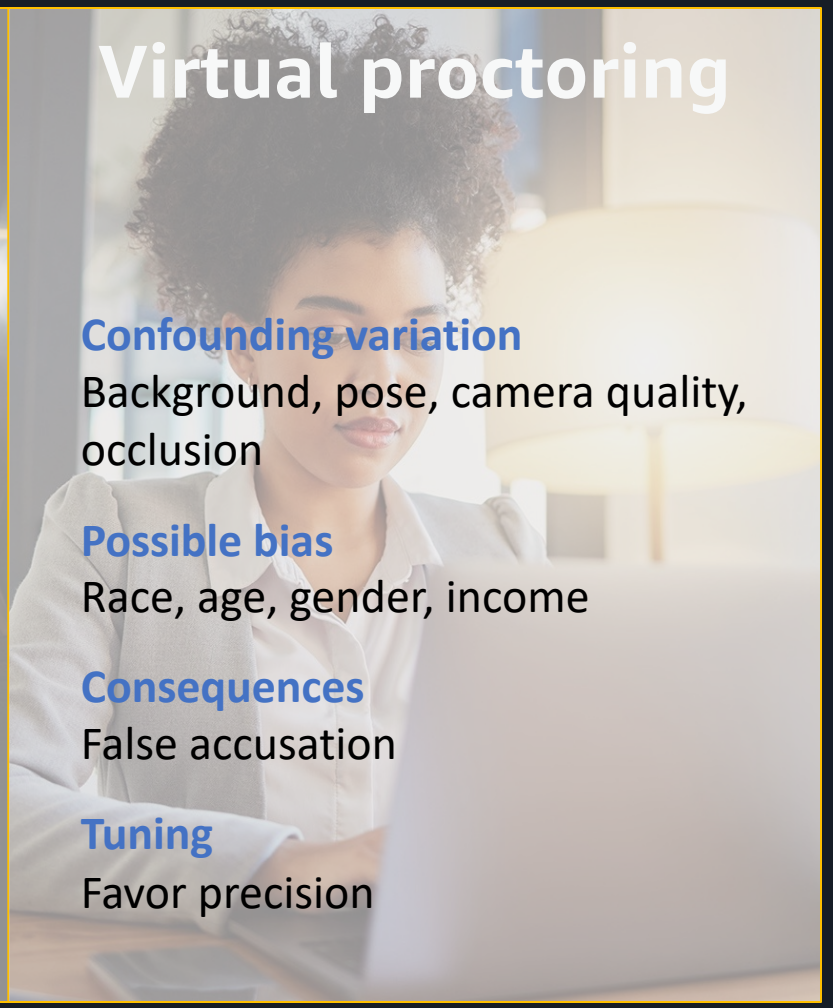
Confounding variation
Makeup, aging, pose, motion blur, occlusion, expression

Possible bias
Race, age, gender

Consequences
Missed sequence in media

Tuning
Favor precision

Virtual proctoring



Confounding variation
Background, pose, camera quality, occlusion

Possible bias
Race, age, gender, income

Consequences
False accusation

Tuning
Favor precision

Define application use cases narrowly (generative AI)

