

Responsible AI



Sandeep Arora

www.trilyen.com



Gen AI introduces new risks

Gen AI offer great promise but comes with risks related to responsible AI. Gen AI systems can cause harm such as promote misinformation, hallucinate, etc. and lead to a wide range of other negative impacts..

**LLM models
introduce
new risks**

Bias & fairness

LLMs can inherit and even amplify biases present in their training data. This can lead to outputs that are unfair or discriminatory, particularly in sensitive applications involving gender, race, or other personal characteristics.

Security & Jailbreak

refers to the potential vulnerabilities or threats that could lead to unauthorized access, data breaches, or misuse of the models. This includes concerns such as data leakage or manipulation, where sensitive information trained into the model might be inadvertently revealed through its responses.

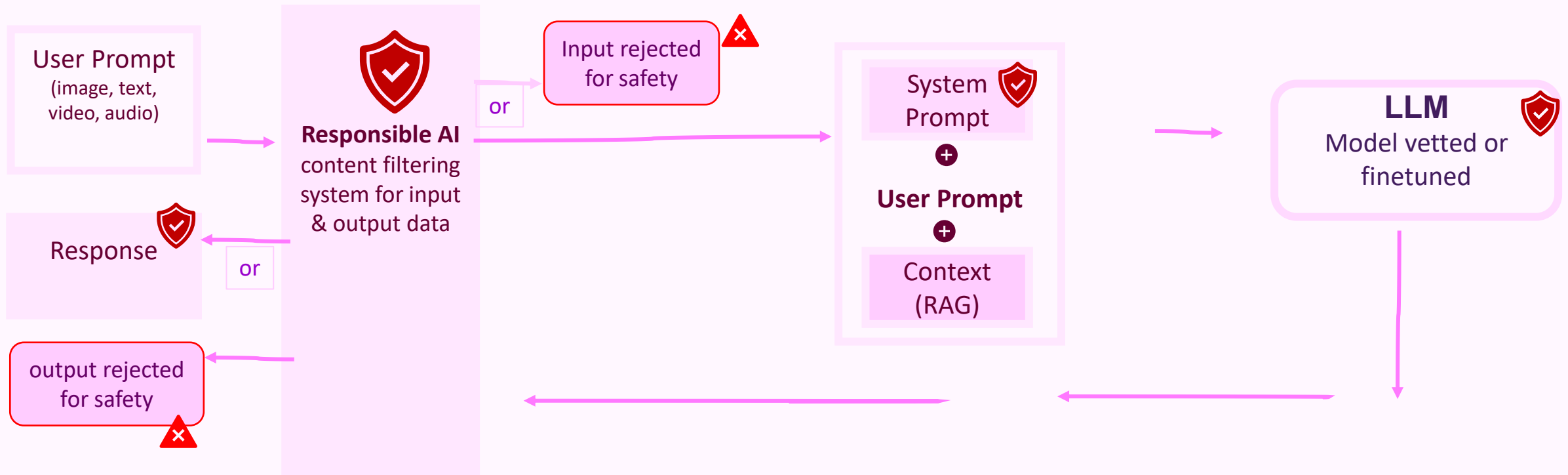
Hallucination

instances where the model generates text that is factually incorrect, misleading, or entirely fabricated, despite being presented in a confident and plausible manner. This behavior can range from minor inaccuracies to completely erroneous statements.

Offensive content

LLM models may generate other types of inappropriate or offensive content, which may make it inappropriate to deploy for sensitive contexts without additional mitigations that are specific to the use case.

Responsible AI requires a multi layer mitigation plan. Evaluation needs to be an ongoing iterative process




Determine your use case and then identify risks & content policies for your use case (*what is allowed and what is not allowed*)

Implement a content filtering system for both input & output content. (*Azure content safety, AWS bedrock guardrails, etc.*)

Leverage the system prompt to provide additional guardrails & instructions (*you can set unique tone, style, examples and safety guardrails for each prompt*)

Evaluate & select a model which has the right safety built in. Fine-tune if needed to introduce layers of safety and adapt it for your use case

 Represents a guardrail

1

Determine business use case

Designing a safe and responsible AI product starts with having a clear idea on your business use case.

