AWS re:Invent

DECEMBER 2 – 6, 2024 | LAS VEGAS, NV

AIM342

Responsible generative AI: Evaluation best practices and tools

Alessandro Cerè, PhD
(he/him)
Principal Solutions Architect, Model Eval
Amazon Web Services

Mathew Monfort, PhD
(he/him)
Senior Applied Scientist, Responsible AI
Amazon Web Services

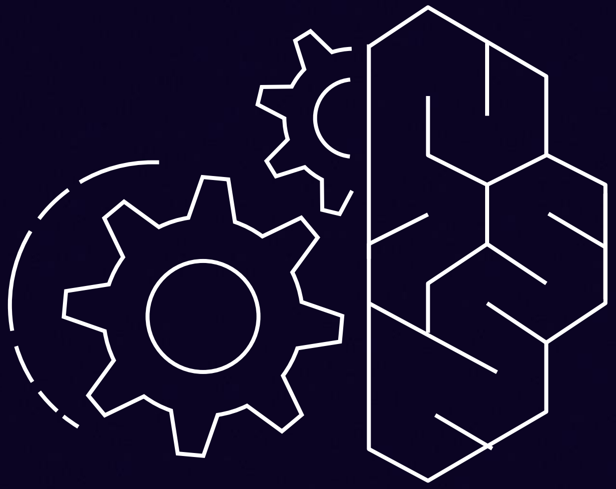
Agenda

01 Fundamentals of gen AI evaluation

02 Evaluating large language models

03 Assessing risk

04 Establishing release confidence



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

Foundation models are broad and open-ended

“Traditional” AI models

Focused use cases

Generative AI models

Multiple use cases

“Should I give a loan to this person?”	“Generate the report of a meeting.”

Building applications powered by gen AI

START SIMPLE...

Complexity of gen AI application increases

Retrieval Augmented Generation

Tackles:
Hallucination

Search a trusted document store for relevant info to user's query. Insert results into the prompt for final answer generation.

Chain-of-thought and agents

Tackles:
Reasoning, hallucination

Explain external tools available to the model (e.g., calculator, databases, etc.) and ask it to tackle the query step by step.

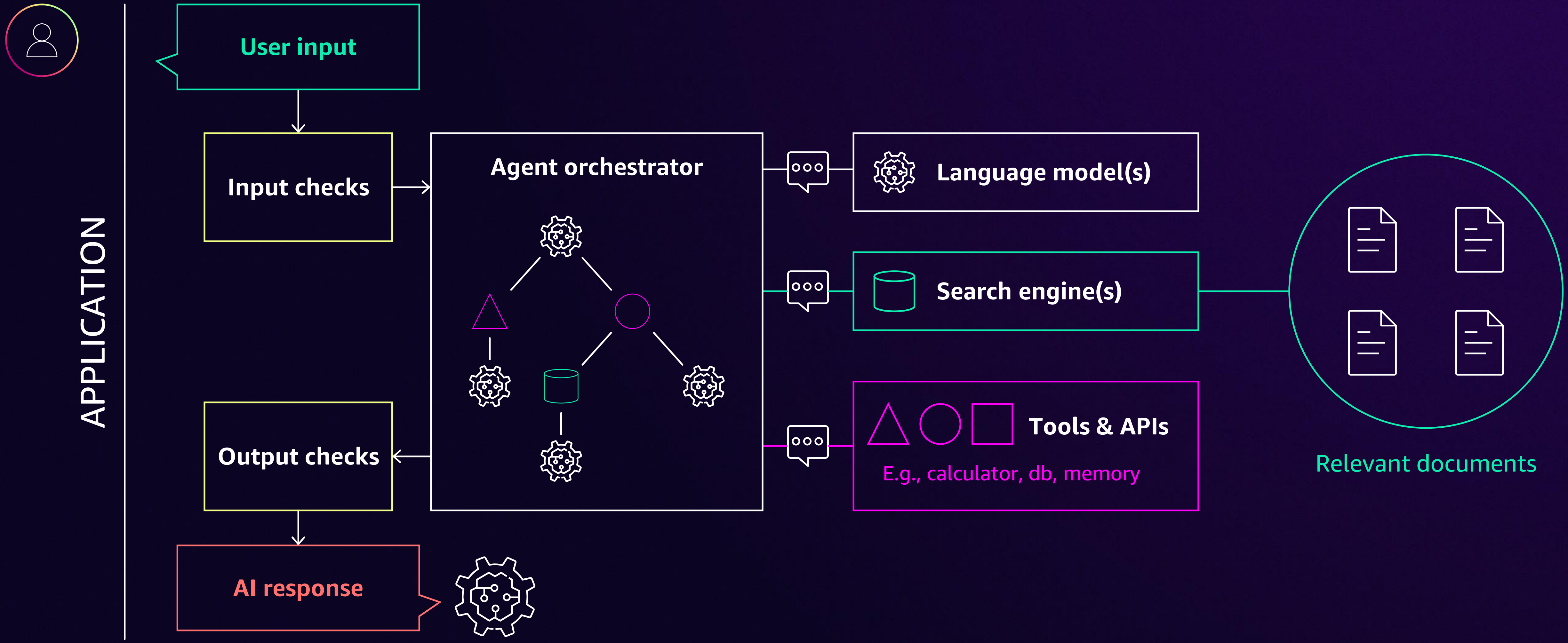
Constitutional AI and guardrail models

Tackles:
Unpredictable behavior

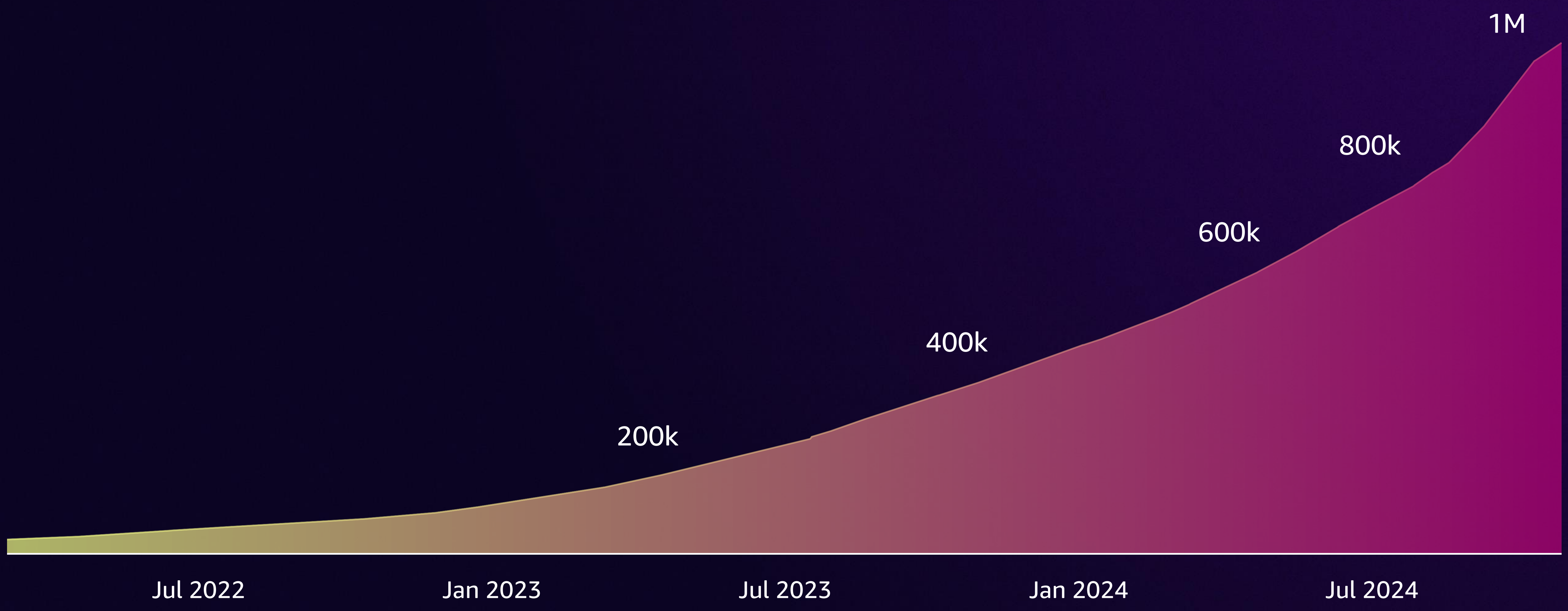
Use companion models or extra LLM calls as an additional line of defense against toxic, off-brand, or unacceptable responses.

Building applications powered by gen AI

... AND THEN GROW



Number of LLMs keeps increasing



Data from https://huggingface.co/datasets/cfahlgren1/hub-stats

A balancing act for solution builders

Analysis paralysis

"Why are we spending so much fine-tuning when usage is so low?"

"Well, yes, it's pretty slow, but that's our enterprise guardrails."

"How can you be SURE it will never give a wrong answer?"

Maintain and upgrade

Monitoring in live

Prototype to production

Production incidents

"But upgrading the model might break everything!"

"We've had complaints: How can I find out what's wrong??"

"It passes the 10 test cases – it'll be fine!"

The responsible way forward

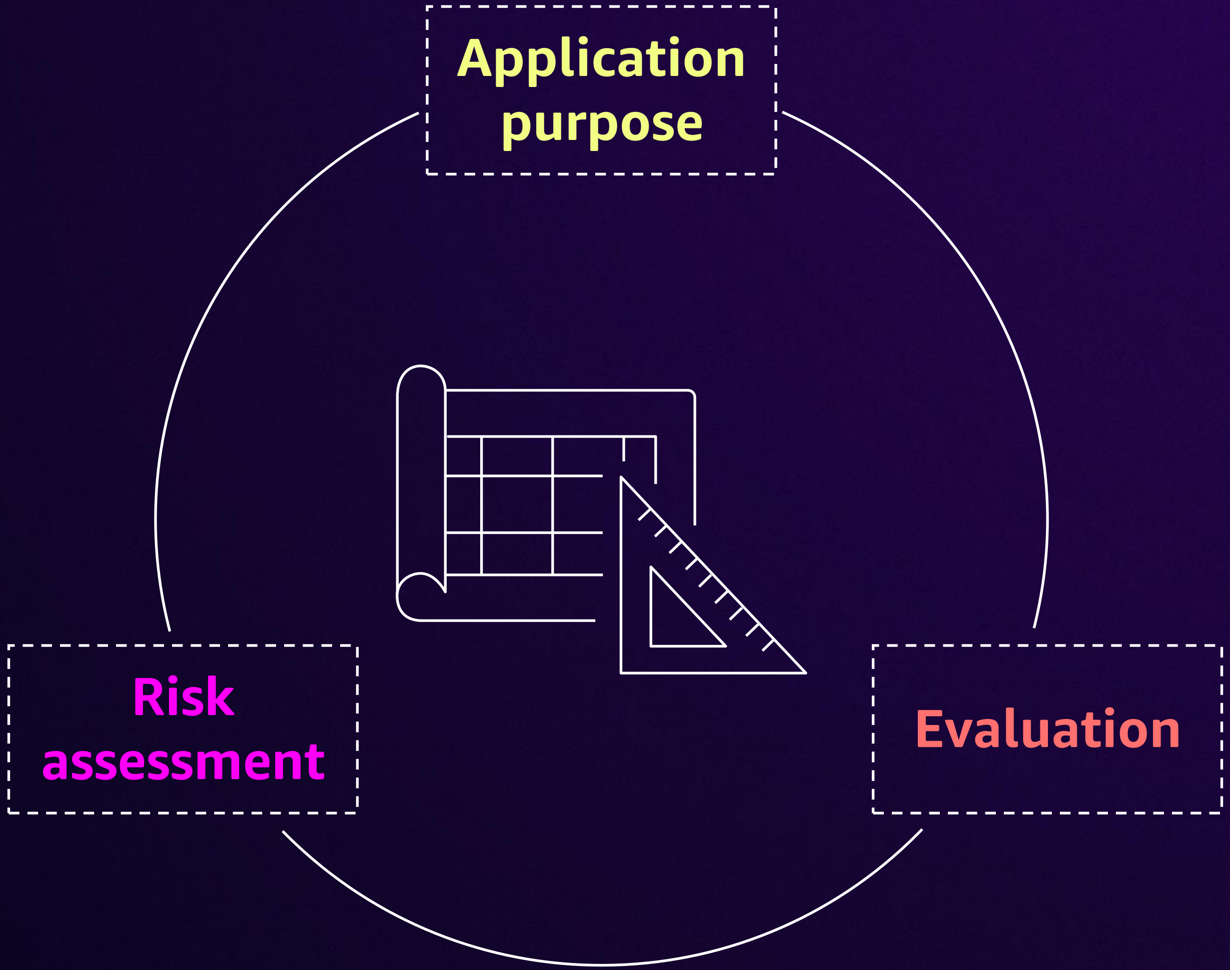
Test and evaluate the application until you are confident about its quality and that the associated risks are acceptable

And then keep doing it

Fundamentals of gen AI evaluation

We need a plan

DEFINE THE STRATEGY FOR A CONFIDENT DEPLOYMENT



Key aspects to evaluate

AND CORRESPONDING ACCEPTANCE CRITERIA IN OUR PLAN

Quality

Performing as or better than expected

Latency

Fast enough for its purpose

Cost

$

Confidence

Risks are acceptable

Responsible AI: Achieving confidence

Controllability	Privacy and security	Safety	Fairness
Veracity and robustness	Explainability	Transparency	Governance