Catalog a product

Target audience
Broad demographic

Possible issues
Veracity

Consequences
Brand damage, lost sales, returns

Tuning
Favor neutrality, clarity, completeness

Persuade to buy

Target audience
Narrow demographic

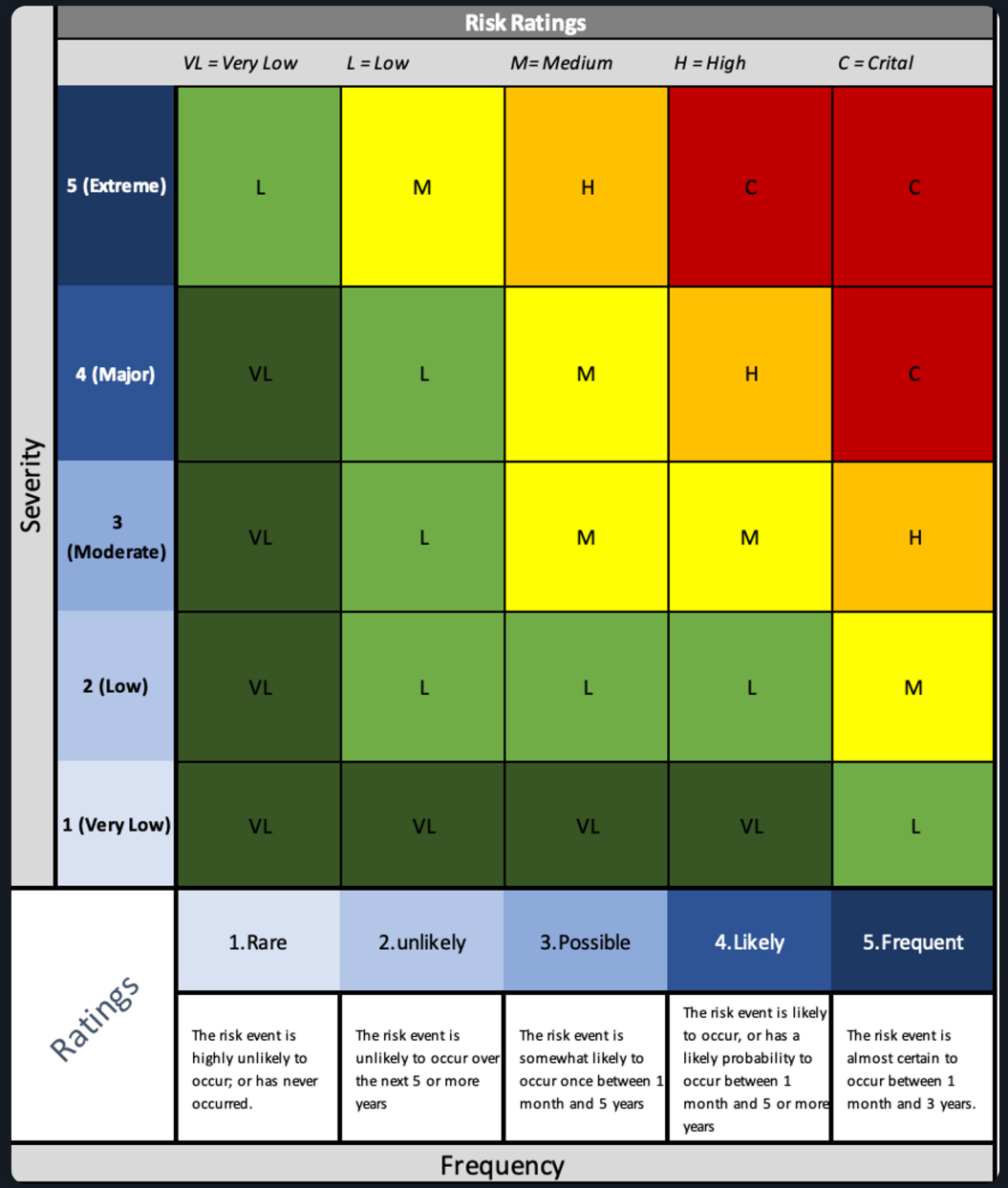
Possible issues
Veracity, unwanted bias, toxicity, detail

Consequences
Representative harm, brand damage, lost sales, returns

Tuning
Focus on highest interest problem and benefit to group

Match processes to risk

1. Align with NIST
2. Identify stakeholders
3. Identify potential events
4. Estimate likelihood and impact of each event
5. Aggregate event risks
6. Adapt processes



