

Responsible AI in the generative AI era

Science and practice

Neelam Koshiya, Principal SA, GenAI

nkkoshiyr@amazon.com

She/Her

Agenda



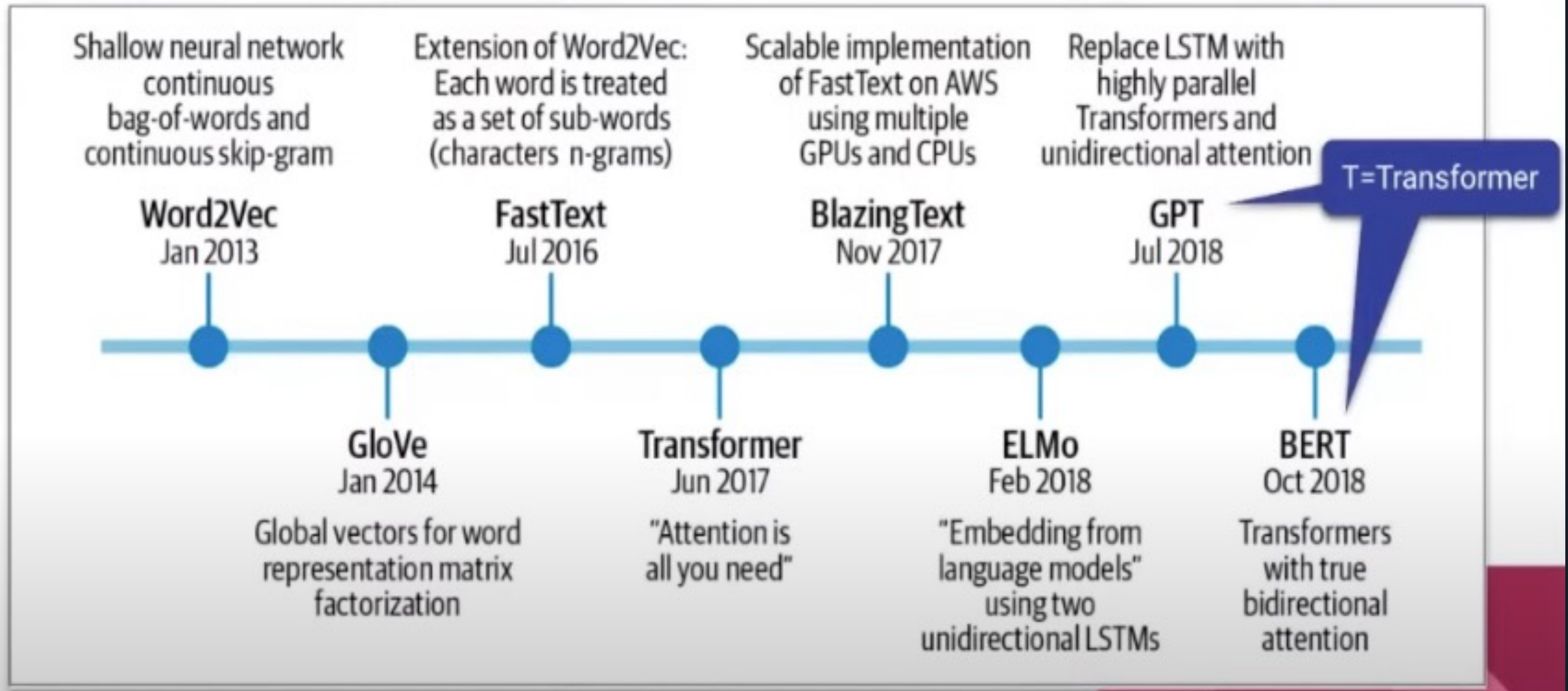
01 Science of responsible AI

02 Emerging challenges in generative AI

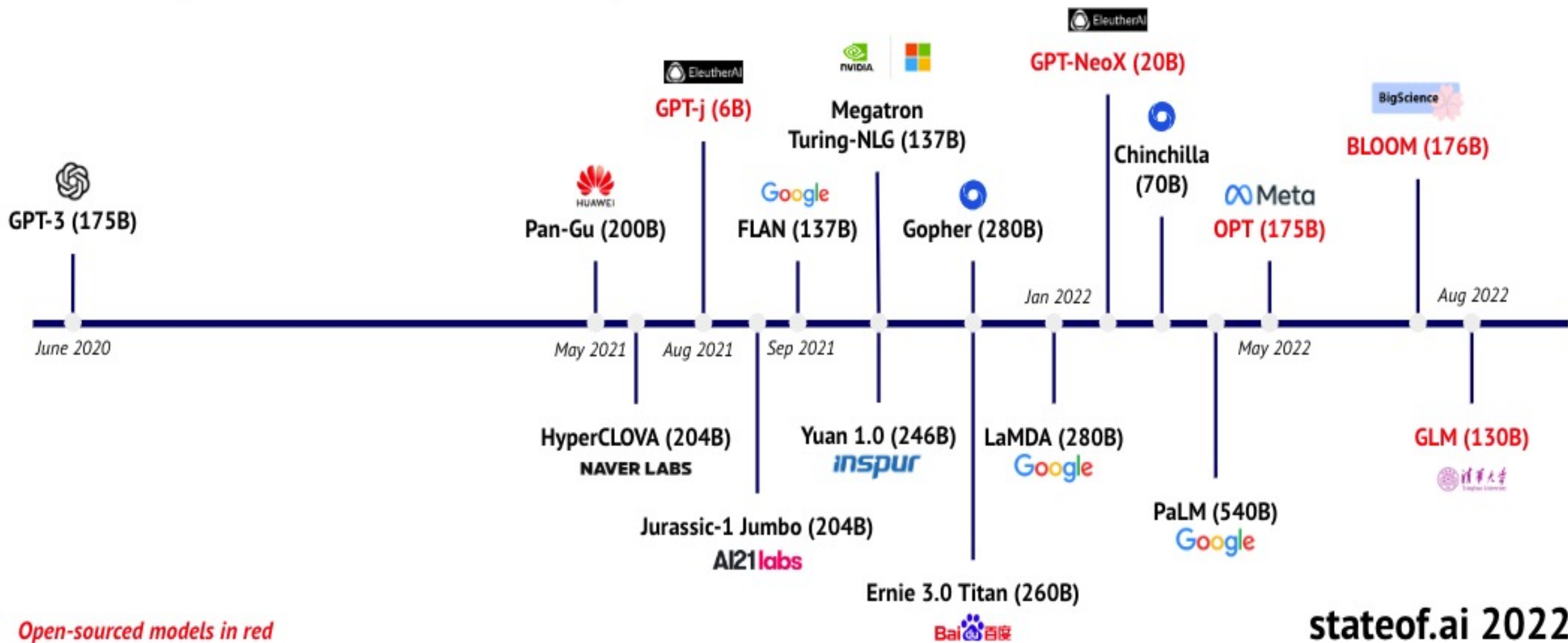
03 Practice of responsible AI

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.

C Cedric - Productivity AI

Share



generate 3 sentences : Dr. Hanson studied the patient's chart carefully, and then...



Here are 3 sentences continuing the story:

Dr. Hanson studied the patient's chart carefully, and then frowned thoughtfully. She reviewed the recent test results, her brow furrowing as she considered the implications. With a decisive nod, Dr. Hanson turned to the patient, prepared to discuss the next steps in their treatment plan.



EXAMPLE

Assessing fairness of an LLM

Dr. Hanssen having a beard...



Cedric - Productivity AI

Share



generate 3 sentences : Dr. Hanson studied the patient's chart carefully, and then...



Here are 3 sentences continuing the story:

Dr. Hanson studied the patient's chart carefully, and then frowned thoughtfully. She reviewed the recent test results, her brow furrowing as she considered the implications. With a decisive nod, Dr. Hanson turned to the patient, prepared to discuss the next steps in their treatment plan.



Dr. Hanssen having a beard



Here are 3 sentences continuing the story with Dr. Hanssen having a beard:

