DATA & AI

GEN AI RISK MANAGEMENT

Developing Safe and Trustworthy Applications

REPLY
CLUSTER

GENERATIVE AI RISKS

The rapid progress of Generative AI models like DALL-E and GPT-4 presents exciting creative opportunities, yet it also brings to light significant safety and security concerns. As these models become increasingly prevalent, it's crucial to develop rigorous methods for assessing potential risks before deploying them in real-world scenarios.

Generative AI exposes organizations to new **Operational**, **Ethical, Security & Safety** risks.

OPERATIONAL

- Inaccurate Output
- Inconsistency
- Protected Material Usage

ETHICAL

- Biased Results
- Toxic Content
- Dangerous Recommendations

SECURITY & SAFETY

- Sensitive Information Exposure
- Prompt Injection

ETHICAL RISKS

Ethical Risks arise when models generate output, violating norms, laws, regulations, or other governance standards.

BIASED RESULTS

Generative AI models trained on biased datasets can perpetuate **unfair outcomes**, reinforcing societal inequalities and potentially generating **discriminatory** or misleading **content**.

TOXIC CONTENT

Generative AI models might **create output** containing **hateful**, **abusive** and **profane** content that can **harm** people interacting with the model leading to reputational harms for the company providing the AI solution.

DANGEROUS RECOMMENDATIONS

Generative AI models might generate **provocative**, radicalizing, and **violent content**, including **instructions** for **dangerous** or **unethical behavior**.

REPLY
CLUSTER

OPERATIONAL RISKS

Ethical Risks arise when models generate output, violating norms, laws, regulations, or other governance standards.

INACCURATE OUTPUT

Inaccurate output (**Hallucinations or Confabulation**) occurs when a Generative AI models perceives nonexistent patterns, **generating false data** and confidently **relying** on it, posing challenges in reliability.

INCONSISTENCY

When provided with the same input, Generative AI models' **outputs** may **vary** quite **significantly** due to the statistical nature of the models, posing challenges in ensuring **consistent** and **predictable** behavior when needed.

PROTECTED MATERIAL USAGE

AI models may **generate existing content** such as copyrighted text or even code without proper authorization, potentially infringing upon copyright laws and leading to legal consequences.

REPLY
CLUSTER

SECURITY & SAFETY RISKS

Security risk refers to Generative AI vulnerabilities that may be exploited maliciously or inadvertently by users to generate an unintended output or force the model to behave incorrectly.

SENSITIVE INFORMATION EXPOSURE

AI models might inadvertently **disclose confidential data** in their outputs, potentially resulting in unauthorized data access, privacy infringements, and security breaches.

PROMPT INJECTION

Attackers may use **Direct Prompt Attack** (**Jailbreak**) to manipulate the system via the prompt, bypassing safeguards and altering its behavior, or employ **Indirect Prompt Attack**, inserting malicious content to deceive the system for unauthorized access or control.

REPLY CLUSTER

HOW TO SOLVE?