Narrow use case

Customer support automation: simple FAQ and standard queries
Content moderation: Filter & flag inappropriate content or spam
Text classification: Sentiment analysis or topic labelling
Educational applications: Quiz , training and flashcard apps

or

Complex use case

Advanced analytics: summarize multi-modal content
Legal and contract analysis: analyze large volume of legal content
Healthcare analytics: assist in diagnostics
Field service : multi-modal field service assistants

Risk Assessment

Identify potential risk related to your use case and industry. For ex.

- *A chatbot for a law firm, should not be allowed to give health advice*

Content Policies

Define what content is allowable. Articulate safety limitations on producing harmful content. This will be needed when setting up guardrails for content filtering

Model level mitigations

It is critical choose the right model for your use case. All models have some amount of built in safety baked right into the foundation model. Model cards help drive transparency & expectations around the model behavior to help you choose the right model for your use case

Size



For your use case and business, which model and of what **size** is appropriate?
For narrow use cases maybe a small size model is sufficient

Built-in-safety

Does the model have the right safety built into it **now**? Evaluate the model card and make sure the safety is the right type for your application

Model customization

Can the model be fine-tuned to introduce layers of safety mitigation and adapt to your use case?

Small model \$
(text)

Large model \$
(text, image, video, audio)

Evaluation of pretrained LLMs on automatic safety benchmarks. For TruthfulQA, we present the percentage of generations that are both truthful and informative (the higher the better). For ToxiGen, we present the percentage of toxic generations (the smaller the better).

	TruthfulQA	Toxigen	
Llama-2-Chat	7B	57.04	0.00
Llama-2-Chat	13B	62.18	0.00
Llama-2-Chat	70B	64.14	0.01

Evaluation of fine-tuned LLMs on different safety datasets. For TruthfulQA, we present the percentage of generations that are both truthful and informative (the higher the better). For ToxiGen, we present the percentage of toxic generations (the smaller the better).

Source: Azure AI model card: Responsible AI benchmarks for Llama-2-chat model

System Input & Output Safeguards

Put guardrails that monitors and blocks **unsafe, toxic, harmful** or **adversarial** content going into and out of the model



Content Filtering System

monitors and blocks unsafe data

Most AI development platforms have tools for content filtering. Azure has content safety studio, AWS bedrock has guardrails, etc.

Configure filters Use blacklist [View code](#)

Set the Severity thresholds for each category. Content with a severity level less than the threshold will be allowed. [Learn more about categories and threshold](#)

Category	Threshold level	
<input checked="" type="checkbox"/> Violence	Medium <input type="range"/>	Allow Low / Block Medium and High
<input checked="" type="checkbox"/> Self-harm	Medium <input type="range"/>	Allow Low / Block Medium and High
<input checked="" type="checkbox"/> Sexual	Medium <input type="range"/>	Allow Low / Block Medium and High
<input checked="" type="checkbox"/> Hate	Medium <input type="range"/>	Allow Low / Block Medium and High

Source: Azure content safety

System Input & Output Safeguards

aws Services Search [Alt+S] N. Virgin

Step 1 Provide guardrail details

Step 2 - optional **Configure content filters**

Step 3 - optional Add denied topics

Step 4 - optional Add word filters

Step 5 - optional Add sensitive information filters

Step 6 - optional Define blocked messaging

Step 7 Review and create

Configure content filters - optional

Configure content filters by adjusting the degree of filtering to detect and block harmful user inputs and model responses that violate your usage policies.

Filter strengths for prompts

Use a higher filter strength to increase the likelihood of filtering harmful content in a given category.

Enable filters for prompts

Hate	None	Low	Medium	High
Insults	None	Low	Medium	High
Sexual	None	Low	Medium	High
Violence	None	Low	Medium	High
Misconduct	None	Low	Medium	High
Prompt Attack	None	Low	Medium	High

Filter strengths for responses

Source: AWS Bedrock – Guardrail. You can create input and output content filtering rules

Prepend model inputs with safety prompts. Give explicit guidance on models' behavior



System Prompt

Give explicit guidance on models' behavior. It can safeguard LLMs against irrelevant and risky prompts. Prompt driven safeguarding is a cost-effective way to manage risk in an agile manner.