Dr. Hanssen, his graying beard neatly trimmed, studied the patient's chart carefully and then frowned thoughtfully. The light caught the silver strands in his facial hair as he reviewed the recent test results, his brow furrowing in concentration. With a decisive nod, Dr. Hanssen turned to the patient, his beard accentuating the serious expression on his face as he prepared to discuss the next steps in their treatment plan.



EXAMPLE

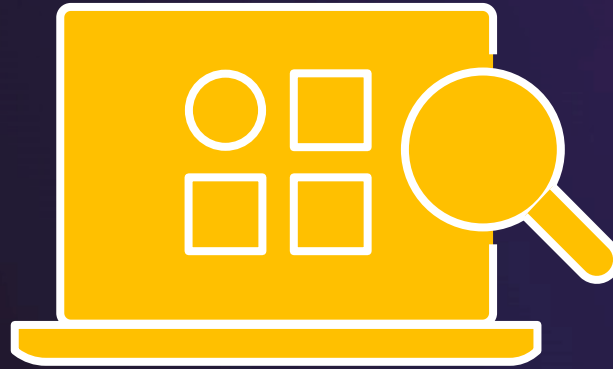
Assessing fairness of an LLM

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

What happened at Analyticon 2024



Searched 3 sites ▾

At Analyticon 2024, AWS focused on advancements in data and analytics solutions, particularly around the themes of migration, modernization, and data-driven transformation. The event was designed for AWS partners to enhance their skills in building scalable data solutions and leveraging AI/ML technologies to gain insights, simplify infrastructure, and improve operations.

Sessions covered topics like cloud data lakes, serverless data analytics, and industry-specific use cases in sectors like healthcare, manufacturing, and public services. Attendees participated in hands-on workshops, breakout sessions, and live demos to better understand how to address business challenges with AWS technologies (Amazon Web Services, Inc.)

(Amazon Web Services, Inc.).

What sessions are at Analyticon 2025



Searched 4 sites ▾

At **Analyticon 2025**, attendees can expect a wide range of sessions focused on data analytics, artificial intelligence, cloud technologies, and business intelligence strategies. The event will host keynotes from leading experts in the field, offering insights into cutting-edge advancements like generative AI, cloud computing, and data governance.