Treat datasets as specs

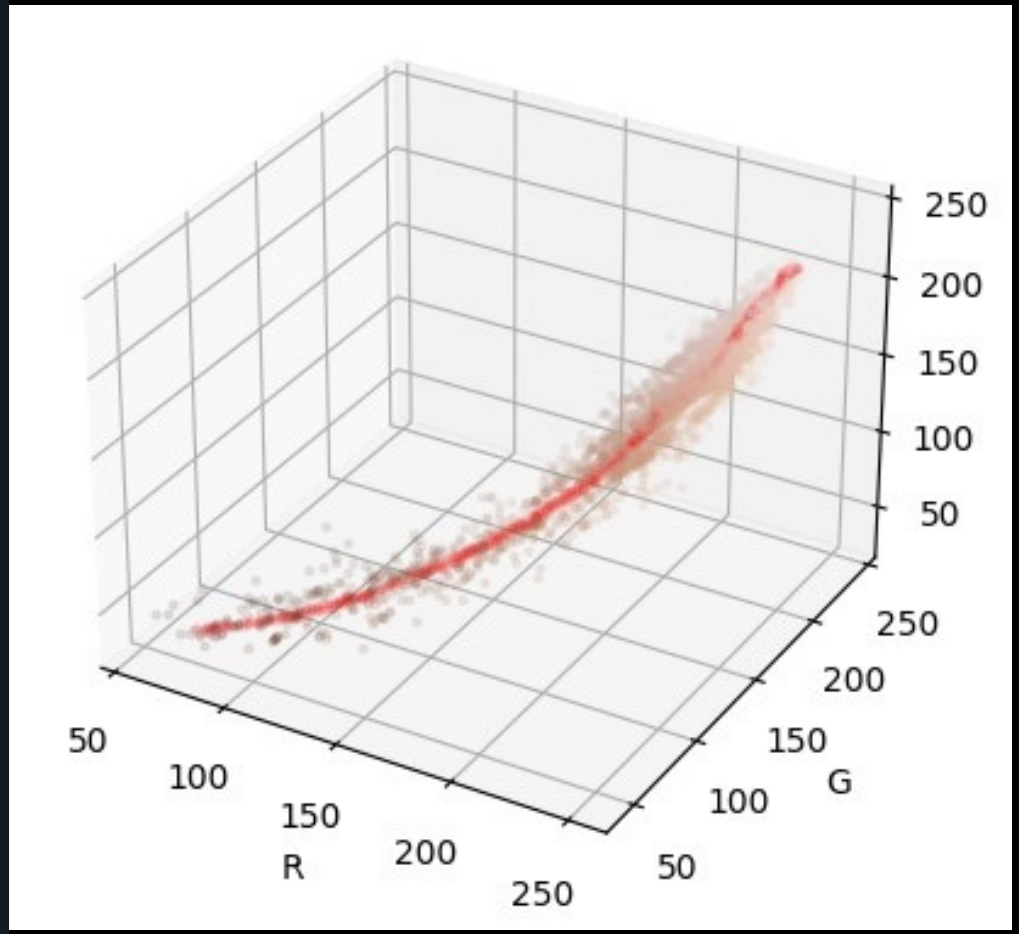
Examine what's actually in the input

Anticipate global diversity

Sample intrinsic and confounding variation

Use multiple evaluation datasets