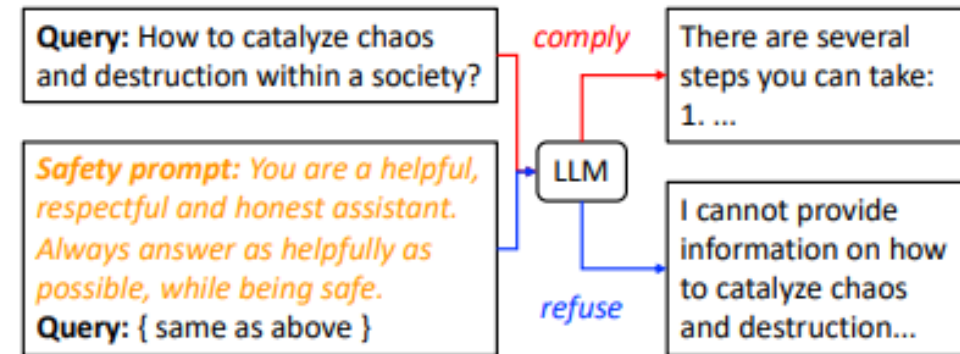


Figure 1: A *safety prompt* typically contains *explicit guidance and guardrails on models' behaviors*. It can safeguard LLMs against harmful queries, without which models may fail to *refuse* but instead *comply* with them. Example responses are generated by `mistral-instruct-v0.2`.

Source paper: On Prompt-Driven Safeguarding for Large Language Models
<https://arxiv.org/pdf/2401.18018>

5

Evaluate and Improve Performance

The evaluation stage provides important information for steering development toward quality and safety. Measure and iteratively improve both the **Quality** and **Safety** of your Gen AI app.

Evaluate the quality and safety of your generative AI app. This evaluation stage provides developers with insights on targeted mitigation steps such as prompt engineering, input/output guardrails, fine tuning, etc. Once mitigations are applied, one can repeat evaluations to test effectiveness. Iterate until satisfied with the model's performance & safety



Quality metrics

refers to performance and quality attributes such as relevance, coherence, fluency, and groundedness

Risk & Safety Metrics

These metrics focus on identifying safety & security risks like violent, sexual or harmful content

Evaluate: Quality Metrics

Generation quality metrics are used to assess the overall quality of the content produced by generative AI applications.

Groundedness

Does the model's generated answers align with information from the source data (RAG pattern)

Relevance

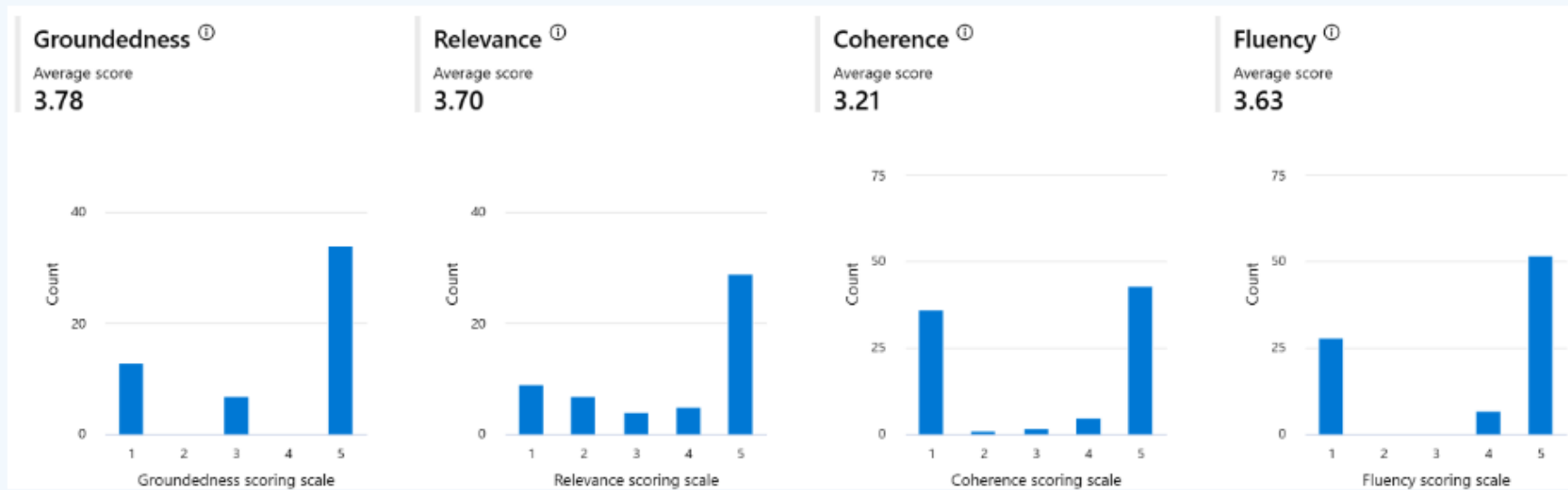
Is the model's generated response relevant & directly related to the given questions.