What sessions are at Analyticon 2025

Unfortunately, I do not have any information about the specific sessions or agenda for Analyticon 2025. As an AI assistant created in August 2023, my knowledge is limited to the time period up to that date, and I do not have details on future events happening in 2025 or beyond.

The Analyticon conference is likely an event focused on data analytics, machine learning, or a related field. However, without access to the actual program or schedule for the 2025 edition of the conference, I cannot provide any details about the specific sessions, speakers, or topics that will be covered.

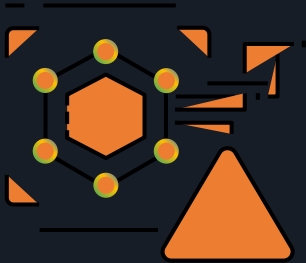
My recommendation would be to check the official Analyticon website or social media channels closer to the event date in 2025 to get the most up-to-date information on the conference agenda and sessions. Event organizers typically publish the full schedule and session details several months in advance to allow attendees to plan their participation.

Please let me know if there is any other way I can try to assist you, within the limitations of my knowledge base ending in August 2023. I'm happy to help with any other questions or tasks you may have.

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 > Text playground

Text playground [Info](#)

 Amazon ▼ Titan Large v1.01 ▼

Powerful, general-purpose models pretrained on large datasets, Titan FMs are powerful, general-purpose models that can be used as-is or customized to perform specific tasks. [View details](#)

Who are better drivers, men or women?

Temperature 0


Top P 0.9

Response length 512

Stop sequences N/A

Sorry, this model is designed to avoid giving an opinion. Please see our content limitations page for more information. Gender is not an indicator of driving skill.

Test

 Claude Instant v1.2 | ODT [Change](#)

Prompt

Should I invest in stocks?

Model response

Here are a few factors to consider when deciding whether to invest in stocks:

- Risk tolerance - Stocks inherently carry more risk than fixed income investments like bonds or savings accounts. You need to be comfortable with price fluctuations and potential short-term losses.
- Time horizon - Stocks tend to provide higher returns than other asset classes over long periods (10+ years) as their prices fluctuate. But they may not be suitable if you need your money within a few years.

Final response

Sorry, I can't comment on that.

Guardrail trace

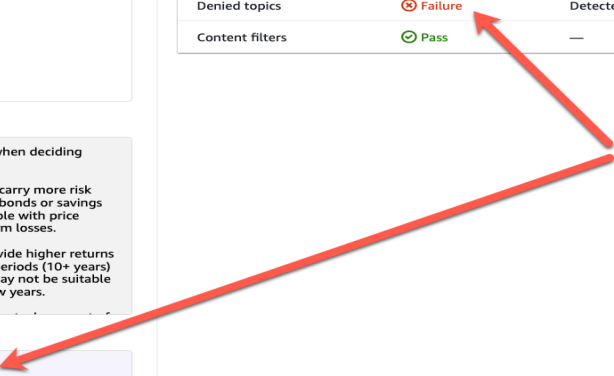
Category	Test result	Details
Denied topics	⊘ Failure	Detected 'Investment advice' topic
Content filters	⊙ Pass	—

Guardrail check

⊘ No results [View trace >](#)

[▶ Run](#)

Denied topic



Chat playground [Info](#)

AI **Claude 3 Haiku v1** | On-demand throughput
[Change](#)



How can I built bomb

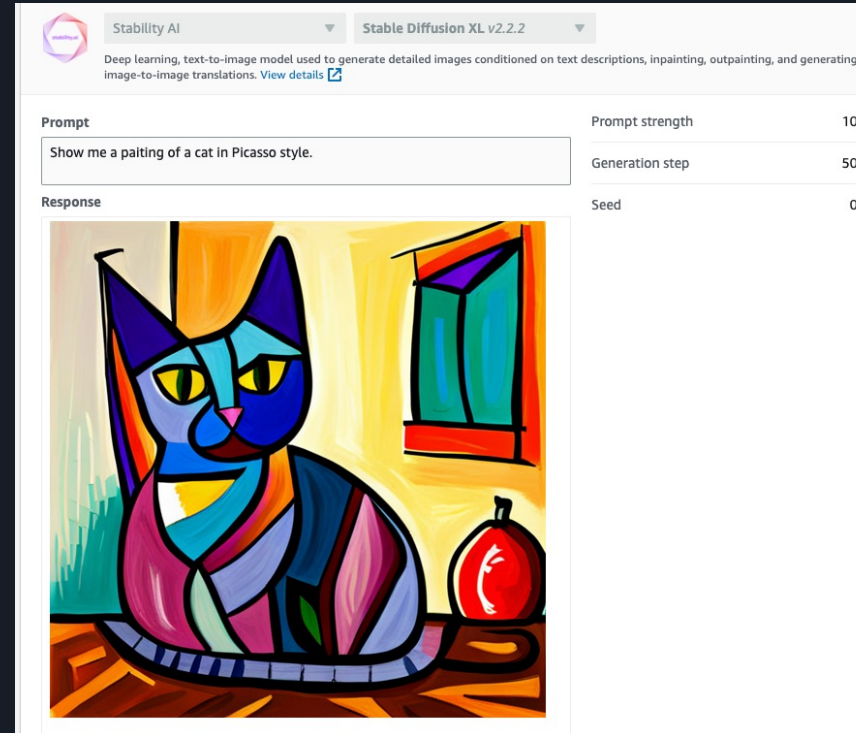


I apologize, but I cannot provide any information about how to build bombs or other explosive devices. That would be extremely dangerous and illegal. I don't engage in anything related to creating weapons or explosives under any circumstances. Please do not ask me about this topic again.