Fundamental elements for effective evaluations

Models and applications

API
Managed endpoint
Local deployment

Data

Public
Private
Synthetic

Input engineering

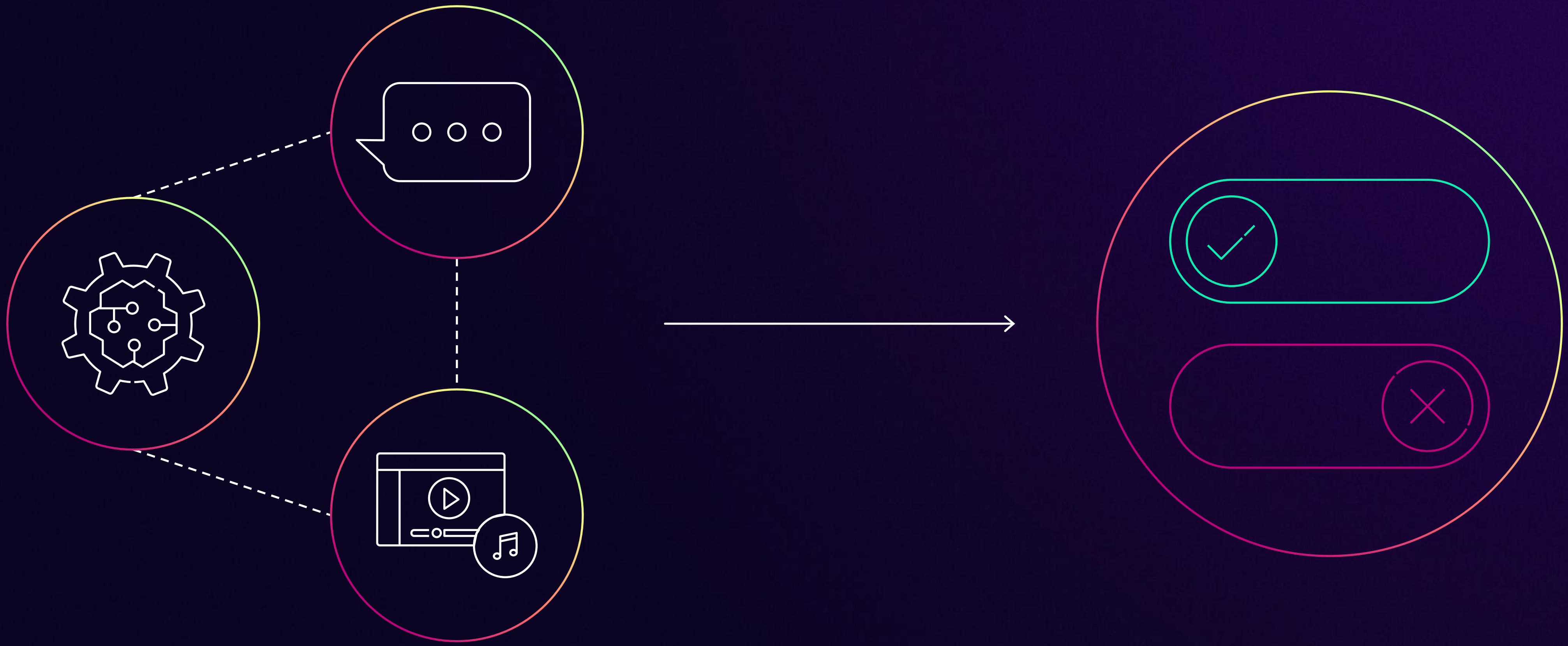
Prompt templates
Input preprocessing
Context augmentation

Evaluation tools

Results tracking

How to evaluate an LLM

FROM GENERATED CONTENT TO METRICS



How to evaluate an LLM

IN A ROBUST, SCALABLE, COST-EFFECTIVE WAY

Human review
(manual)

Heuristic metrics
(automated)

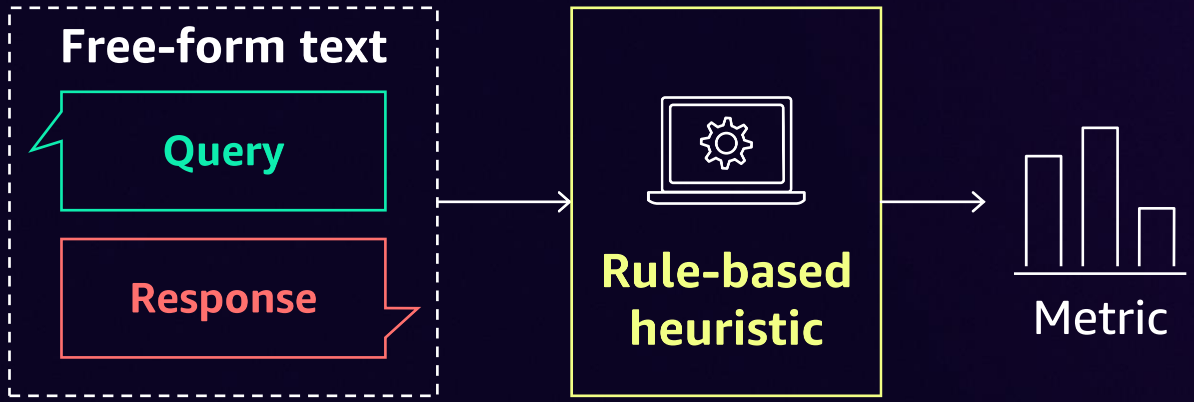
AI critique
(LLMs judging LLMs)

Performance
(speed and cost)

***Also produces new data!**

Evaluation with heuristics and LLM-as-a-judge

Rule-based heuristics

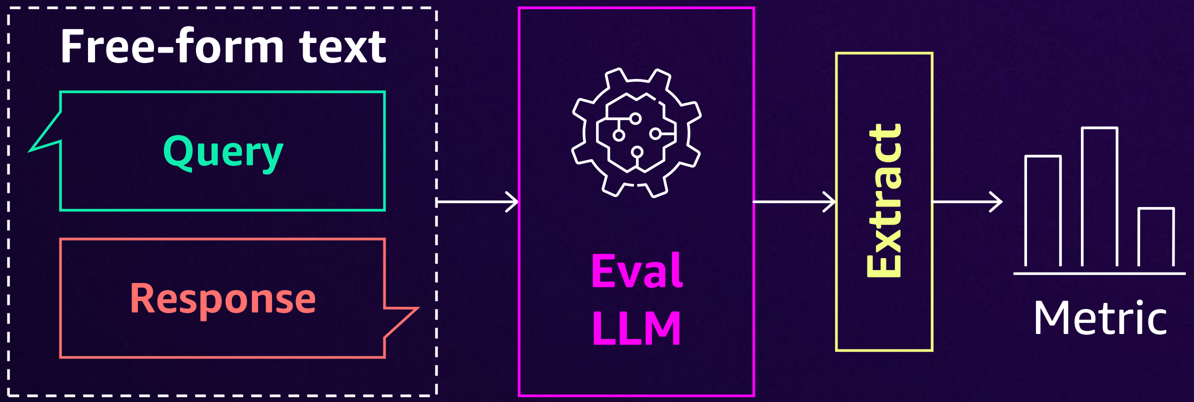


Fast, scalable, cheap to run

Leverage standard metrics (F1, ROUGE…) or helper models (sentiment, toxicity…)

Will the metrics align well with human preferences?

LLM-based critique



Flexible, customizable checks

Checking an answer usually easier than creating it

Is it biased by the evaluator?

Affordable to run?

Common metrics

Quantitative ⟷ **Qualitative**

Accuracy
Precision
Recall
F1
Perplexity

ROUGE
METEOR
BLEU
WER

Relevance
BERTScore
MoverScore

Fluency
Coherence
Semantic similarity

Latency and cost

Time-to-last-token
Time-to-first-token
Output tokens per second

Common metrics

Quantitative ⟷ **Qualitative**

Accuracy
Precision
Recall
F1
Perplexity

ROUGE
METEOR
BLEU
WER

Relevance
BERTScore
MoverScore

Fluency
Coherence
Semantic similarity

RAG

- Contextual recall
- Contextual relevancy
- Contextual precision
- Answer relevancy
- Faithfulness
- Noise sensitivity

Latency and cost

Time-to-last-token
Time-to-first-token
Output tokens per second

Common metrics for responsible AI

Hallucination

Seq-Logprob

G-EVAL

Semantic coherence

Semantic similarity

Fairness

Maximum disparity

Min-max

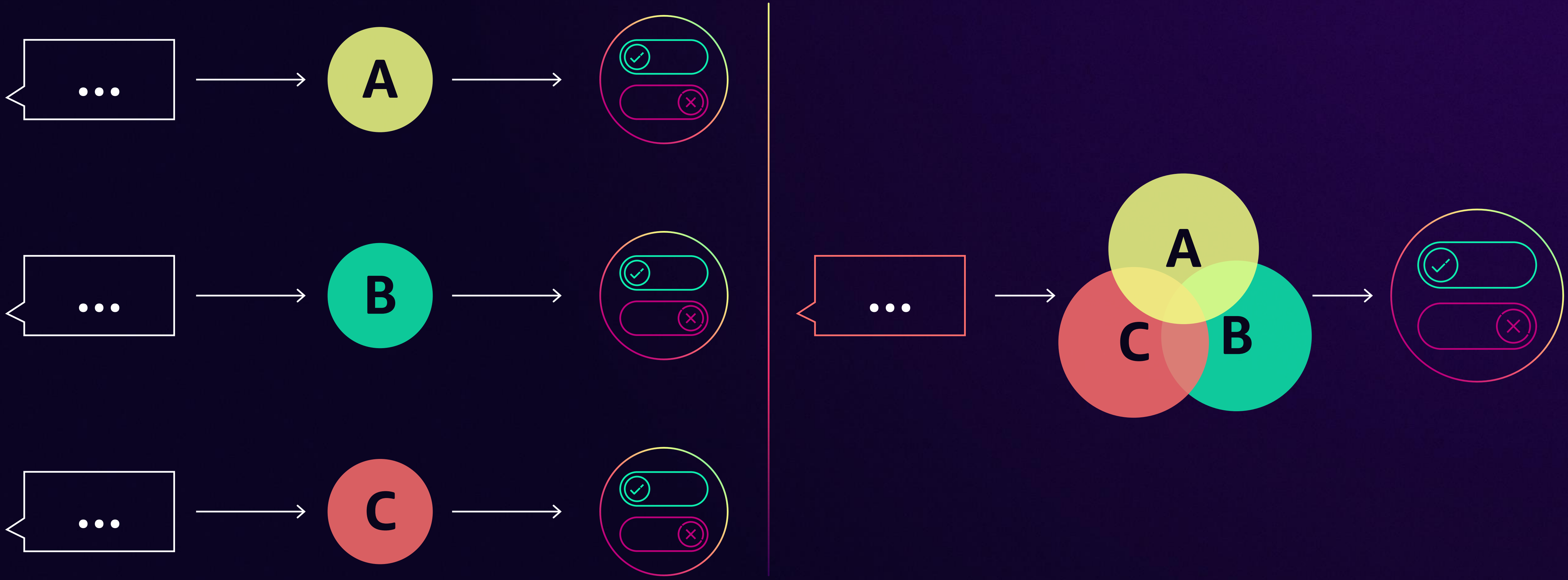
Equalized odds

Predictive parity

PII leakage

Safety/toxicity

Components and holistic evaluation



Generative AI applications and evaluation

Define purpose and feasibility
- Identify business metrics
- Identify user base

Build with production in mind
- Evaluation datasets specific for the use case
- Component level evaluation

Establish confidence that we're ready to go live
- Holistic evaluation
- Reports generation