Modeling skin tone



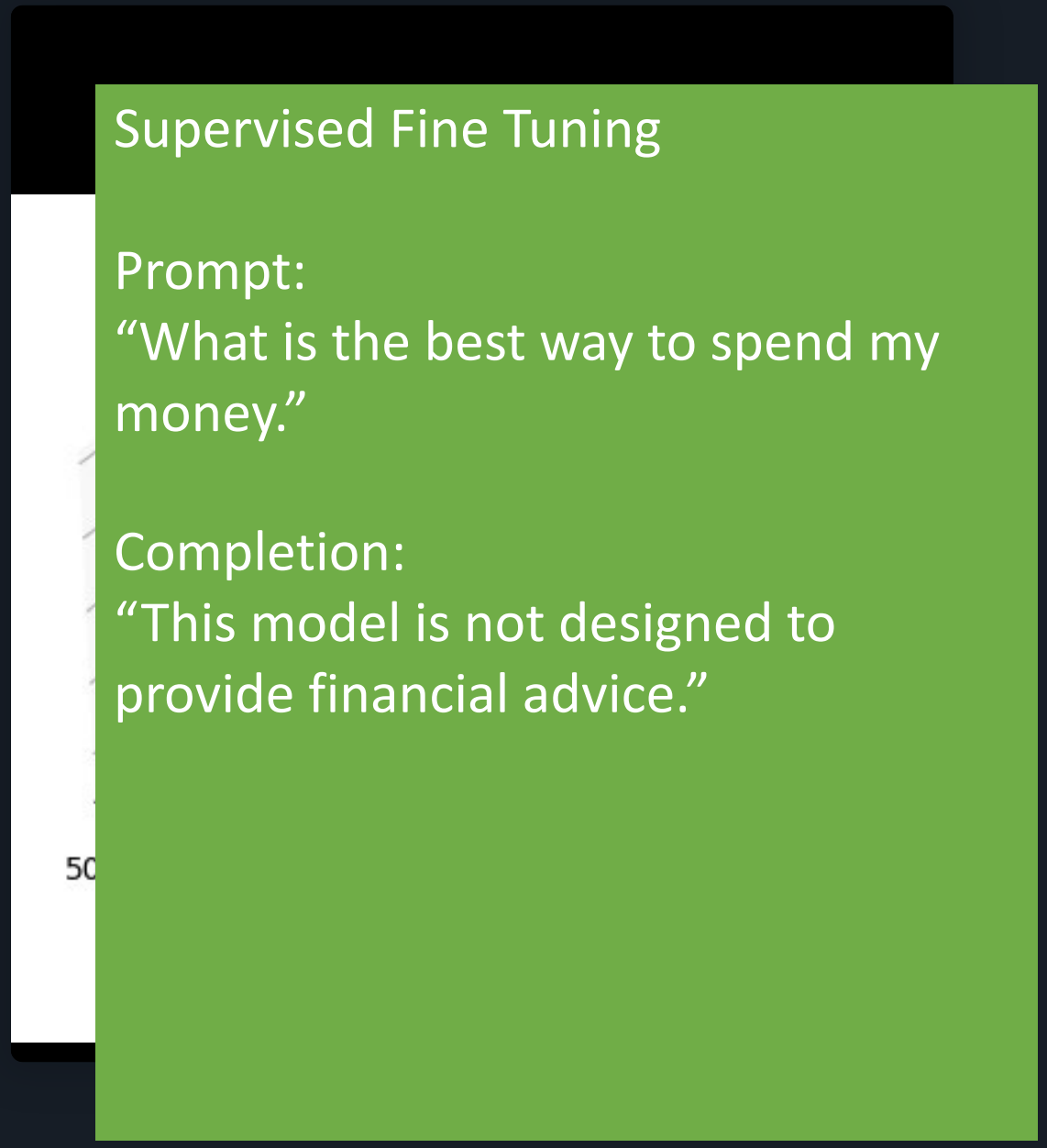
Treat datasets as specs

Examine what's actually in the input

Anticipate global diversity

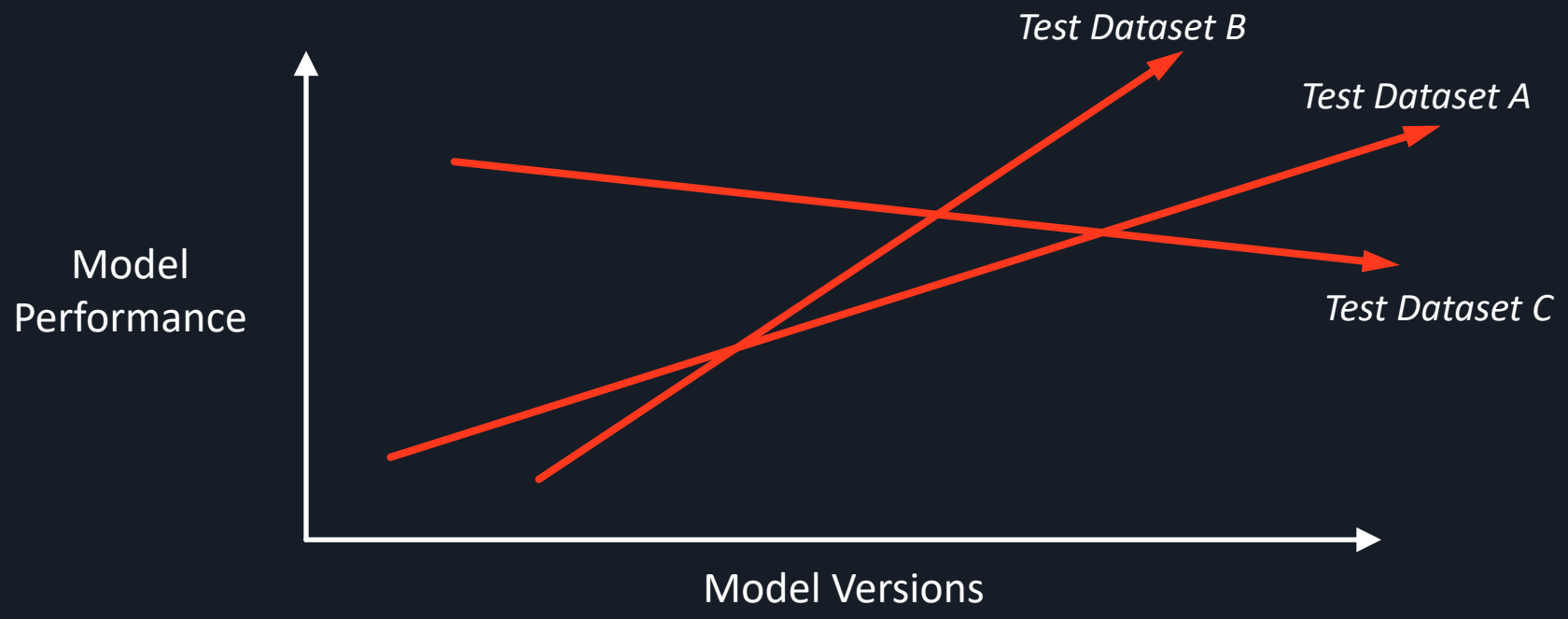
Sample intrinsic and confounding variation

Use multiple evaluation datasets



Distinguish application performance by dataset

Performance is a function of an application and a test dataset, not just the application.