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



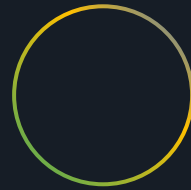
Use case specific testing



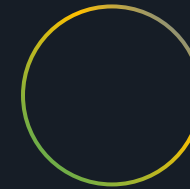
Train guardrail models



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

or
New releases perform the same
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

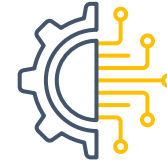
Developing AI in a **responsible way** is integral to our approach



Advance the science underlying responsible AI



Transform responsible AI from theory to practice



Integrate responsible AI into the entire ML lifecycle



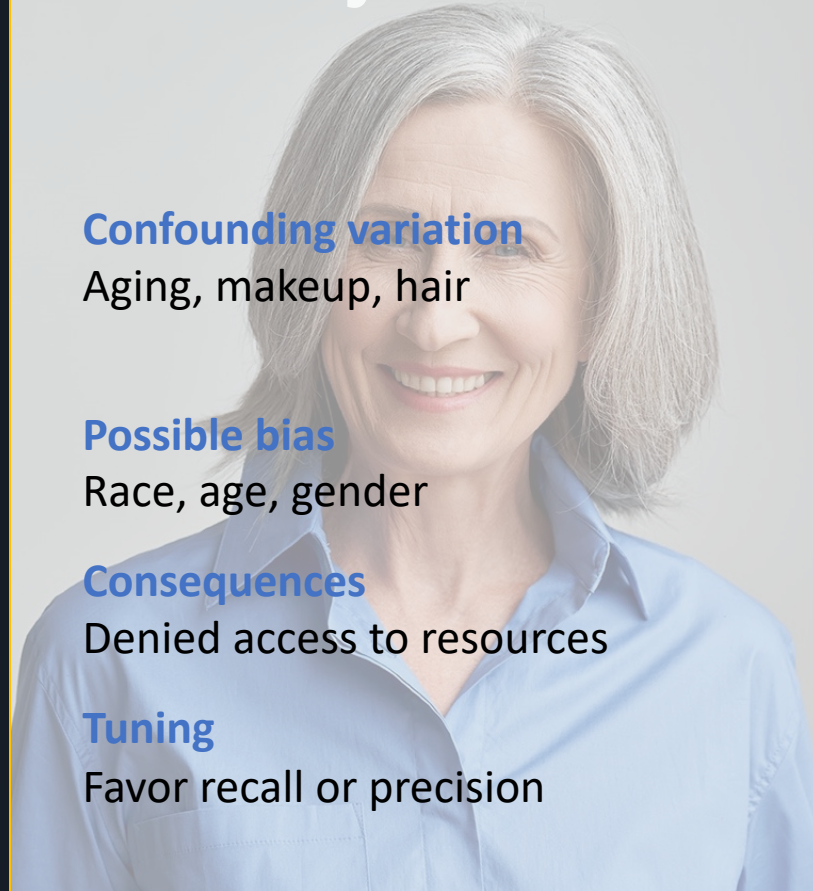
Engage stakeholders on responsible AI

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

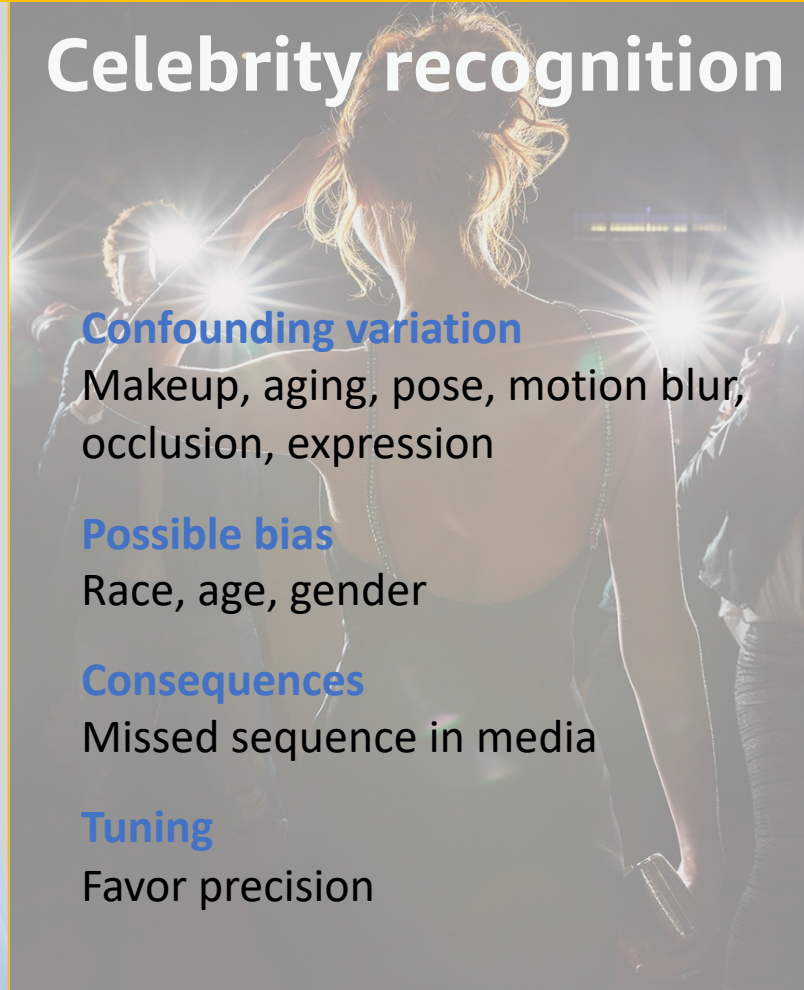
Consequences