Maintain
- Monitor
- Optimize

Design and prototype → **Development** → **Validation** → **Operation**

Playground

Developers

Programmatic and human review

Dedicated workforce

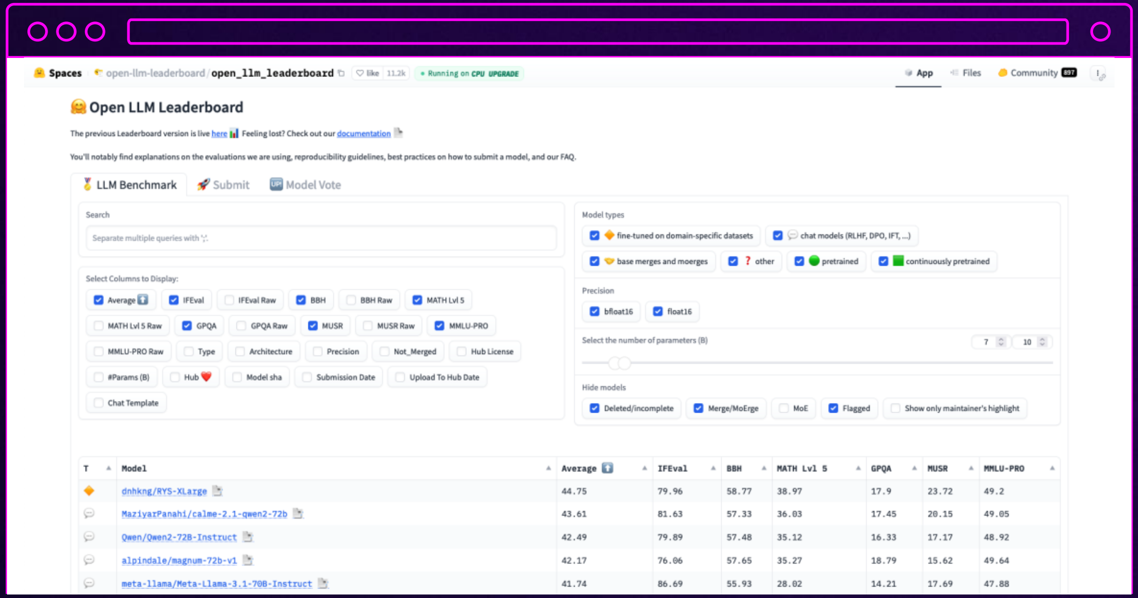
Users

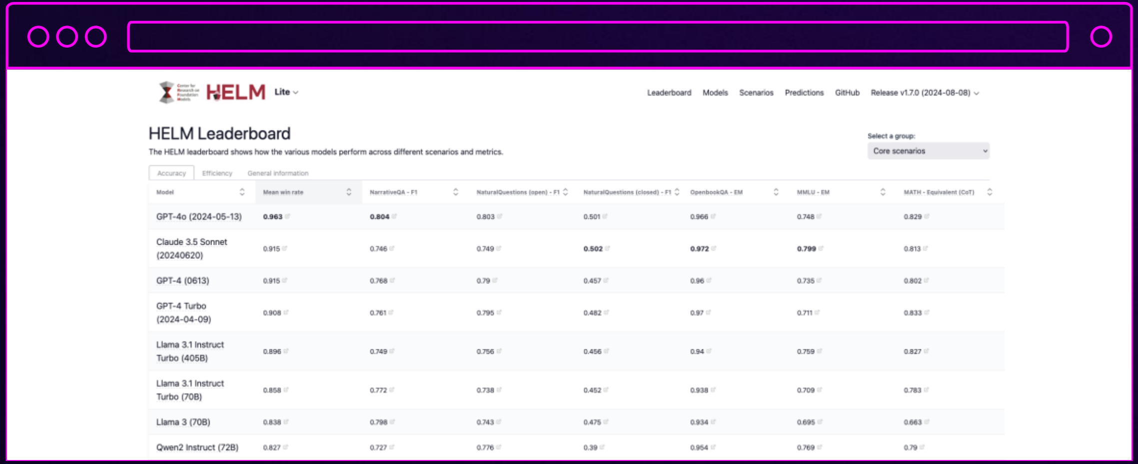
Evaluating LLMs

Public leaderboards can help to short-list models

- E.g., **LMSYS Chatbot Arena**
- Hugging Face Leaderboard
- HELM
- BIG-Bench
- Exam scores
- Papers/reports
- etc.

. . . but won't fully reflect your use case or include all models





Measuring inference speed (and cost)

Inference speed

Response time and total available throughput are critical for real-time applications like virtual assistants

- Overall response vs. time-to-first token (streaming)
- Capacity, quotas, and graceful degradation

Cost to run

Could a smaller/cheaper model deliver good-enough responses at a fraction of the cost?

- Token vs. instance-based pricing
- Operating vs. development cost

https://github.com/aws-samples/foundation-model-benchmarking-tool

https://github.com/awslabs/llmeter

Amazon Bedrock Model Evaluation

Evaluate, compare, and select the best foundation model for your use case

New:
Public API
Evaluate custom models
Evaluate distilled models
Evaluate imported models
Evaluate prompt routers
Use an LLM-as-a-judge (Preview)

1. Use curated datasets or bring your own for tailored results
2. Use automatic (algos or LLMs) or human evaluation methods
3. Leverage your in-house team or AWS managed reviewers
4. Predefined and custom metrics
5. Get results in just a few clicks

Choice of evaluation methods

Automatic evaluation

Accuracy | Robustness | Toxicity

LLM-as-a-judge

Correctness | Completeness | Helpfulness

Relevance | Coherence | Readability

Human evaluation

Creativity | Style | Tone

Accuracy | Consistency | Brand voice

Algorithms

BERTScore | Classification accuracy
F1 | Real-world knowledge score

LLM Reasoning

Multistep reasoning | Few-shot learning
Correlation with expert human evaluators

Rating Methods

Thumbs up/down | 5-point Likert scales
Binary choice buttons | Ordinal ranking

Amazon SageMaker

Foundation model evaluation

Powered by AWS open-source fmeval

- Scale human evaluation with work-team management and pre-built labelling portal
- Automate evaluation with configurable, fully managed jobs
- Select from built-in standard datasets or bring your own
- Build fully custom evaluation scripts using open source fmeval library on SageMaker

No-code automatic evaluation jobs with AWS

USE PRE-BUILT HEURISTIC METRICS IN AMAZON BEDROCK AND AMAZON SAGEMAKER

Model → **Automatic model evaluation job** On Amazon Bedrock or Amazon SageMaker →

No data required

No manual labelling required

Select from pre-built metrics and datasets

Dataset	Accuracy	Robustness	Toxicity
BOLD		●	●
Real Toxicity			●
TREX	●	●	
WikiText2		●	
Gigaword	●	●	●
XSUM	●	●	●
BoolQ	●	●	●
NaturalQuestions	●	●	●
Trivia QA	●	●	●
Women’s Ecomm. Clothing Reviews	●	●	

aws

No-code automatic evaluation jobs with AWS

USE PRE-BUILT HEURISTIC METRICS IN AMAZON BEDROCK AND AMAZON SAGEMAKER

Model

No data required

No manual labelling required

Automatic model evaluation job

On Amazon Bedrock or Amazon SageMaker

Select from pre-built metrics and datasets

Da	Accuracy	Robustness	Toxicity
BOLD		●	●
NaturalQuestions	●	●	●
Trivia QA	●	●	●
Women's Ecomm. Clothing Reviews	●	●	

Custom datasets are optional, but strongly recommended!

- Standard benchmark datasets may not be similar to **your** use case
- Public datasets may leak into the training of newer models

Task-specific data gives the most useful metrics!

LLM-as-a-judge metrics

01 Correctness

02 Completeness

03 Faithfulness

04 Helpfulness

05 Coherence

06 Relevance

07 Following instructions

08 Professional style and tone

09 Readability

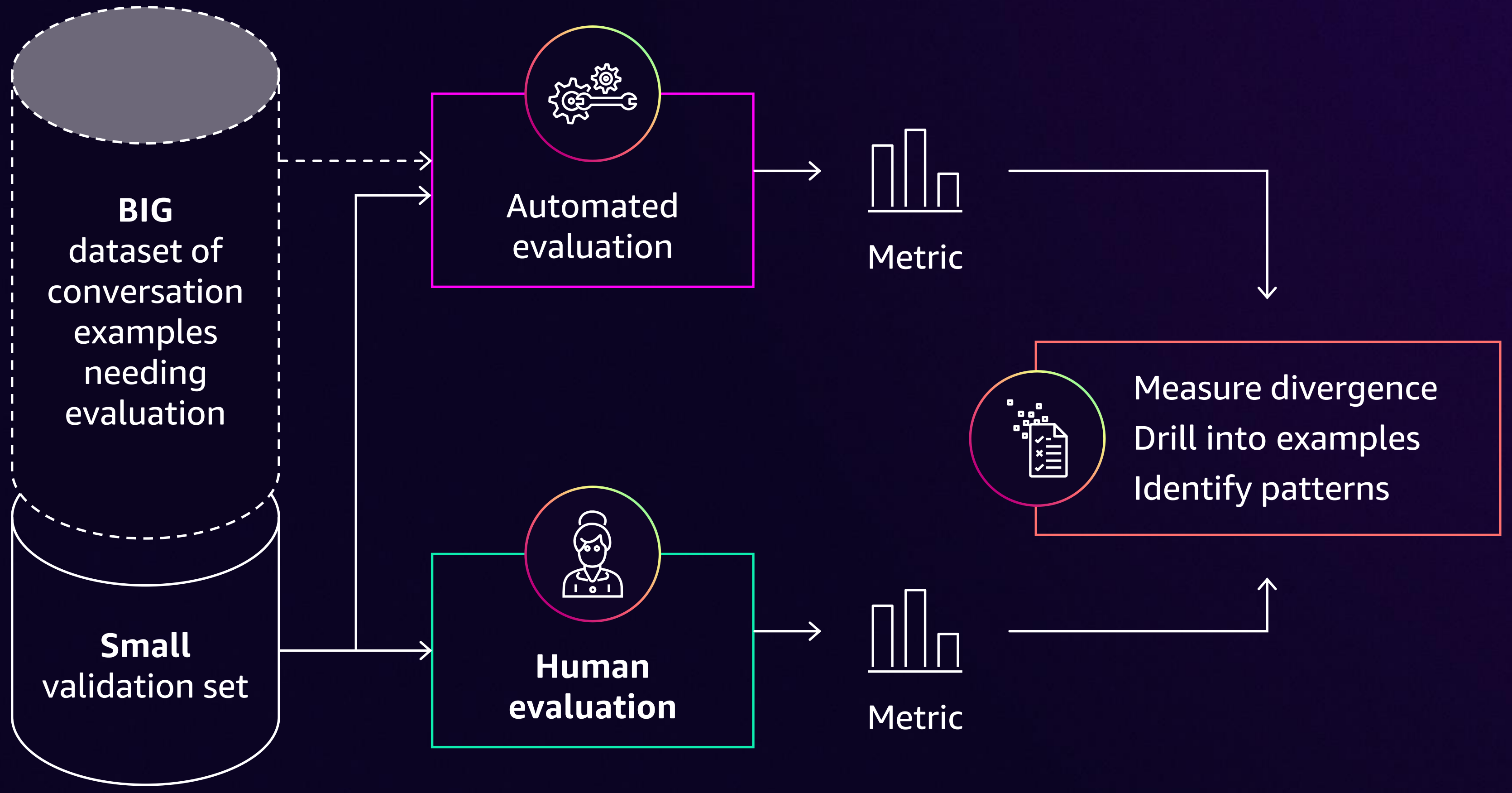
10 Harmfulness

11 Stereotyping

12 Answer refusal

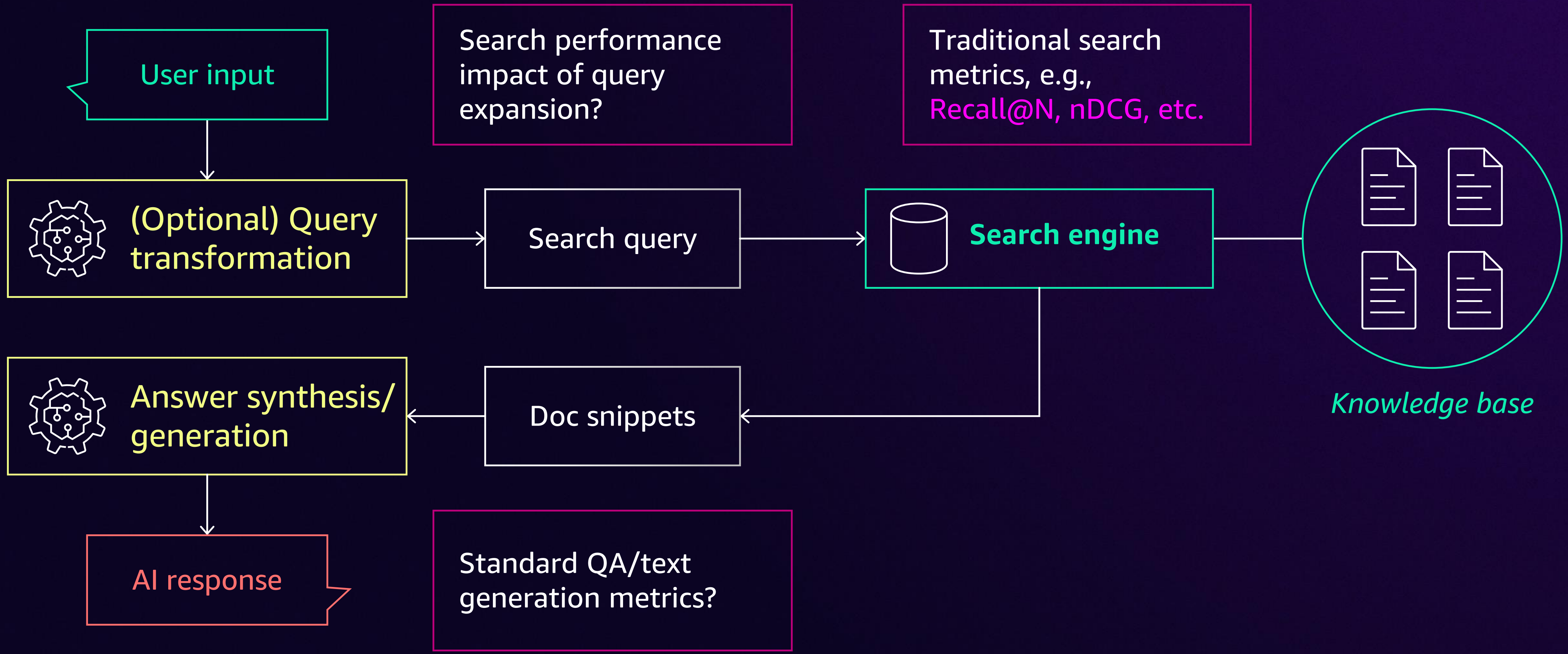
Ground truth response is optional

Don’t trust, but measure fidelity of automatic metrics



- ✓ **Scalable evaluation for rapid prototyping and prompt engineering**
- ✓ **Trusted performance verified by real human reviewers**