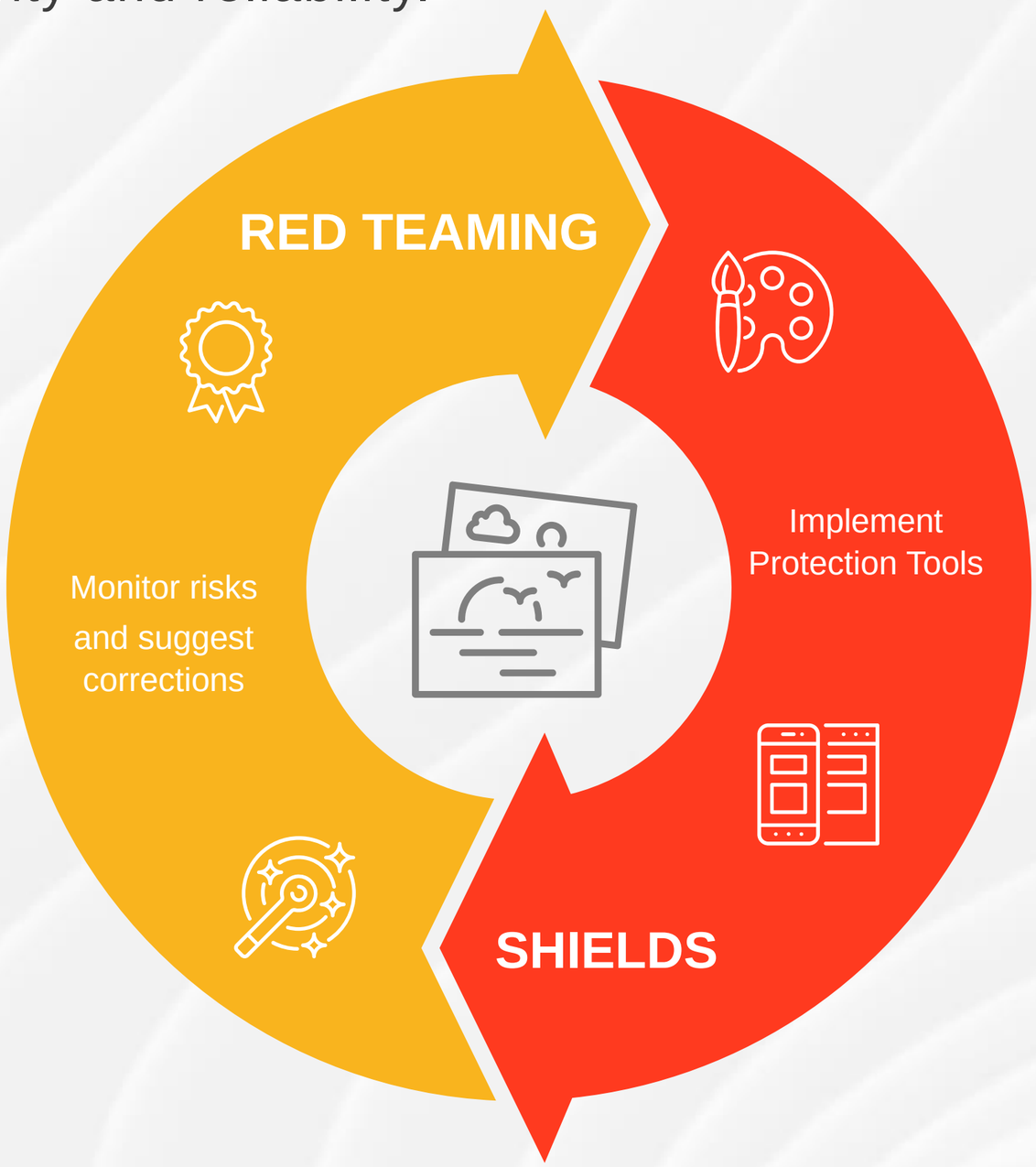
To mitigate the aforementioned risks associated with Generative AI, two essential approaches should be implemented:

AI Validation (**Red Teaming)**: Employing rigorous **testing methods** to uncover potential flaws and vulnerabilities within AI systems.

AI Protection: Implementing a range of **security measures** to protect AI systems to ensure their integrity and reliability.

AI VALIDATION

Apply behavioral tests to assess how the model responds to various risks.



AI PROTECTION

Apply Real Time Protection using tools as **Azure AI Content Safety**

OUR SOLUTION

To make sure your application is safe and secure, we provide support in **monitoring** and **validating** your application before publishing it.
Once running, we **protect** it through services such as Microsoft **Azure AI Content Safety**.

AI VALIDATION

- Apply **Red Teaming** through specific test before go-live
- **Identify** critical security, ethical, operational **risks**
- Develop customized **dashboard** for risk monitoring
- Periodically **review** and **test** model in production
- Suggest **improvements** for AI solutions

AI PROTECTION

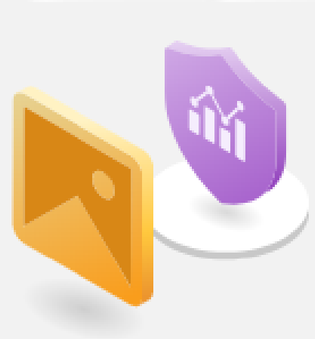
- Apply **Real-Time protection** with services as **Azure AI Content Safety**
- **Filter** input presenting **jailbreak** risks
- **Check** models' output and users' input to filter **toxic content**
- **Identify** and **protect** intellectual property filtering copyrighted material
- **Detect ungroundedness** in models' output

REPLY CLUSTER

AZURE AI CONTENT SAFETY - FEATURES

Azure AI Content Safety helps you **detect** and **filter** harmful user-generated and AI-generated **content** in your applications and services.