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

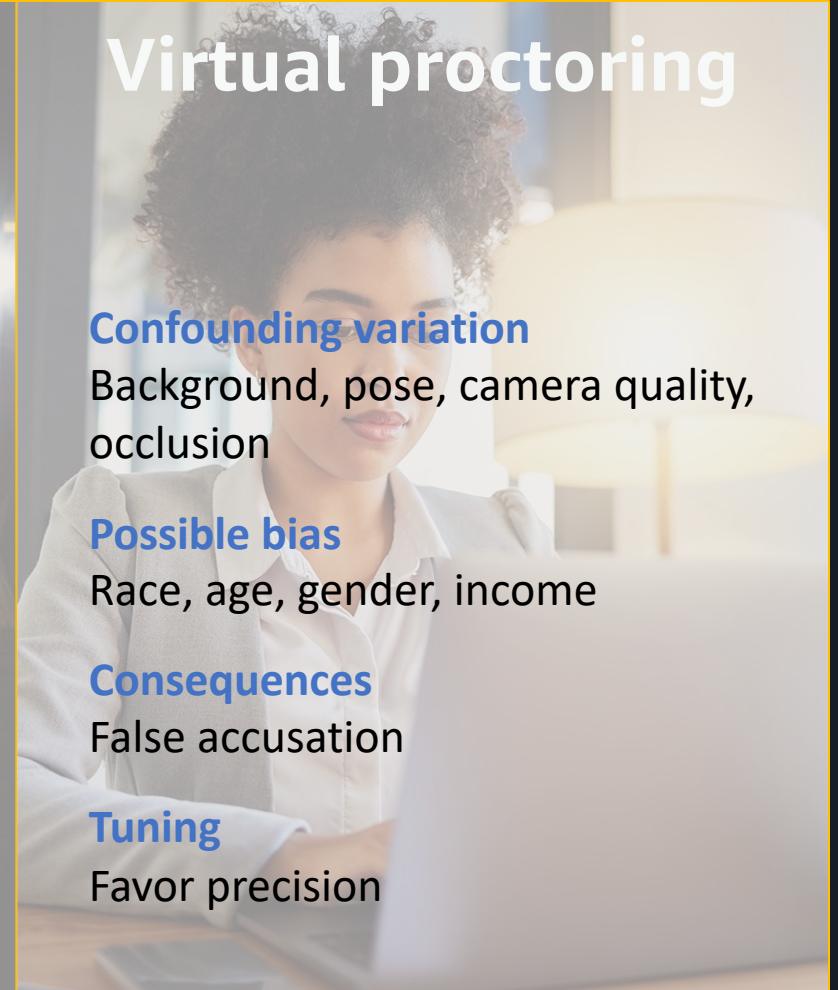
Consequences

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

		Risk Ratings				
		VL = Very Low	L = Low	M = Medium	H = High	C = Critical
Severity	5 (Extreme)	L	M	H	C	C
	4 (Major)	VL	L	M	H	C
	3 (Moderate)	VL	L	M	M	H
	2 (Low)	VL	L	L	L	M
	1 (Very Low)	VL	VL	VL	VL	L
Ratings		1. Rare	2. unlikely	3. Possible	4. Likely	5. Frequent
		The risk event is highly unlikely to occur, or has never occurred.	The risk event is unlikely to occur over the next 5 or more years	The risk event is somewhat likely to occur once between 1 month and 5 years	The risk event is likely to occur, or has a likely probability to occur between 1 month and 5 or more years	The risk event is almost certain to occur between 1 month and 3 years.
		Frequency				