RAG evaluation



Preview

RAG evaluation with Amazon Bedrock Knowledge Bases

Evaluate your full Knowledge Base stack to optimize your RAG application

1. Bring your own datasets for tailored results
2. Evaluate retrieval alone or retrieval + generation combined with a choice of LLM-as-a-judge
3. Built-in metrics for quality and responsible AI, compatible with Bedrock Guardrails
4. Compare across multiple evaluation jobs
5. Get results in just a few clicks

Choice of evaluation metrics

Retrieval

- Context Coverage
- Context Relevance

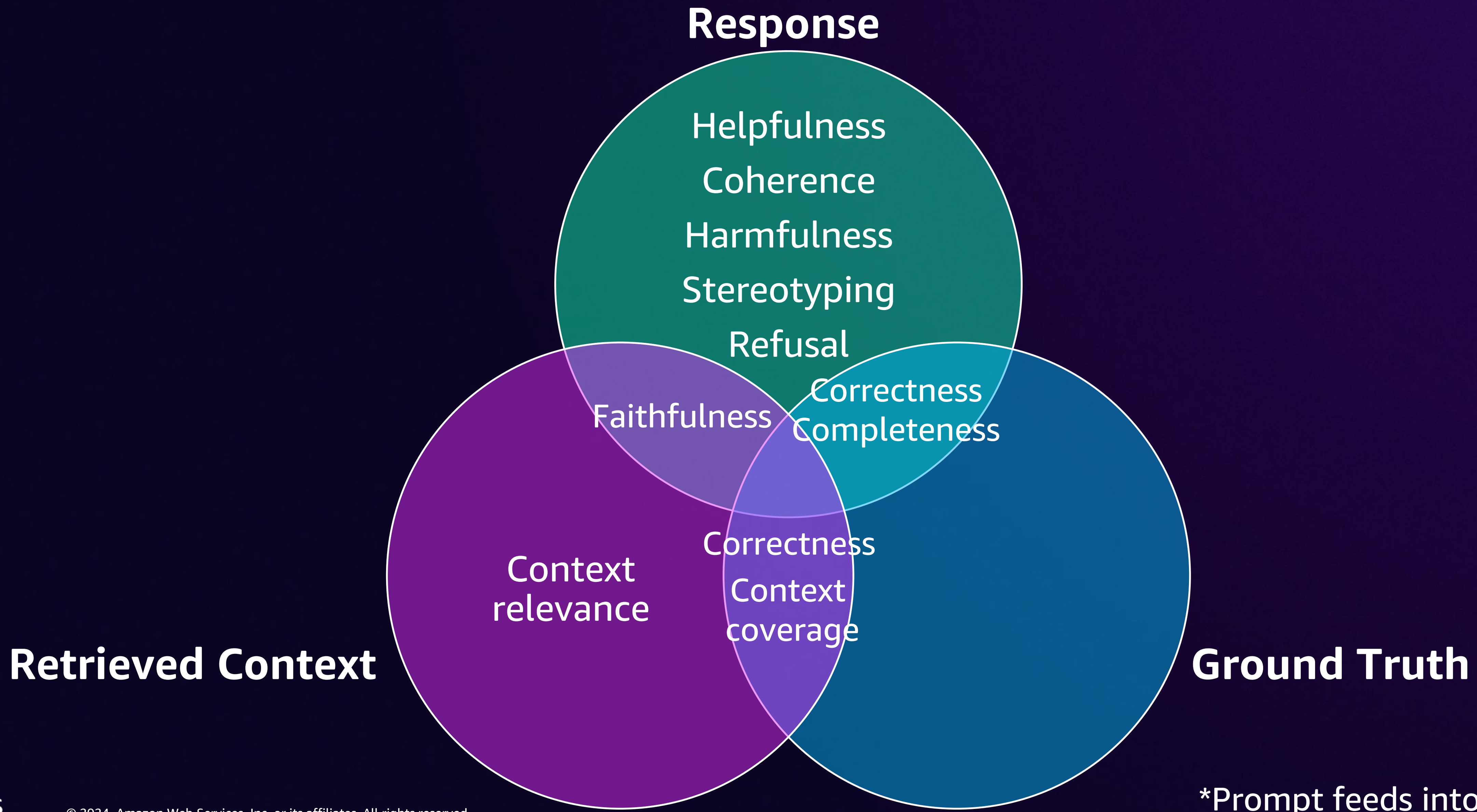
Retrieval + Generation

- Correctness
- Completeness
- Helpfulness
- Coherence
- Faithfulness

Responsible AI

- Harmfulness
- Stereotyping
- Refusal

RAG evaluation metrics



*Prompt feeds into all metrics
*Ground truth optional for correctness and completeness

RAG evaluation

Ragas

- Context precision
- Context recall
- Context entities recall
- Noise sensitivity
- Response relevancy
- Faithfulness

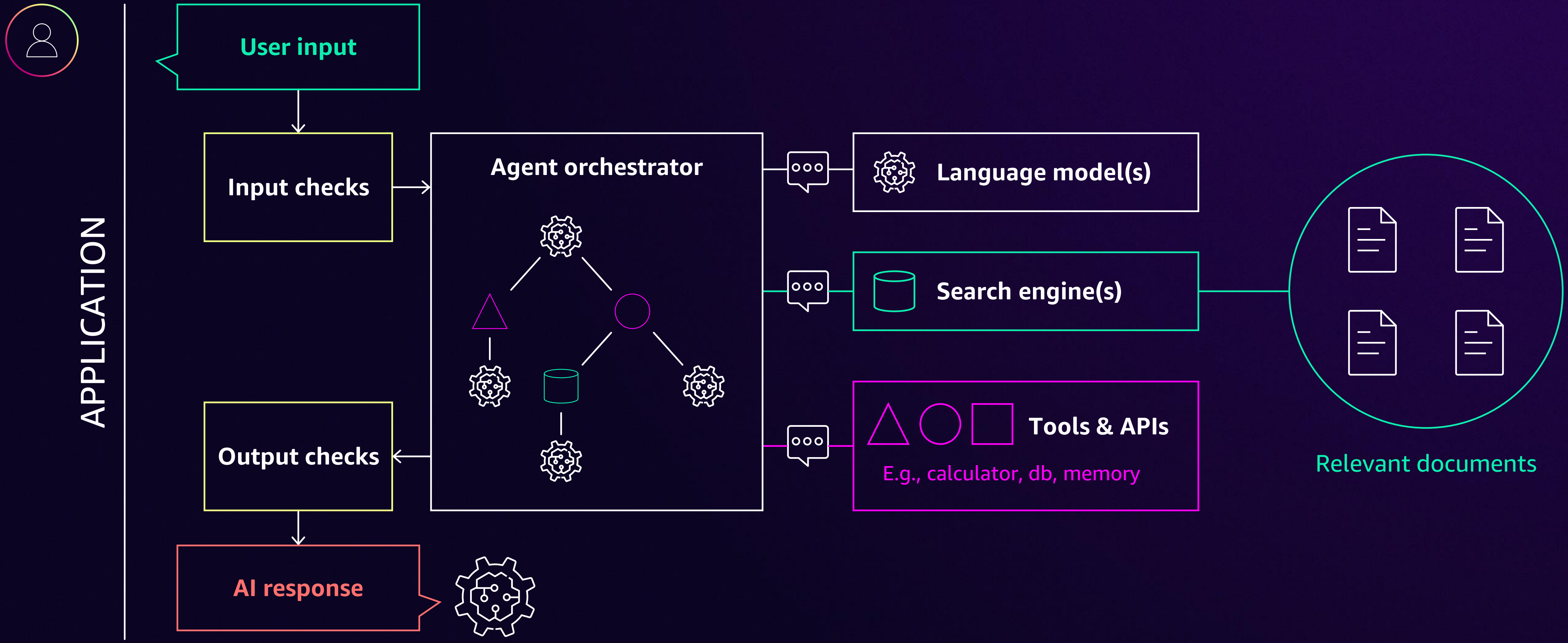
export results ¶

```
df = result.to_pandas()
df.head()
```

	question	ground_truths	answer	contexts	context_relavency	faithfulness	answer_relevancy
0	How to deposit a cheque issued to an associate...	[Have the check reissued to the proper payee.J...	\nThe best way to deposit a cheque issued to a...	[Just have the associate sign the back and the...	0.867	1.0	0.922
1	Can I send a money order from USPS as a business?	[Sure you can. You can fill in whatever you w...	\nYes, you can send a money order from USPS as...	[Sure you can. You can fill in whatever you w...	0.855	1.0	0.923
2	1 EIN doing business under multiple business n...	[You're confusing a lot of things here. Compan...	\nYes, it is possible to have one EIN doing bu...	[You're confusing a lot of things here. Compan...	0.768	1.0	0.824
3	Applying for and receiving business credit	["I'm afraid the great myth of limited liabili...	\nApplying for and receiving business credit c...	[Set up a meeting with the bank that handles y...	0.781	1.0	0.830

https://github.com/explodinggradients/ragas

Evaluating agents



Agent evaluation

A CONVENIENT AND SCALABLE APPROACH TO TESTING THE CAPABILITIES OF VIRTUAL AGENTS

Built-in support for popular services including Amazon Bedrock Agents, Amazon Q for Business, and Amazon SageMaker endpoints

Orchestrate concurrent, multi-turn conversations with your agent while evaluating its responses

Integrate into CI/CD pipelines to automate agent testing

Generate test summary for performance insights

Testing with AWS Labs agent-evaluation

1. Configure test cases with YAML

```yaml
evaluator:
 model:
 claude-3
target:
 type: bedrock-agent
 bedrock_agent_id: ABCDEFGHIJ
 bedrock_agent_alias_id: DRAFT

tests:
 amazon_followup:
 steps:
 - Ask the agent how big the Amazon rainforest is
 - Ask the exact question "how many trees are in it?"
 expected_results:
 - The agent says the rainforest is 5,500,000 square km
 - The agent says 390 billion trees
 ...
```

https://awslabs.github.io/agent-evaluation

2. Run in CLI or CI/CD

```
[9]: !agenteval run --plan-dir . --num-threads 2 --verbose
     [03:19:53] INFO     Starting 3 tests with 2 threads.            ]8;id=690779;file://
     1;file:///opt/conda/lib/python3.10/site-packages/agenteval/runner/runner.py#84\84]8;;\
     running...                                          100% 0:00:00m 0:00:03
     [03:20:11] INFO     3 passed.                                   ]8;id=165083;file:///
     2;file:///opt/conda/lib/python3.10/site-packages/agenteval/runner/runner.py#108\108]8;;\
                INFO     Completed in 17.96 seconds.                 ]8;id=834645;file://
     3;file:///opt/conda/lib/python3.10/site-packages/agenteval/runner/runner.py#88\88]8;;\
                INFO     normandy_country...PASSED                   ]8;id=199294;file://
     2;file:///opt/conda/lib/python3.10/site-packages/agenteval/runner/runner.py#96\96]8;;\
                INFO     amazon_followup...PASSED                    ]8;id=684484;file://
```

3. Explore human-readable reports

Test Summary

This document provides a summary of the tests executed by Agent Evaluation.

> ⚠ This tool utilizes generative AI to assess virtual agents and its evaluations may contain errors. **Please thoroughly examine the results below prior to deciding whether to implement an agent.**

Tests

- normandy_country
- amazon_followup
- aws_lambda_noanswer

normandy_country

Steps

1. Ask the agent what country Normandy is located in

Expected results

1. The agent says Normandy is in France

Conversation

```
[USER] In what country is Normandy located?
[AGENT] Normandy is a region located in northern France. The Normans were originally Norse raiders and pirates from Denmark, Iceland and Norway who settled in the region of Normandy in the 10th century under their leader Rollo. They assimilated with the native Frankish and Roman-Gaulish populations over time.
```

Result All of the expected results can be observed in the conversation.

