INSIDE GENAI

n° 03

The key security and safety challenges in adopting GenAI

CRIF
Together to the next level

Table of Contents

1 Introduction

Analyzing the reliability of generative artificial intelligence (GenAI) systems is crucial in today's technological landscape. The rapid proliferation of GenAI, known for its extraordinary ability to generate complex and realistic content, combined with its "black box" nature, presents unprecedented challenges.

The number of providers is increasing, while the gap between open source and proprietary models is decreasing, as are the inference costs. In the early stages of GenAI's evolution, **reasoning capability** was the most important criterion for selecting one model over another; now, additional factors such as **price**, **speed**, and **security** play a crucial role.

In this context and considering that it is expected that the number of GenAI-based applications in production systems will increase in 2025[1], the creation of a **robust evaluation framework** is fundamental to correctly guide architectural choices.

Addressing these challenges requires a concerted effort across research, regulation, and technological innovation to ensure that the benefits of GenAI can be fully realized without compromising the **security** and **integrity** of systems.

Unfortunately, for the AI architectures that power most of the GenAI-based applications, it is impossible to prevent all attacks.

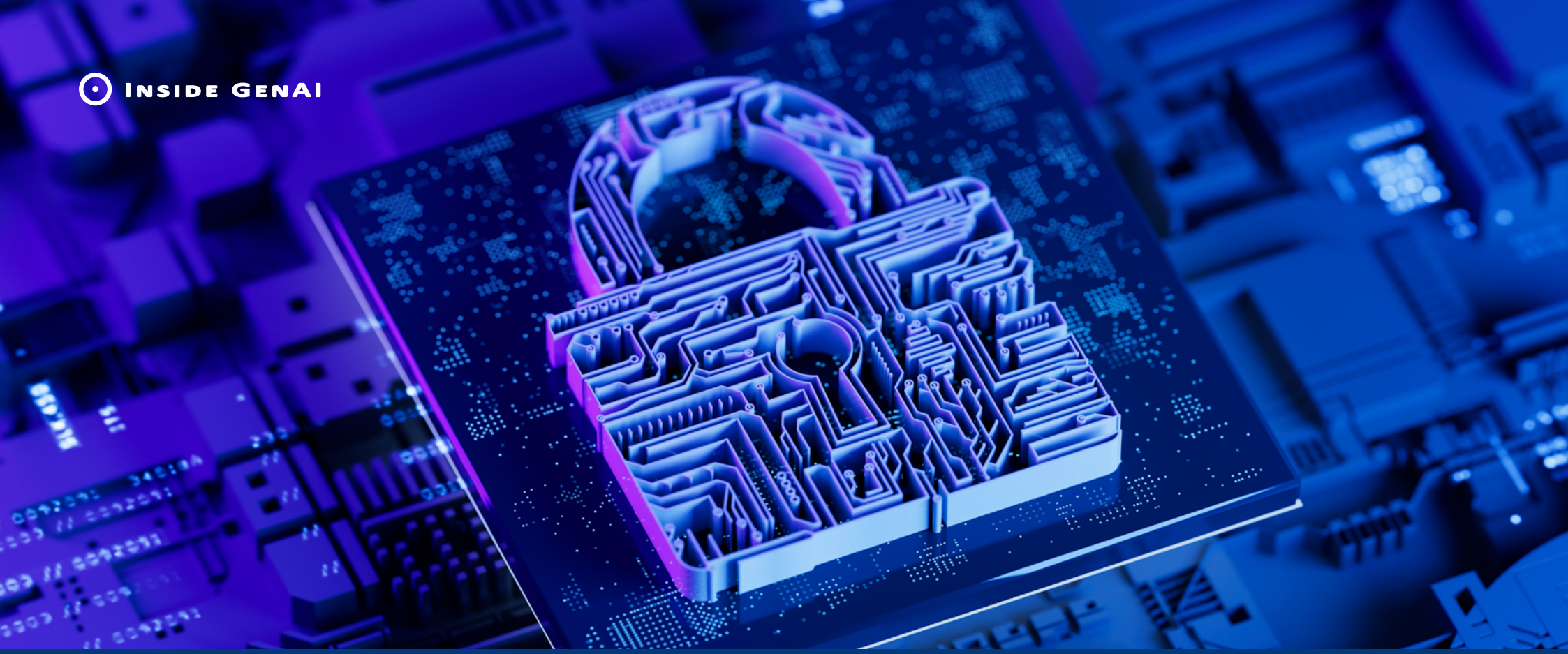
For example, crafted inputs can manipulate the model into producing undesired or harmful outputs, such as unsafe content, while mitigation strategies such as input sanitization, adversarial tuning, and moderation models can strongly reduce these risks[2], but do not eliminate them.

The environment is further complicated by the **lack of clear standards** and the prevalence of **trade secrets**, making independent and transparent evaluations difficult. Additionally, the opacity of both AI algorithms and the organizations developing them makes it difficult to assign legal responsibility, further obstructing governance and regulatory enforcement.