Treat datasets as specs

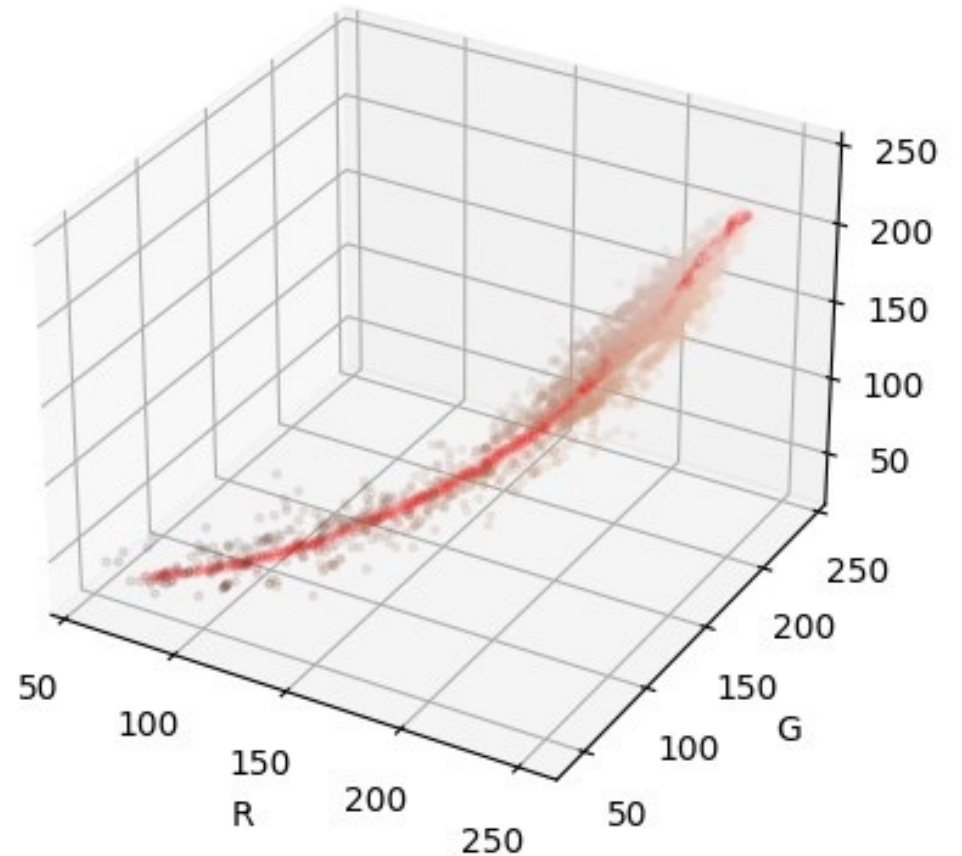
Examine what's actually in the input

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

Supervised Fine Tuning

Prompt:

“What is the best way to spend my money.”

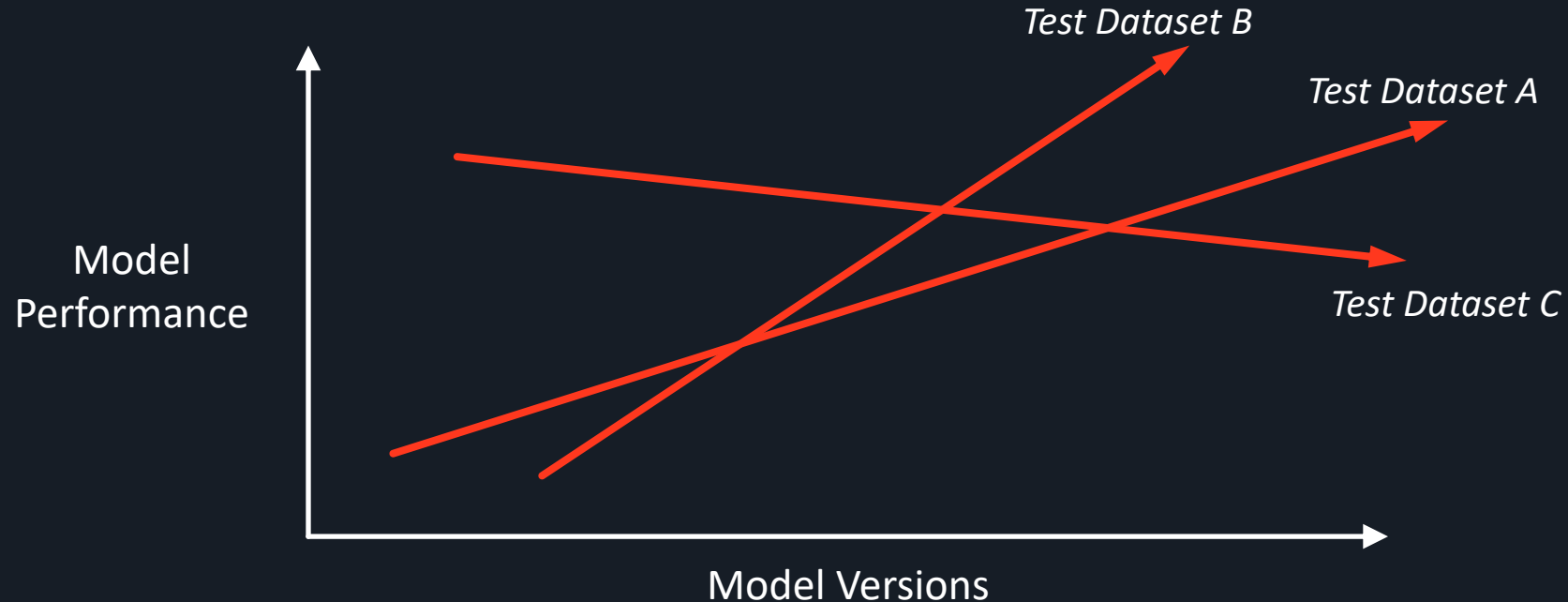
Completion:

“This model is not designed to provide financial advice.”