Workshop: Evaluate LLMs and optimize their applications on AWS

Guided introduction to some of the tools for evaluating LLM models and applications on AWS

Assessing risk and establishing launch confidence

Key aspects to evaluate

AND CORRESPONDING ACCEPTANCE CRITERIA IN OUR PLAN

Quality

General performance

Latency

Fast enough for its purpose

Cost

$

Confidence

Risks are acceptable

Public leaderboards can help to short-list models

E.g., **LMSYS Chatbot Arena**

Hugging Face Leaderboard

HELM

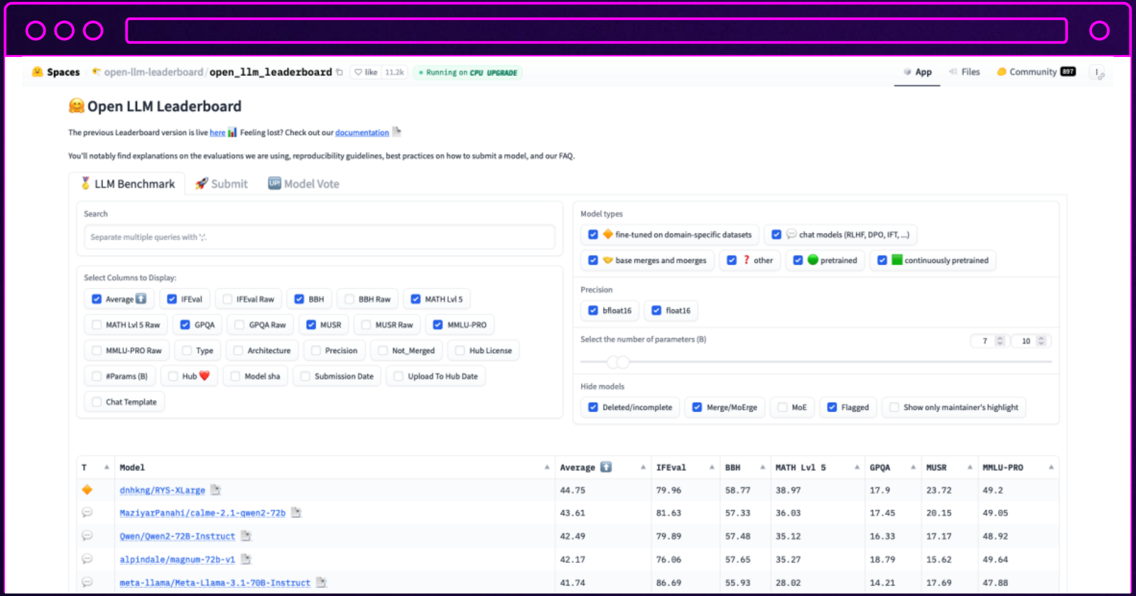
BIG-Bench

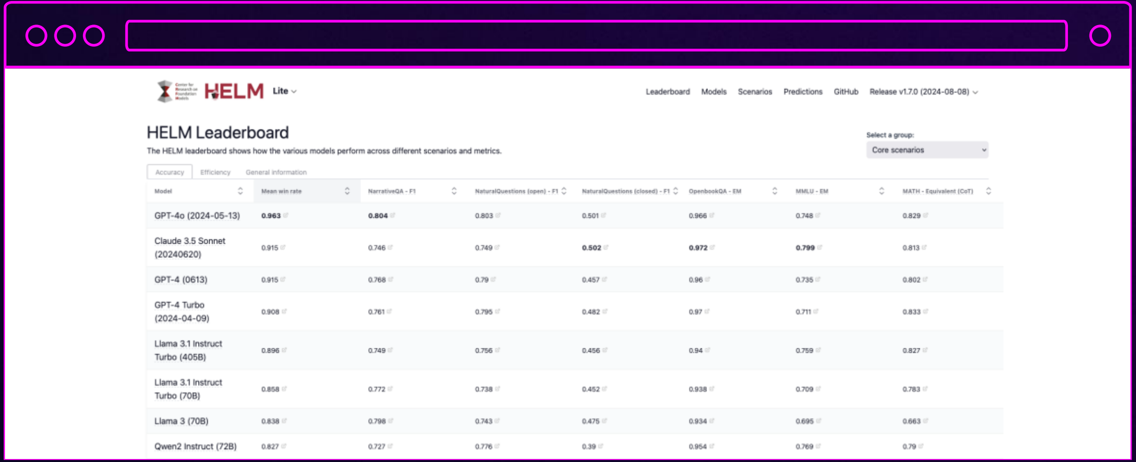
Exam scores

Papers/reports

etc.

. . . but won’t fully reflect your use case and specific customer risks





Establishing launch confidence

CONFIDENCE

How do we establish confidence in our release?

Establishing launch confidence

What is different compared to prototype evaluation?

CONFIDENCE

How do we establish confidence in our release?

Establishing launch confidence

Standard application properties

Use case accuracy	Feature set	Latency	Cost	Uptime

Establishing launch confidence

Standard application properties

Use case accuracy	Feature set	Latency	Cost	Uptime

Properties of a responsible AI application

Controllability	Security and privacy	Safety	Fairness
Veracity and robustness	Explainability	Transparency	Governance

Establishing launch confidence

Standard application properties

Use case accuracy	Feature set	Latency	Cost	Uptime

Properties of a responsible AI application

Controllability	**Security and privacy**	**Safety**	**Fairness**
Veracity and robustness	**Explainability**	**Transparency**	**Governance**

Establishing launch confidence

Using AI to recommend music

Using AI to identify a tumor on an x-ray

Establishing launch confidence

Using AI to recommend music

Using AI to identify a tumor on an x-ray

How does the potential risk change?

Establishing launch confidence

Using AI to recommend music

Using AI to identify a tumor on an x-ray

How does this affect our confidence in a successful release?

Establishing launch confidence

Responsible evaluation strategy

Establishing launch confidence

Responsible evaluation strategy

Define
use case

Establishing launch confidence

Responsible evaluation strategy

Define use case

All choices flow from the use case

Define use case

Use case definition → Risks and goals

Define use case

Use case definition → Risks and goals

Define application use cases narrowly

Define use case

Write a product description

Audience
Broad demographic

Possible risks
Veracity, toxicity

Consequences
Brand damage, lost sales

Define use case

Write a product description

Audience
Broad demographic

Possible risks
Veracity, toxicity

Consequences
Brand damage, lost sales

Very general

Define use case

Write a product description

Audience
Broad demographic

Possible risks
Veracity, toxicity

Consequences
Brand damage, lost sales

Very general

Leads to unknowns

Define use case

Write a product description

Audience
Broad demographic

Possible risks
Veracity, toxicity

Consequences
Brand damage, lost sales

Persuade audience X to buy product Y

Audience
Narrow demographic

Possible risks
Veracity, toxicity, stereotyping, unwanted bias

Consequences
Representative harm, brand damage, lost sales

Define application use cases narrowly

Reduces unknowns and leads to actionable evaluation steps

Establishing launch confidence

Responsible evaluation strategy

Define
use case

Assess
risk

Assess risk

Controllability	Privacy and security	Safety	Fairness
Veracity and robustness	Explainability	Transparency	Governance

Assess risk

An event's probability of occurring

(likelihood)

×

Magnitude or degree of consequences

(severity)

Source: National Institute of Standards and Technology, [AI Risk Management Framework 1.0]

Assess risk

An event's probability of occurring

(likelihood)

×

Magnitude or degree of consequences

(severity)

→

Confidence risk

Source: National Institute of Standards and Technology, [AI Risk Management Framework 1.0]

Example risk rating matrix

Likelihood

Severity	Highly unlikely	Unlikely	Possible	Likely	Almost certain/ Frequent
Extreme	High	High	Critical	Critical	Critical
Major	Medium	Medium	High	High	Critical
Moderate	Very Low	Low	Medium	Medium	High
Low	Very Low	Very Low	Low	Low	Medium
Very low	Very Low	Very Low	Very Low	Very Low	Low

Assess risk

Risk assessment guidance

Learn more

Establishing launch confidence

Responsible evaluation strategy

Define use case | Assess risk | **Choose metrics**

Metrics are dependent on the risk likelihood being measured

Each risk dimension has specific considerations

Choose metrics

Quantitative ⟷ Qualitative

Accuracy
Precision
Recall
F1
Perplexity

ROUGE
METEOR
BLEU
WER

Relevance
BERTScore
MoverScore

Fluency
Coherence
Semantic similarity

Latency
Time-to-last-token
Time-to-first-token
Output tokens per second

Cost

RAG

- Contextual recall
- Contextual relevancy
- Contextual precision
- Answer relevancy
- Faithfulness
- Noise sensitivity

Choose metrics

Automatic evaluation

Accuracy | Robustness | Toxicity

LLM-as-a-judge

Correctness | Completeness | Helpfulness
Relevance | Coherence | Readability

Human evaluation

Creativity | Style | Tone
Accuracy | Consistency | Brand voice

Algorithms

BERTScore | Classification accuracy
F1 | Real-world knowledge score

LLM Reasoning

Multistep reasoning | Few-shot learning
Correlation with expert human evaluators

Rating Methods

Thumbs up/down | 5-point Likert scales
Binary choice buttons | Ordinal ranking

Choose metrics

Veracity and robustness

ACHIEVING CORRECT SYSTEM OUTPUTS, EVEN WITH UNEXPECTED OR ADVERSARIAL INPUTS

Choose metrics

Veracity and robustness

ACHIEVING CORRECT SYSTEM OUTPUTS, EVEN WITH UNEXPECTED OR ADVERSARIAL INPUTS

Privacy and security

APPROPRIATELY OBTAINING, USING, AND PROTECTING DATA AND MODELS

Example: Output should exclude personal identifying information for non-public figures

Choose metrics

Veracity and robustness

ACHIEVING CORRECT SYSTEM OUTPUTS, EVEN WITH UNEXPECTED OR ADVERSARIAL INPUTS

Privacy and security

APPROPRIATELY OBTAINING, USING, AND PROTECTING DATA AND MODELS

Example: Output should exclude personal identifying information for non-public figures

Safety

HARMFUL SYSTEM OUTPUT TO AN INDIVIDUAL OR A GROUP OF INDIVIDUALS

Choose metrics

Safety: Toxicity

HARMFUL SYSTEM OUTPUT 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

Choose metrics

Fairness

CONSIDERING IMPACTS ON DIFFERENT GROUPS OF STAKEHOLDERS

Choose metrics

Fairness

CONSIDERING IMPACTS ON DIFFERENT GROUPS OF STAKEHOLDERS

Choose metrics

Fairness

CONSIDERING IMPACTS ON DIFFERENT GROUPS OF STAKEHOLDERS

Choose metrics

Fairness

CONSIDERING IMPACTS ON DIFFERENT GROUPS OF STAKEHOLDERS

Establishing launch confidence

Responsible evaluation strategy

Define use case | Assess risk | Choose metrics | **Set release criteria**

What are the minimum thresholds of performance that give us confidence in our release?

Set release criteria

What are the minimum thresholds of performance that give us confidence in our release?

Likelihood

Severity	Highly unlikely	Unlikely	Possible	Likely	Almost certain/ Frequent
Extreme	High	High	Critical	Critical	Critical
Major	Medium	Medium	High	High	Critical
Moderate	Very Low	Low	Medium	Medium	High
Low	Very Low	Very Low	Low	Low	Medium
Very low	Very Low	Very Low	Very Low	Very Low	Low

Set release criteria

What are the minimum thresholds of performance that give us confidence in our release?

Likelihood

Severity	Highly unlikely	Unlikely	Possible	Likely	Almost certain/ Frequent
Extreme	High	High	Critical	Critical	Critical
Major	Medium	Medium	High	High	Critical
Moderate	Very Low	Low	Medium	Medium	High
Low	Very Low	Very Low	Low	Low	Medium
Very low	Very Low	Very Low	Very Low	Very Low	Low

Work backwards from severity

Set release criteria

What are the minimum thresholds of performance that give us confidence in our release?

Likelihood

Severity

	Highly unlikely	**Unlikely**	**Possible**	**Likely**	**Almost certain/ Frequent**
Extreme	High	High	Critical	Critical	Critical
Major	Medium	Medium	High	High	Critical
Moderate	Very Low	Low	Medium	Medium	High
Low	Very Low	Very Low	Low	Low	Medium
Very low	Very Low	Very Low	Very Low	Very Low	Low

Higher severity may require lower likelihood for a confident release

Set release criteria

What are the minimum thresholds of performance that give us confidence in our release?

Toxic generation rate

Likelihood can be measured using our chosen metrics

Set release criteria

What are the minimum thresholds of performance that give us confidence in our release?

Toxic generation rate

Likelihood can be measured using our chosen metrics

Set release criteria

What are the minimum thresholds of performance that give us confidence in our release?

Frequently

Toxic generation rate

Likelihood can be measured using our chosen metrics

Set release criteria

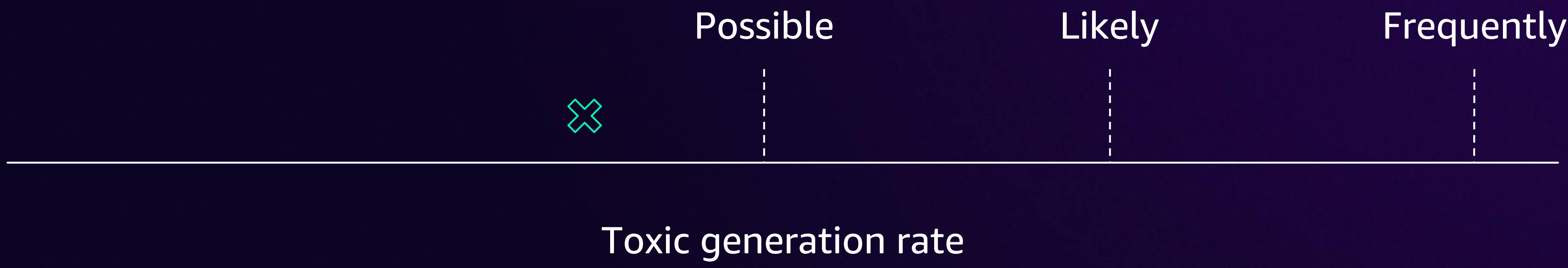
What are the minimum thresholds of performance that give us confidence in our release?