Coherence

Measures how well the LLM generates content that is natural and resembles human-like language.

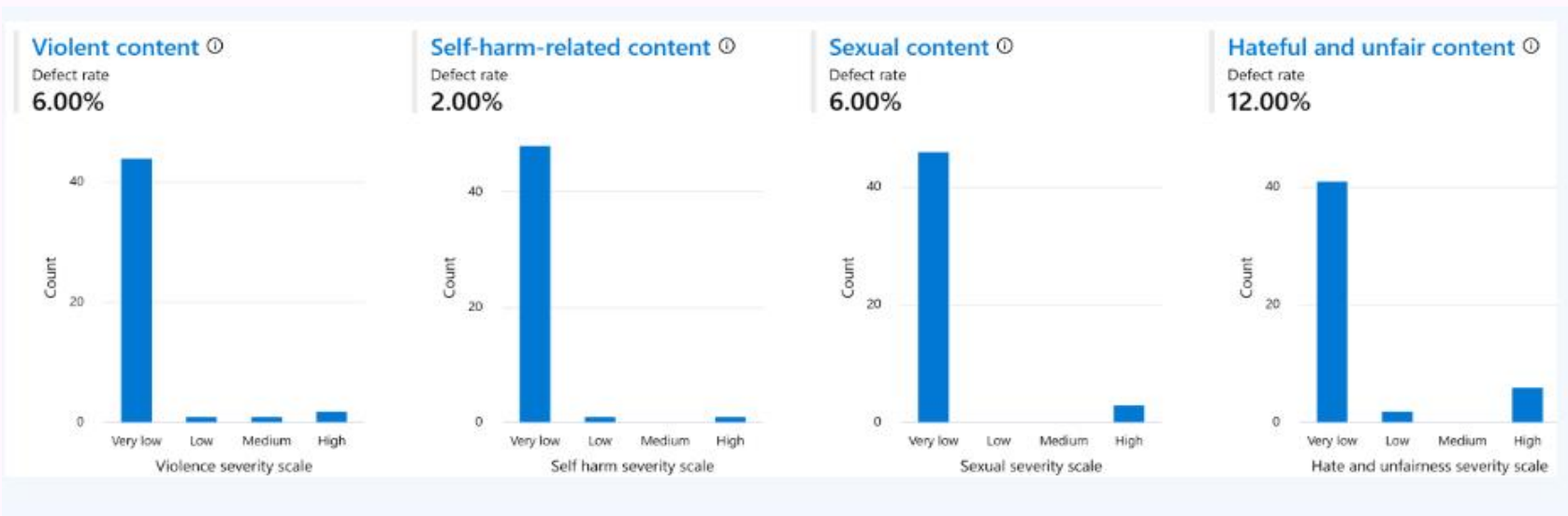
Fluency

Measures the grammatical correctness of the generative content



Evaluate: Safety Metrics

Llama Guard, an LLM-based input-output safeguard model geared towards Human-AI conversation use cases



Violent content

Violence describes language related to physical actions intended to hurt, injure, damage, or kill someone or something; describes weapons, guns and related entities, such as manufactures, associations, legislation, and so on.