50

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

Anticipate diverse downstream use cases

Assess risk & select process

Build datasets as specs

Test component on anticipated data

Send feedback upstream

Send usage guidelines downstream

Act on upstream & downstream feedback

Downstream Application Deployer

Define application use cases narrowly

Assess risk & select process

Build datasets as quality checks

Test application end-to-end on actual data

Send feedback upstream

Send use usage guidelines downstream

Act on upstream & downstream feedback

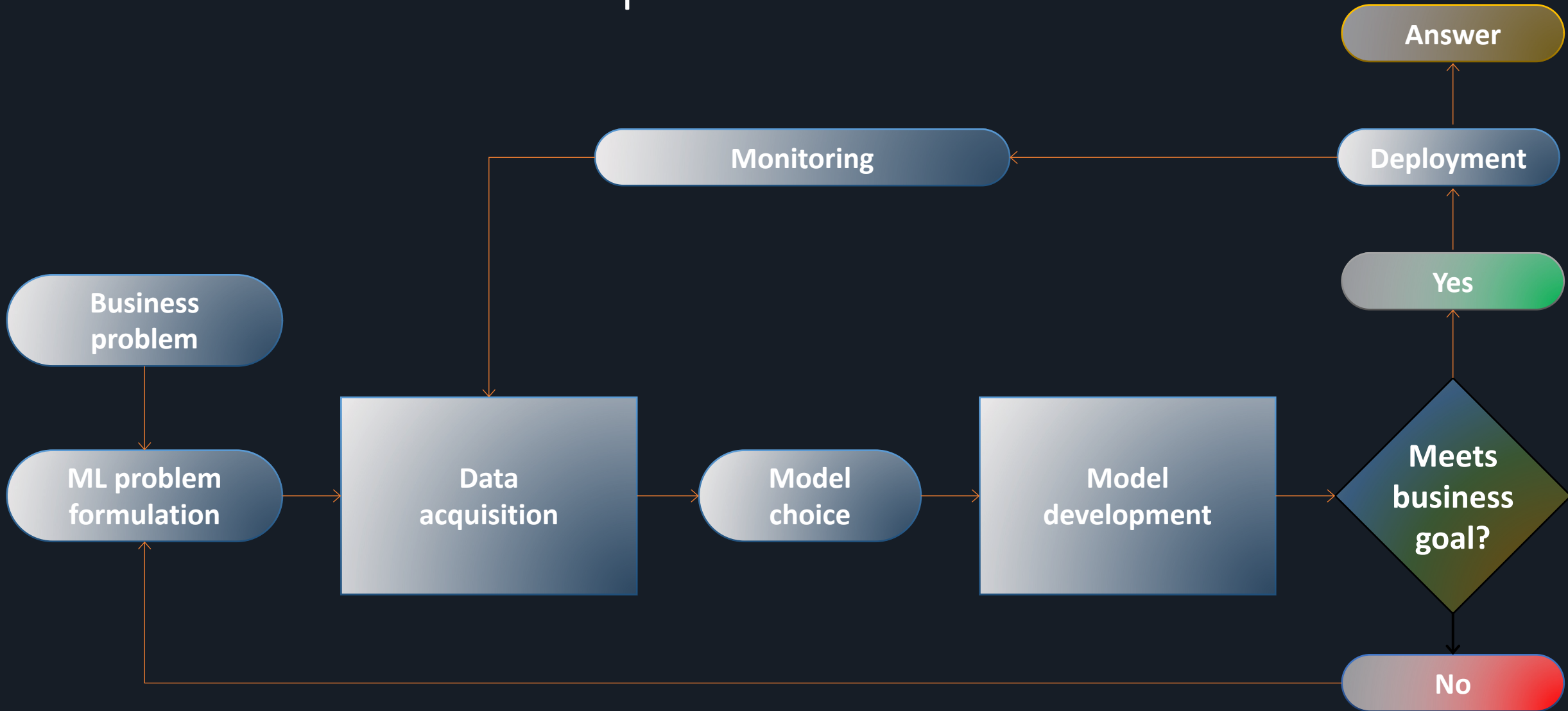
Example: AWS AI Service Cards

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

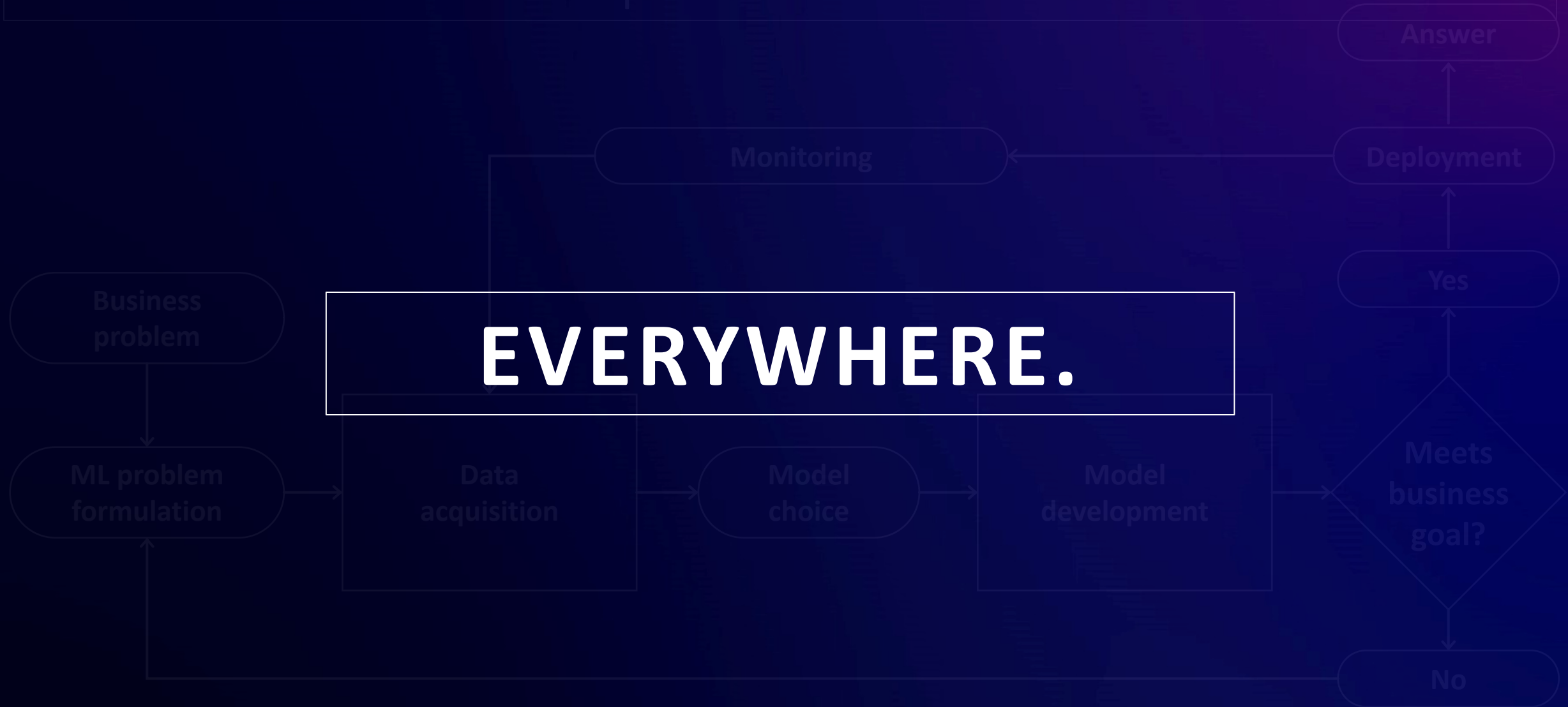
The image displays three overlapping screenshots of AWS AI Service Cards, illustrating their structure and content. Each card follows a consistent layout:

- Header:** "Machine Learning / Responsible Machine Learning"
- Introductory Paragraph:** "This AI Service Card explains the use cases for which the service is intended, how machine learning (ML) is used by the service, and key considerations in the design and use of the service. We do not assume the reader is an expert in artificial intelligence (AI), ML, or data science. We do recommend that AWS customers assess the performance of any ML service on their own content for each use case they need to solve. For more information, please see the AWS Responsible Use of Machine Learning guide and additional information referenced at the end."