Toxic generation rate

Likelihood can be measured using our chosen metrics

Set release criteria

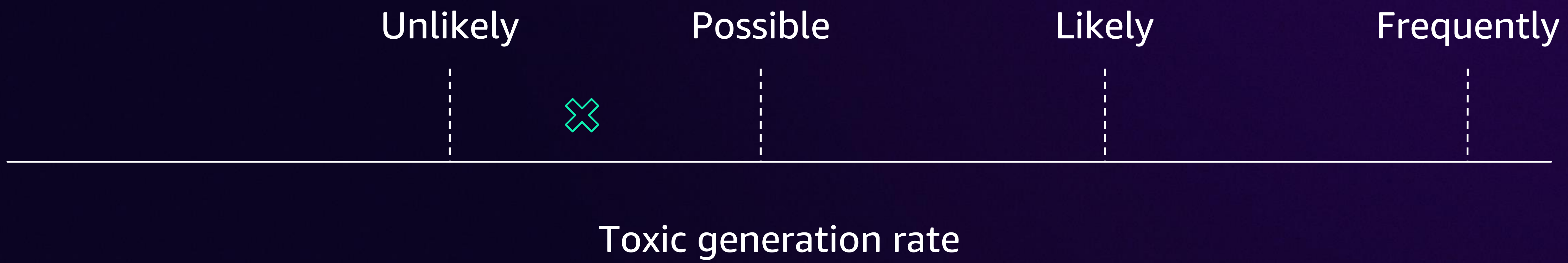
What are the minimum thresholds of performance that give us confidence in our release?



Likelihood can be measured using our chosen metrics

Set release criteria

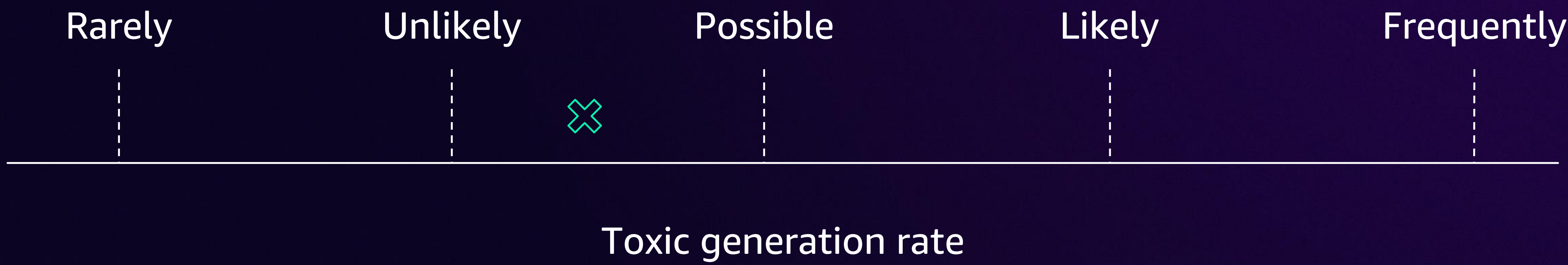
What are the minimum thresholds of performance that give us confidence in our release?



Likelihood can be measured using our chosen metrics

Set release criteria

What are the minimum thresholds of performance that give us confidence in our release?



Likelihood can be measured using our chosen metrics

Establishing launch confidence

Responsible evaluation strategy

Define use case

Assess risk

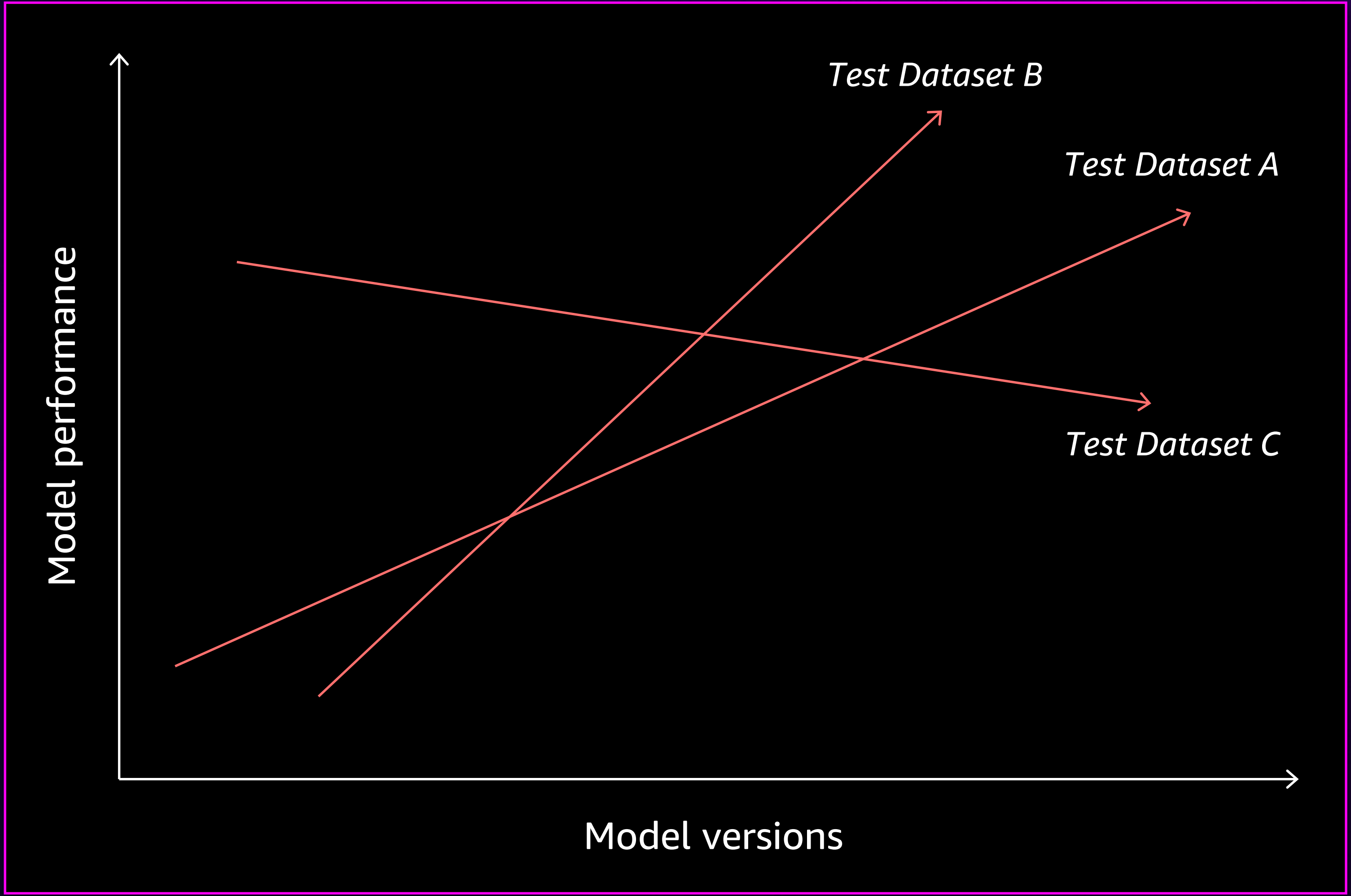
Choose metrics

Set release criteria

Design evaluation dataset

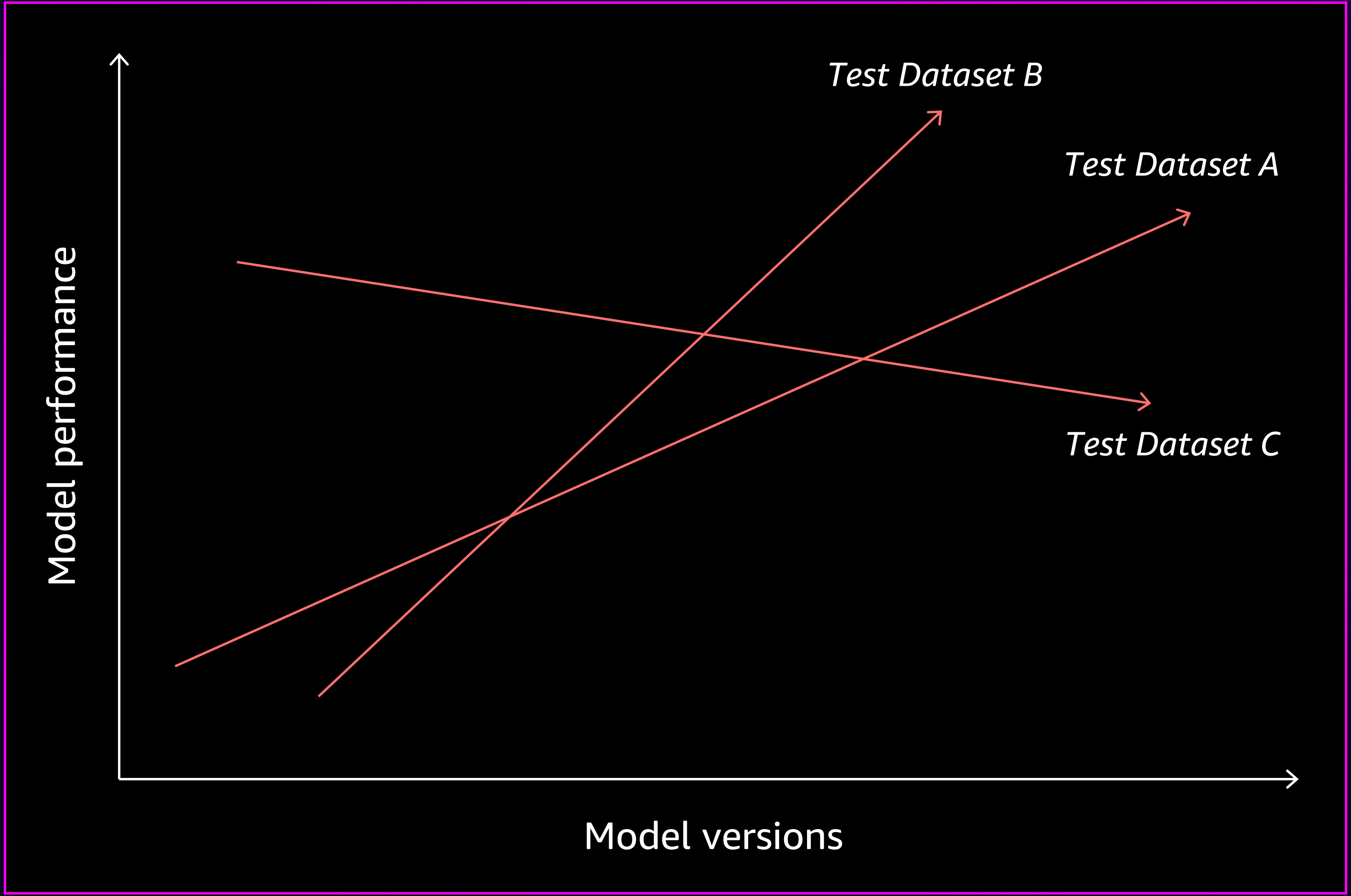
Design evaluation dataset

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



Design evaluation dataset

Performance is a function of an application, a risk dimension, a test dataset, and a metric



Establishing launch confidence

Responsible evaluation strategy

Define use case	Assess risk	Choose metrics	Set release criteria	Design evaluation dataset(s)	**Generate metrics**

Generate metrics

Quantitative ⟷ Qualitative

Accuracy
Precision
Recall
F1
Perplexity

ROUGE
METEOR
BLEU
WER

Relevance
BERTScore
MoverScore

Fluency
Coherence
Semantic similarity

RAG

- Contextual recall
- Contextual relevancy
- Contextual precision
- Answer relevancy
- Faithfulness
- Noise sensitivity

Latency and cost

Time-to-last-token
Time-to-first-token
Output tokens per second

Amazon SageMaker

Foundation model evaluation

Scale human evaluation with work-team management and pre-built labelling portal

Automate evaluation with configurable, fully managed jobs

Select from built-in standard datasets or bring your own

Build fully custom evaluation scripts using open source fmeval library on SageMaker

Powered by AWS open-source fmeval

Amazon Bedrock Model Evaluation

Evaluate, compare, and select the best foundation model for your use case

New:
Public API
Evaluate custom models
Evaluate distilled models
Evaluate imported models
Evaluate prompt routers
Use an LLM-as-a-judge (Preview)

1. Use curated datasets or bring your own for tailored results
2. Use automatic (algos or LLMs) or human evaluation methods
3. Leverage your in-house team or AWS managed reviewers
4. Predefined and custom metrics
5. Get results in just a few clicks

Establishing launch confidence

Responsible evaluation strategy

Define use case	Assess risk	Choose metrics	Set release criteria	Design evaluation dataset(s)	Generate metrics	**Interpret results**

Do my evaluation results establish confidence in my release?

Interpret results

Do my evaluation results establish confidence in my release?

Interpret results

Do my evaluation results establish confidence in my release?

Interpret results

Do my evaluation results establish confidence in my release?

Interpret results

Do my evaluation results establish confidence in my release?