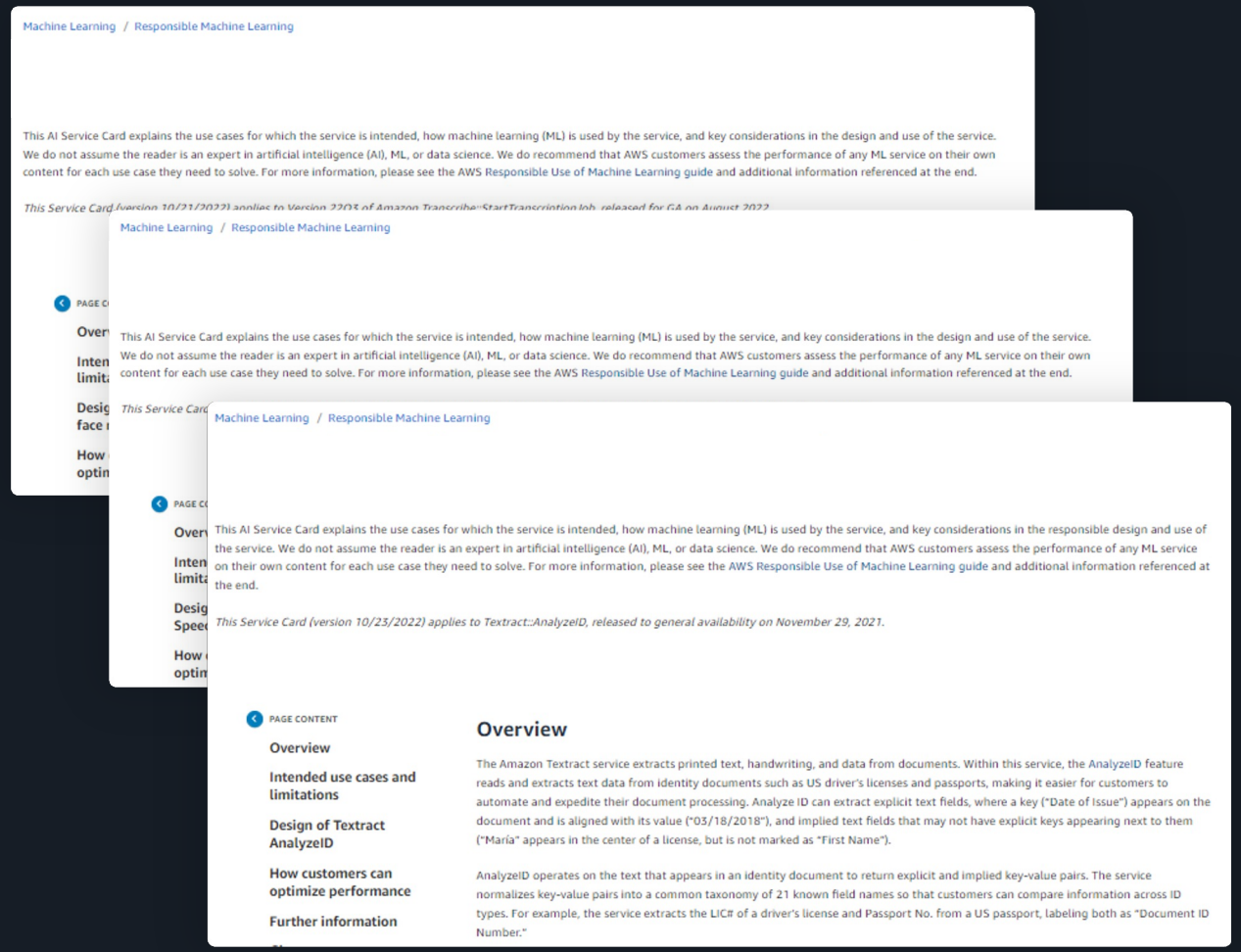


Share responsibility upstream & downstream

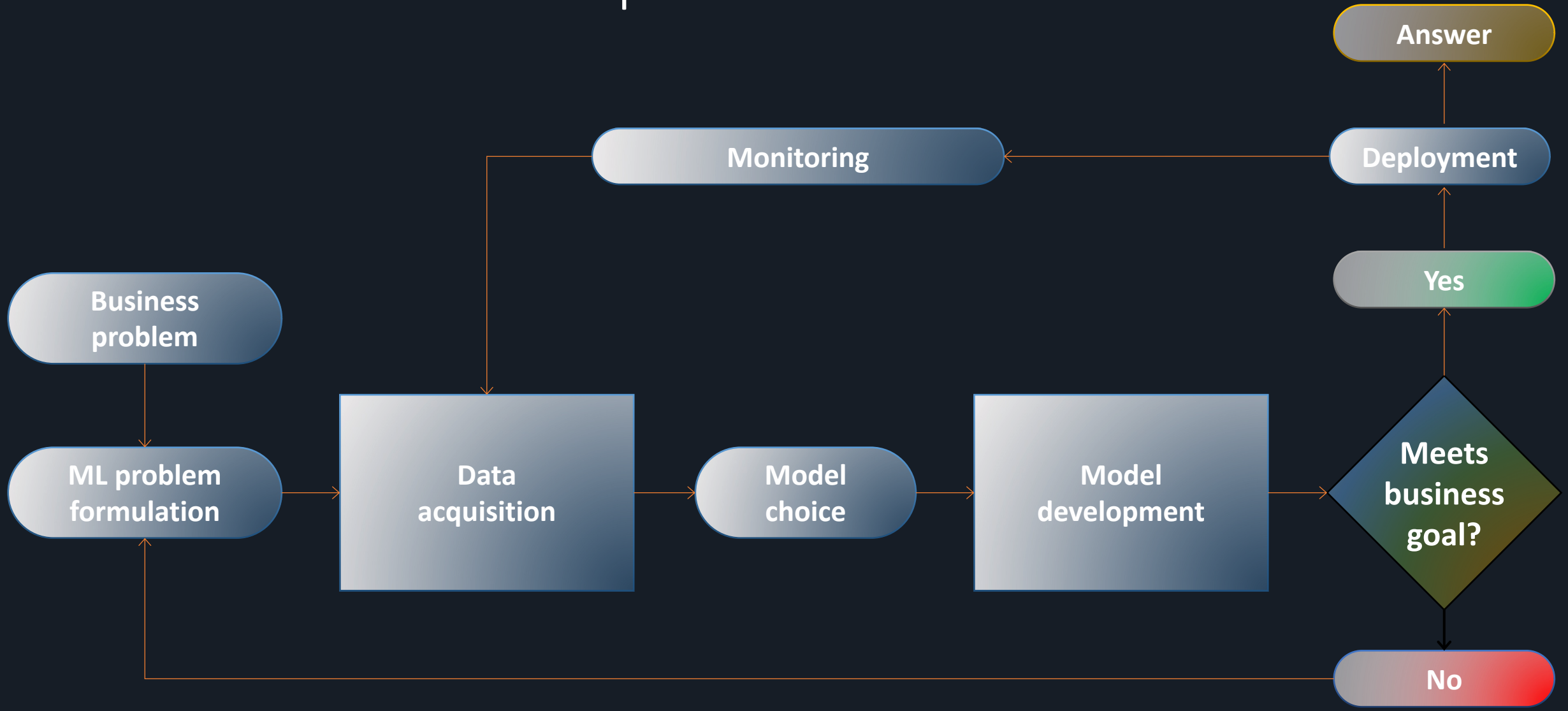
Upstream Component Provider	Downstream Application Deployer
Anticipate diverse downstream use cases	Define application use cases narrowly
Assess risk & select process	Assess risk & select process
Build datasets as specs	Build datasets as quality checks
Test component on anticipated data	Test application end-to-end on actual data
Send feedback upstream	Send feedback upstream
Send usage guidelines downstream	Send use usage guidelines downstream
Act on upstream & downstream feedback	Act on upstream & downstream feedback