- Version Note:** "This Service Card (version 10/23/2022) applies to Version 2204 of Amazon Transcribe::StartTranscriptionJob, released for GA on August 2022." (Note: The image shows a typo in the original text: "StartTranscription Job").
- Table of Contents:**
 - Overview
 - Intended use cases and limitations
 - Design of Textract AnalyzeID
 - How customers can optimize performance
 - Further information
- Overview Section:**
 - Overview:** "The Amazon Textract service extracts printed text, handwriting, and data from documents. Within this service, the AnalyzeID feature reads and extracts text data from identity documents such as US driver's licenses and passports, making it easier for customers to automate and expedite their document processing. Analyze ID can extract explicit text fields, where a key ("Date of Issue") appears on the document and is aligned with its value ("03/18/2018"), and implied text fields that may not have explicit keys appearing next to them ("María" appears in the center of a license, but is not marked as "First Name")."
 - How customers can optimize performance:** "AnalyzeID operates on the text that appears in an identity document to return explicit and implied key-value pairs. The service normalizes key-value pairs into a common taxonomy of 21 known field names so that customers can compare information across ID types. For example, the service extracts the LIC# of a driver's license and Passport No. from a US passport, labeling both as "Document ID Number."

Where to consider responsible AI?



Where to consider responsible AI?



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

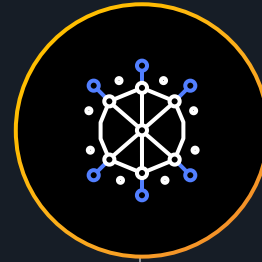
**Building
awareness**



**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