Moderate Text Content
Run moderation tests on text samples to detect unwanted content.



Moderate Image Content
Run moderation tests on images samples to detect unwanted content.

Moderate Multimodal Content
Run moderation tests on images and text together to detect unwanted content.

Groundedness Detection
Detect ungrounded content generated by LLMs

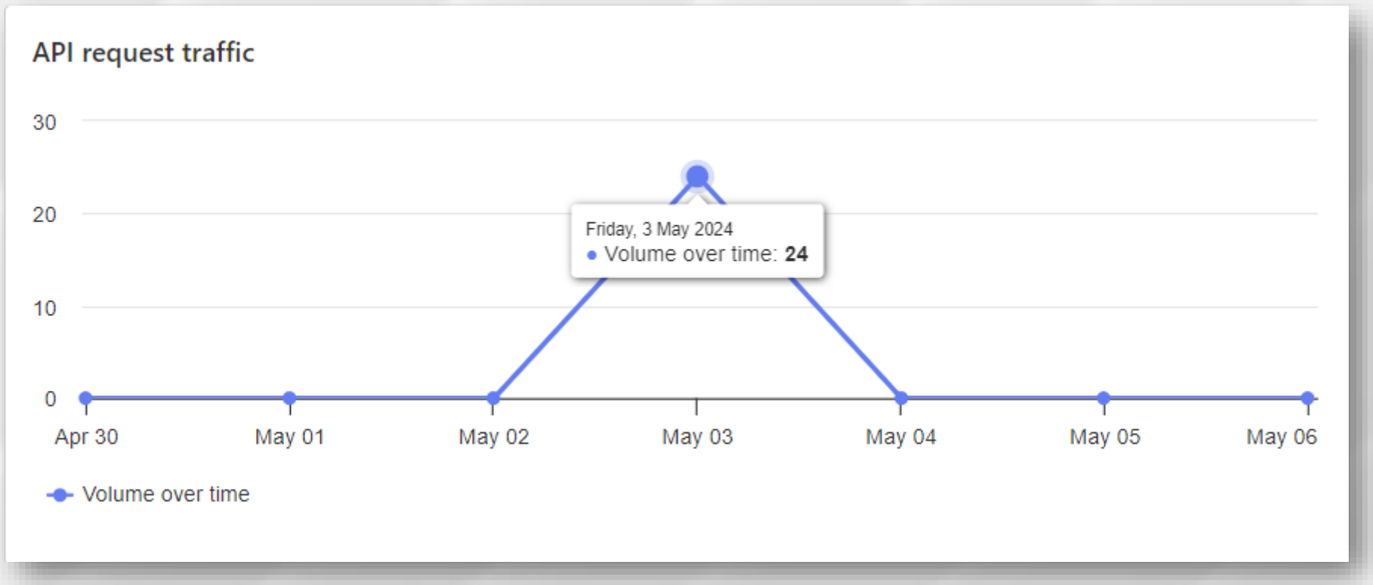
Prompt Shields
Provide APIs to detect and filter jailbreak attempts