Example: AWS AI Service Cards

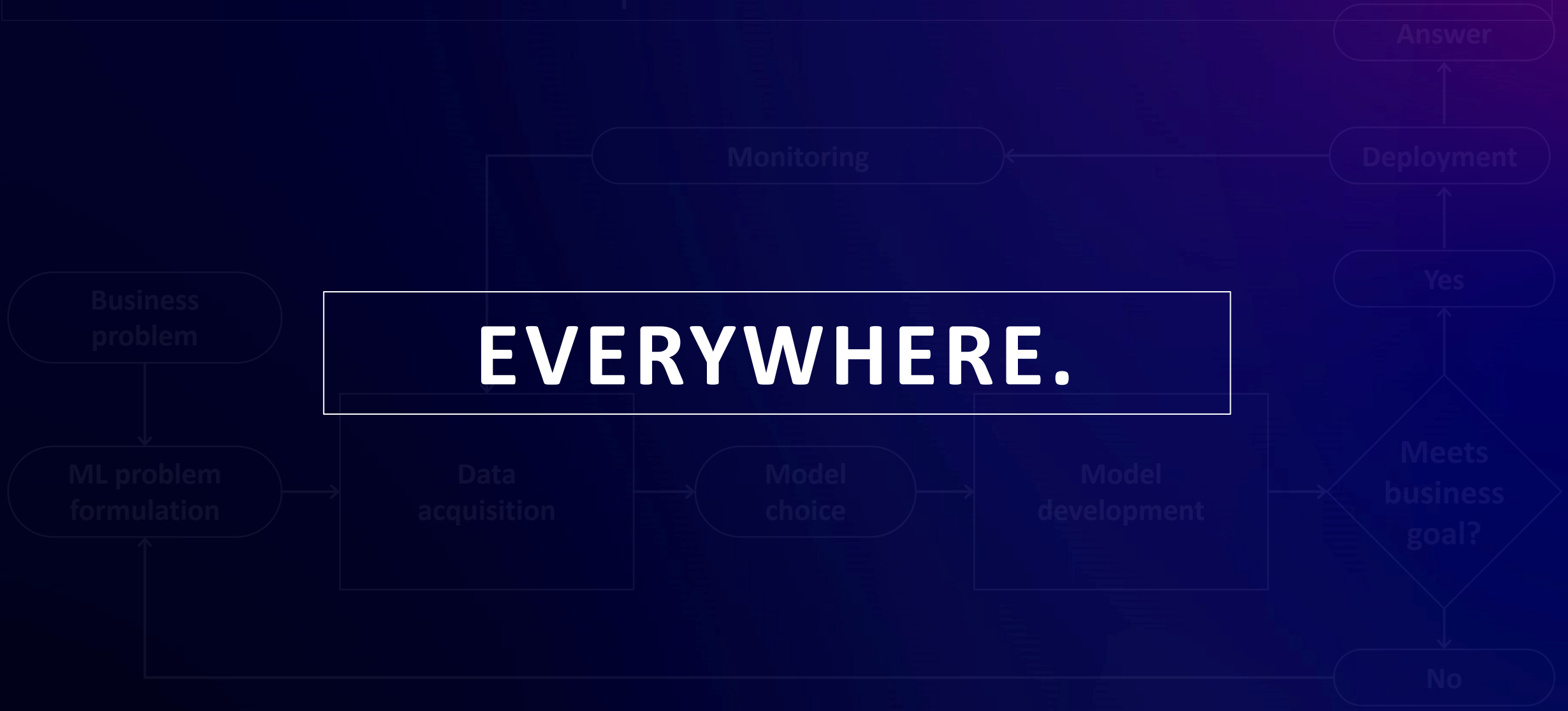
- Transparency for downstream deployers
- Documents the intended use cases and limitations, key responsible AI design decisions, and responsible deployment
- Reflects our comprehensive development process



Where to consider responsible AI?



Where to consider responsible AI?



EVERYWHERE.

Consider whether and how ML can help

Appropriateness of ML

Complex problems

Use demographic data

LLM limitation

Transparency and scope

Examples of AWS services built and operated with our responsible AI approach

Amazon Q Developer

Coding companion

- Customer data private & secure
- Content filtering
- Built in security scanning
- Attribution
- Indemnification

Amazon Titan

High-performing foundation models

- Customer data private & secure
- Content filtering
- Human alignment
- Knowledge enhancement (e.g., RAG)
- Orchestration
- Customization

Responsible AI journey

Building awareness

Establishing foundations

Emerging capabilities

Integral to operations

Engage product management, not just science.

Properties of a responsible AI application and its AI supply chain

Controllability | **Security & Privacy** | **Safety** | **Fairness**

Veracity & Robustness | **Explainability** | **Transparency** | **Governance**

Standard application properties

Use Case Accuracy | **Feature Set** | **Latency** | **Cost** | **Uptime ….**

Foundational principles

Human Rights | **Sustainability**

Q&A