Interpret results

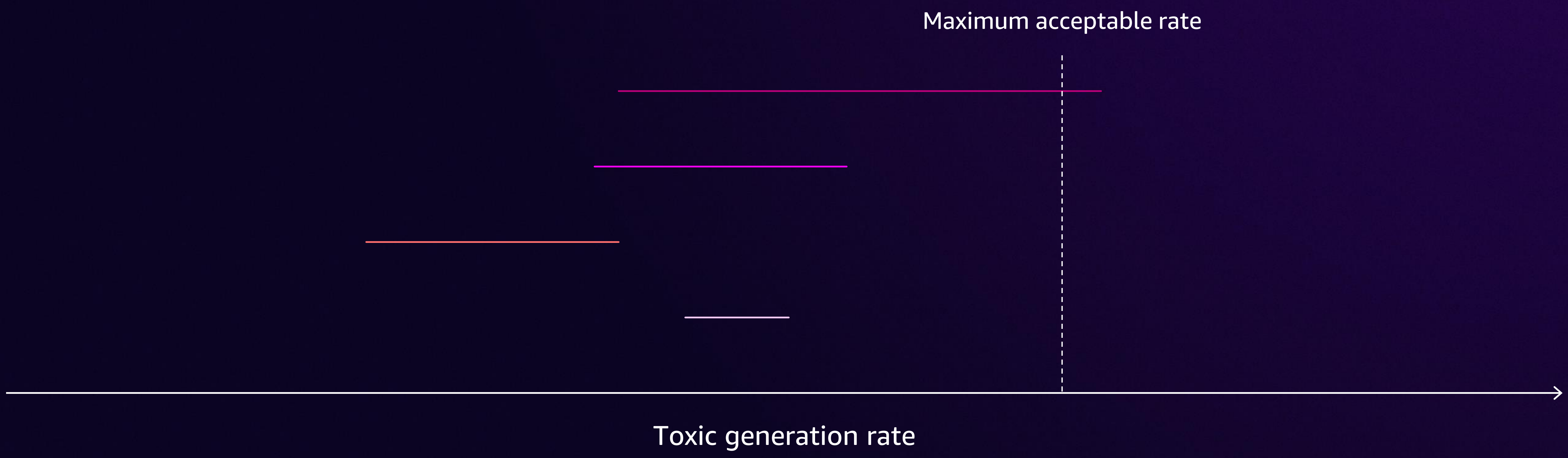
Do my evaluation results establish confidence in my release?

Confidence intervals capture uncertainty in our performance estimation

Interpret results

Do my evaluation results establish confidence in my release?

Confidence intervals capture uncertainty in our performance estimation



Interpret results

Do my evaluation results establish confidence in my release?

One-sided intervals capture uncertainty in meeting our threshold

Interpret results

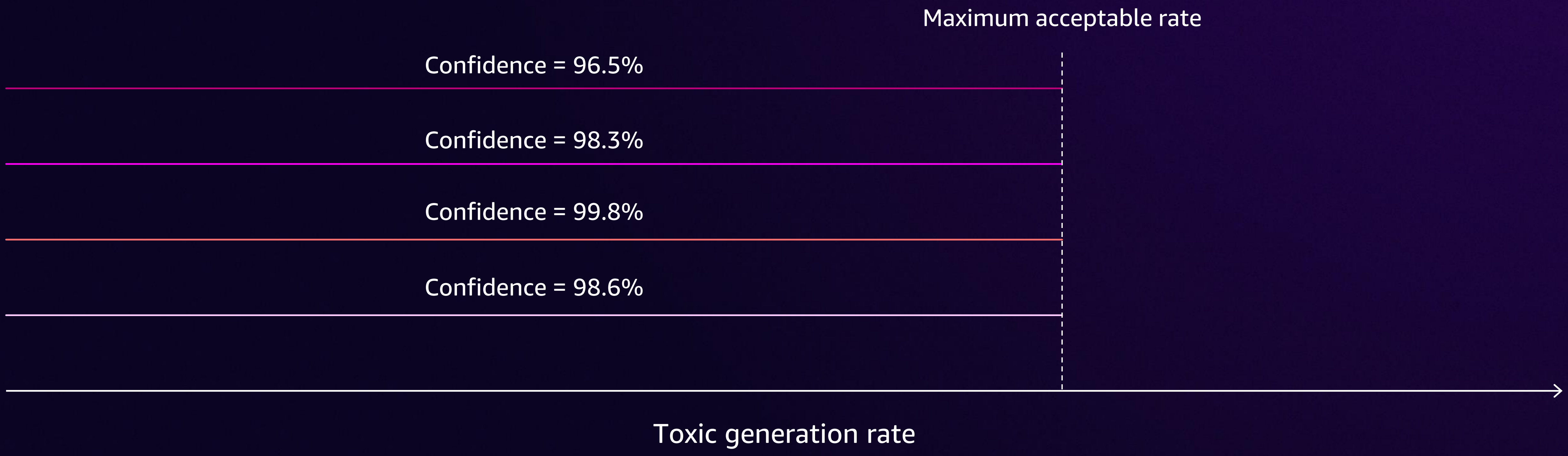
Do my evaluation results establish confidence in my release?

Hypothesis testing (one-tailed)

Interpret results

Do my evaluation results establish confidence in my release?

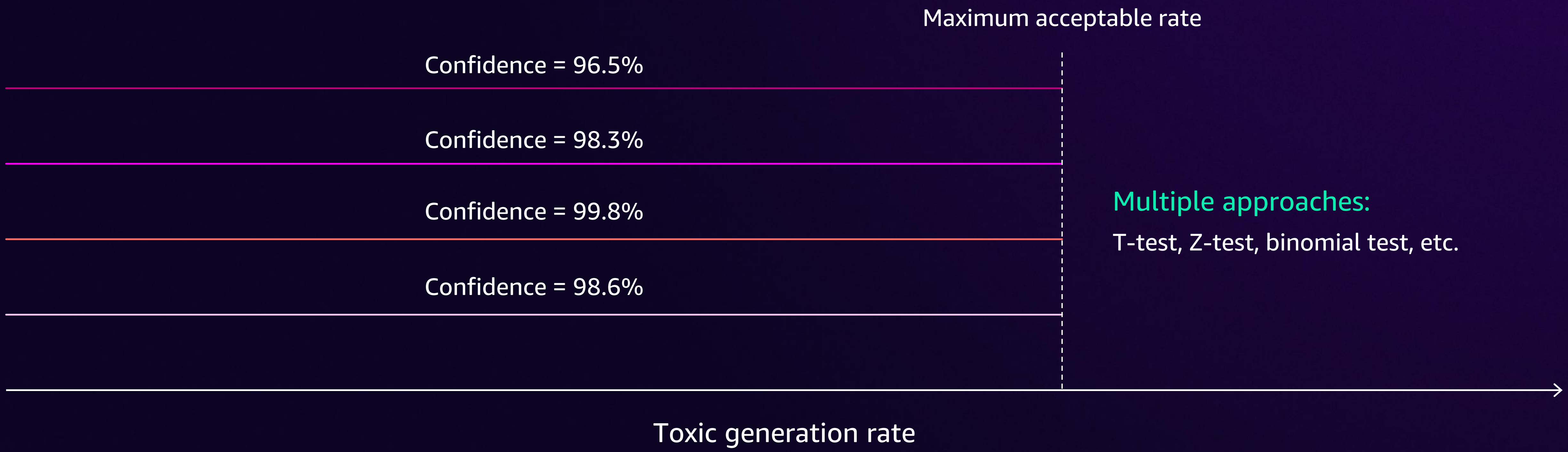
Hypothesis testing (one-tailed)



Interpret results

Do my evaluation results establish confidence in my release?

Hypothesis testing (one-tailed)



Multiple approaches:

T-test, Z-test, binomial test, etc.

Interpret results

Do my evaluation results establish confidence in my release?

Confidence intervals capture uncertainty in our performance estimation

Interpret results

Do my evaluation results establish confidence in my release?

Confidence intervals capture uncertainty in our performance estimation

Interpret results

Do my evaluation results establish confidence in my release?

Confidence intervals capture uncertainty in our performance estimation

Interpret results

Do my evaluation results establish confidence in my release?

Hypothesis testing (disparity)

Testing for significant disparities across 2 or more groups:

Analysis of variance (ANOVA): Does a disparity exist?

Interpret results

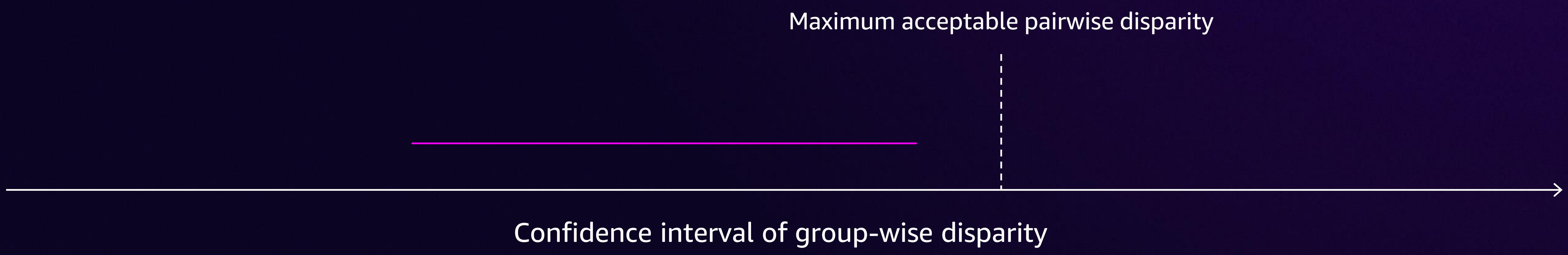
Do my evaluation results establish confidence in my release?

Hypothesis testing (disparity)

Testing for significant disparities across 2 or more groups:

Analysis of variance (ANOVA): Does a disparity exist?

Tukey Honest Significant Differences (Tukey's HSD): What is the disparity?



Interpret results

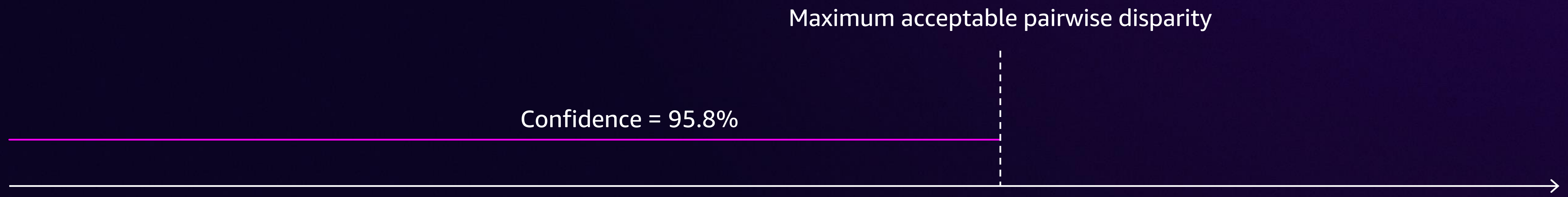
Do my evaluation results establish confidence in my release?

Hypothesis testing (disparity)

Testing for significant disparities across 2 or more groups:

Analysis of variance (ANOVA): Does a disparity exist?

Tukey Honest Significant Differences (Tukey's HSD): What is the disparity?



One-sided confidence interval of group-wise disparity

Interpret results

Do my evaluation results establish confidence in my release?