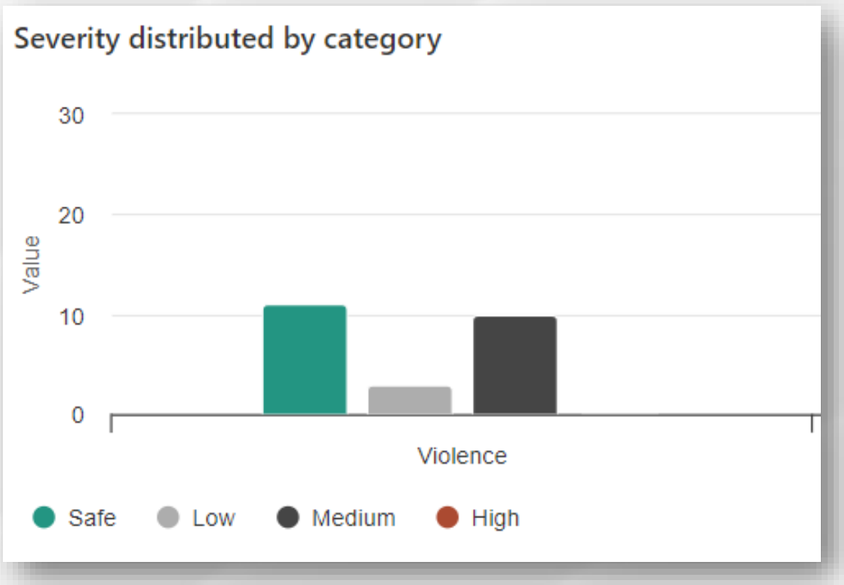
Protected Material Detection
Detect and protect third-party text material

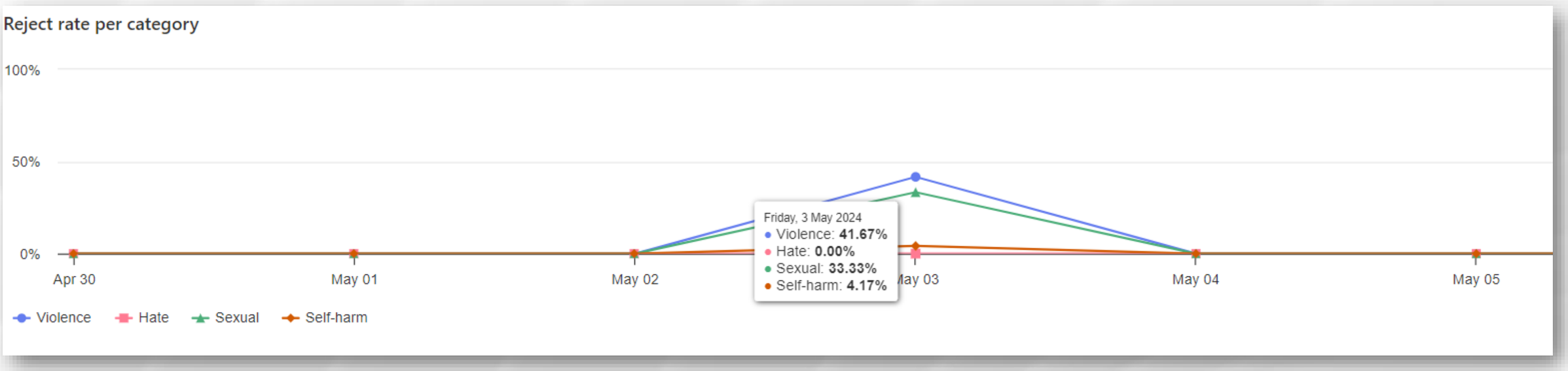
REPLY CLUSTER

AZURE AI CONTENT SAFETY - MONITORING

For optimal security assurance, Azure AI Safety presents comprehensive monitoring dashboards. Designed for clarity, these dashboards offer intuitive visuals, empowering you with deep insights into the security status of your applications







REPLY CLUSTER

BUSINESS GOAL

OUTPUT RELIABILITY

Ensure that generative models produce **accurate** and **reliable results**, minimizing occurrences of inaccurate or misleading output that could compromise system **reliability**.

DATA PROTECTION

Ensure the security of AI implementing measures to prevent the exposure of **sensitive information**, the **injection** of harmful prompts, and the unauthorized use of **protected materials**.

ETHICS AND RESPONSIBILITY

Implement strategies to mitigate ethical risks such as **reducing bias** and preventing **toxic** or **discriminatory content**, that might **damage** company's **reputation**

COMPLIANCE & REGULATIONS

Ensure compliance with regulations regarding AI ethics and safety, minimizing the risk of **regulatory violations.**

DATA & AI

REPLY
CLUSTER

Contact us!

Claudio Motisi
c.motisi@reply.it

Giulio Cammarata
g.cammarata@reply.it

WWW.CLUSTERREPLY.IT