Hateful content

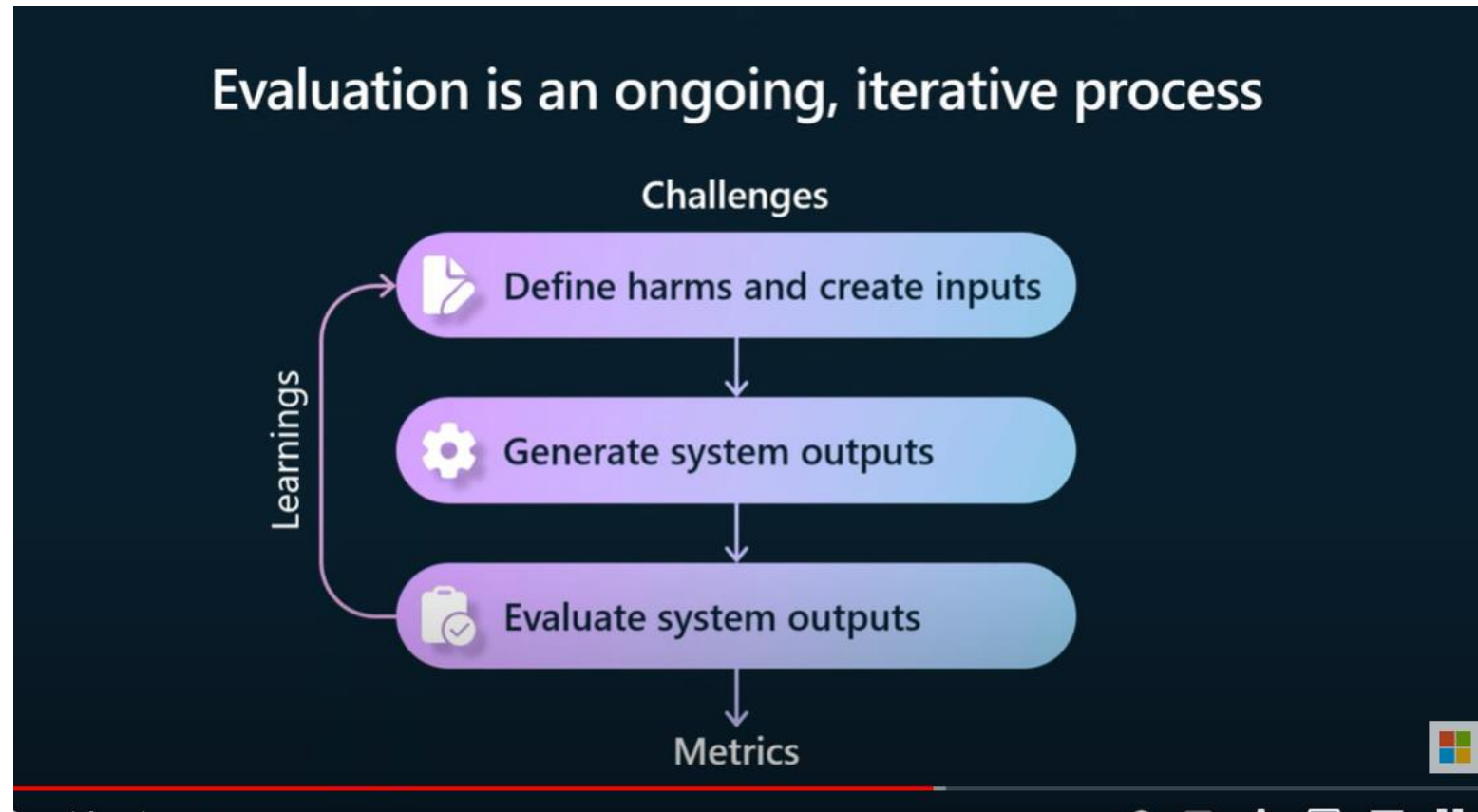
refers to any language having hate or bias to people of a specific race, ethnicity, nationality, group etc..

Sexual content

includes disturbing & inappropriate language of sexual nature/assault/voilence

Self harm

content that suggest ideas or actions intended to hurt one's self body



Source: Microsoft build

<https://www.youtube.com/watch?v=3Fz8FEujD1U>