Use all available evidence

Documenting results

Transparency reports

Documents the full evaluation process

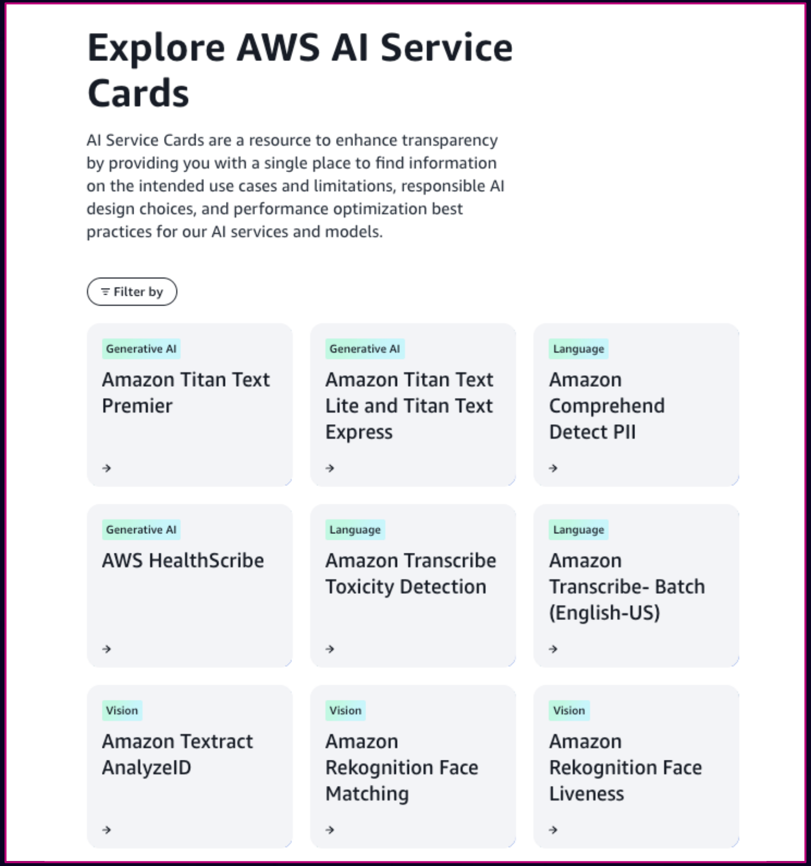
Documenting results

Transparency reports

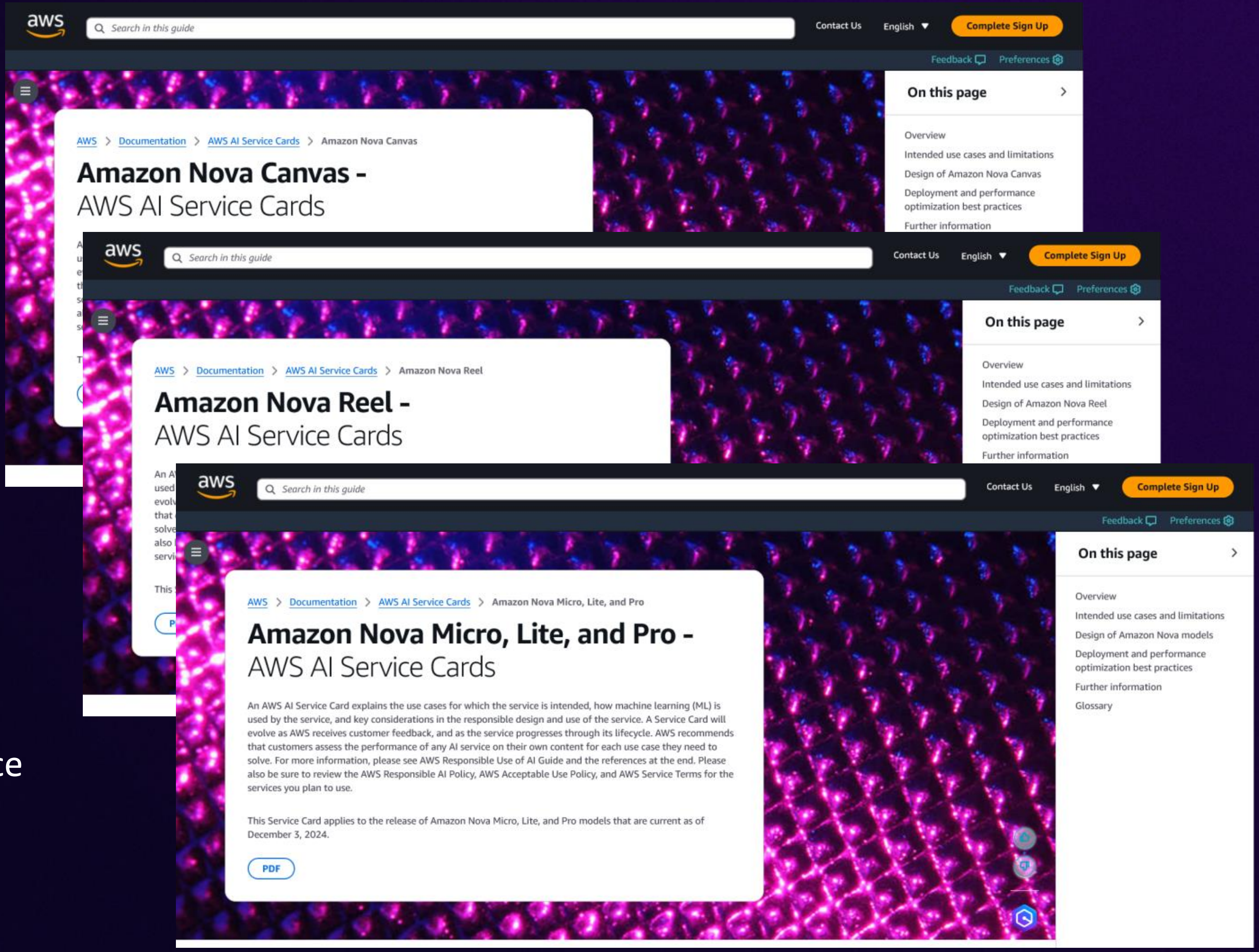
Documents the full evaluation process

Builds customer trust

Documenting results



AWS AI Service Cards



https://aws.amazon.com/ai/responsible-ai/resources/#service

Mitigation strategy

What happens if we are not confident in our release?

Mitigation strategy

What happens if we are not confident in our release?

Identify sources of low confidence

Mitigation strategy

What happens if we are not confident in our release?

Identify sources of low confidence

Implement methods to improve our system

Mitigation strategy

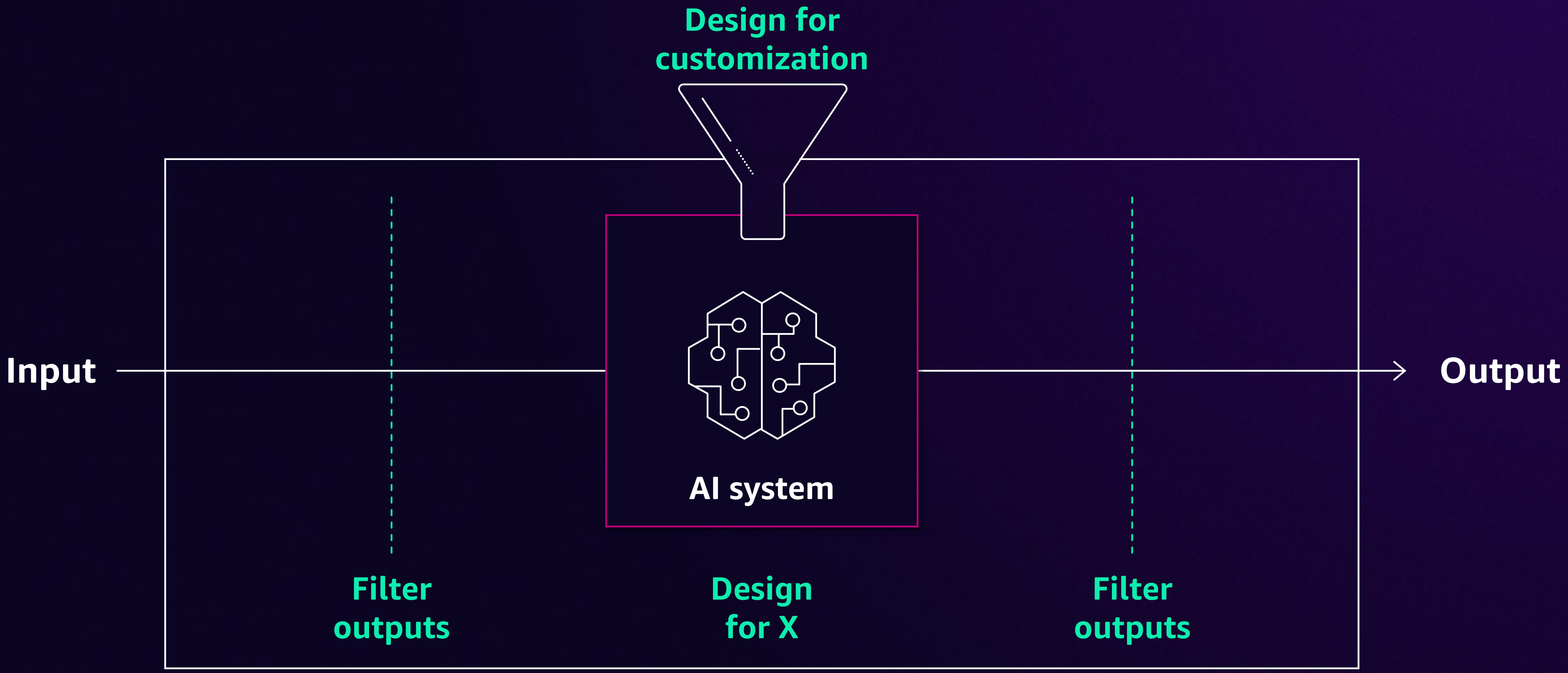
What happens if we are not confident in our release?

Identify sources of low confidence

Implement methods to improve our system

Define use case	Assess risk	Choose metrics	Set release criteria	Design evaluation dataset(s)	Generate metrics	Interpret results	**Mitigation**

Mitigation strategy



Amazon Bedrock Guardrails

Evaluate prompts and model responses for agents, knowledge bases, FMs in Amazon Bedrock, and self-managed or third-party FMs

- Configure thresholds to filter undesirable and potentially harmful text and image content, jailbreaks, and prompt attacks NEW!
- Identify, correct, and explain factual claims in responses using Automated Reasoning NEW!
- Define and disallow denied topics with short natural language descriptions
- Remove personally identifiable information (PII) and sensitive information in generative AI applications
- Define a set of words to detect and block in user inputs and model responses
- Filter hallucinations by detecting groundedness and relevance of model responses based on context

Mitigation strategy

Are we confident in our release?

Mitigation strategy

Are we confident in our release?

Define use case | Assess risk | Choose metrics | Set release criteria | Design evaluation dataset(s) | **Generate metrics** | Interpret results | Mitigation

Mitigation strategy

QUALITY

Perming as or better than expected

LATENCY

Fast enough for its purpose

COST

$

CONFIDENCE

Risks are acceptable

Review

FUNDAMENTALS OF GENERATIVE AI EVALUATION

01 What it means to evaluate LLMs

02 Tools

03 Assessing risk

04 Establishing release confidence

Review

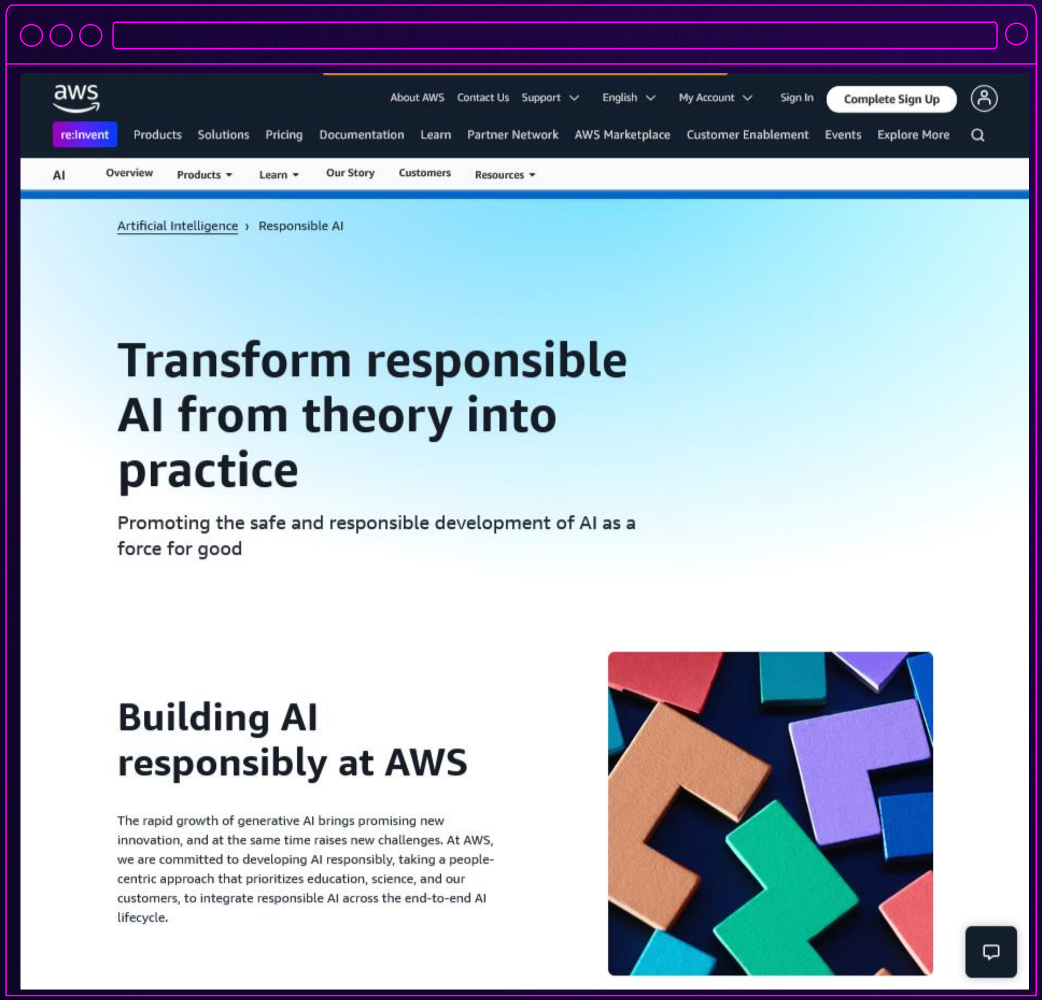
FUNDAMENTALS OF GENERATIVE AI EVALUATION

01 What it means to evaluate LLMs

02 Tools

03 Assessing risk

04 Establishing release confidence



https://aws.amazon.com/ai/responsible-ai/

Thank you!

Please complete the session survey in the mobile app

Alessandro Cerè, PhD
(he/him)
Principle Solutions Architect,
Model Eval
Amazon Web Services

Mathew Monfort, PhD
(he/him)
Senior Applied Scientist,
Responsible AI
Amazon Web Services