For these reasons, CRIF has always been committed to providing secure and r obust products to its clients. Specifically, the CRIF Engineering and Data Science teams focus on three key development principles: Security, Safety, and Accuracy. CRIF's aim is to deliver precise applications while mitigating operational risks.

[1] **IBM**. 5 Trends for 2025. IBM. [Online] 2025. [Cited: 02 12, 2025.]

[2] Constitutional Classifiers: Defending against Universal Jailbreaks across Thousands of Hours of Red Teaming. Sharma, Mrinank, Meg Tong, Jesse Mu, Jerry Wei, Jorrit Kruthoff, Scott Goodfriend, Euan Ong, Alwin Peng, Raj Agarwal, Cem Anil, Amanda Askell, Nathan Bailey, Joe Benton, Emma Bluemke, Samuel R. Bowman, Eric Christiansen, Hoagy Cunningham, Andy Dau et al. s.l. : arXiv preprint arXiv:2501.18837, 2025.



2 The three axes: security, safety and accuracy

Security, safety, and accuracy are the three foundational pillars for responsibly implementing a GenAI system. Security ensures the system is protected from technical interference, safety ensures the system avoids harmful outputs, and accuracy ensures the system produces correct and reliable results[3].
In the following section, we will examine the primary types of attacks and their associated issues.

[3] International AI Safety Report. **Bengio, Yoshua**. s.l. : arXiv: 2501.17805, 2025.

2.1 Security

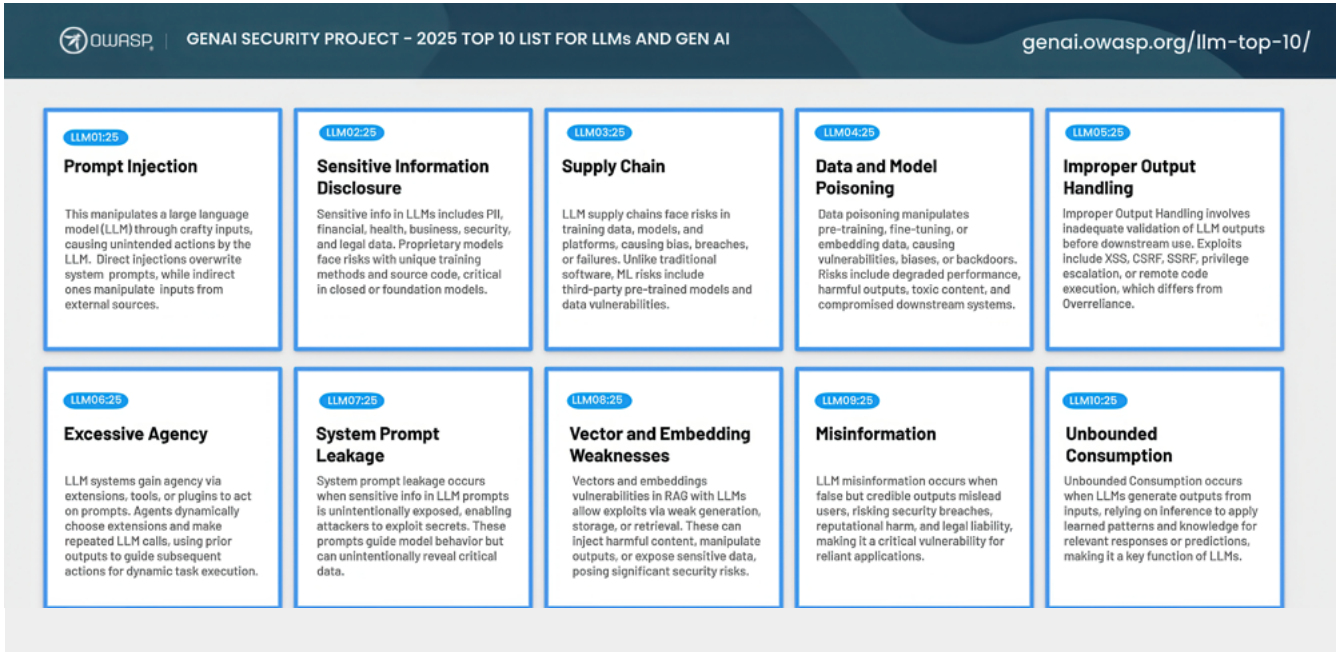
Although tracking the number of attacks on GenAI systems or instances of undesired content leading to economic losses is challenging, notable examples of successful **attacks** include:

- **Exposure of sensitive data**[4]
- **Misinformation**[5]
- **Misleading information**[6]

The rapid evolution of the field makes maintaining an accurate list of the main security risks associated with the application of GenAI in industrial applications difficult. Given the rapid evolution of the GenAI landscape and the increasing popularity of new applications, new vulnerabilities are emerging.

In 2023, the Open Worldwide Application Security Project (OWASP) created the first version of the "Top 10 for LLM Applications", and at the end of 2024 it delivered an updated version for 2025[7] - see Figure 1.

Figure 1. OWASP Top 10 List for LLM and GenAI[7]



[4] Ray, Siladitya. Samsung bans ChatGPT among employees after sensitive code leak. [Online] 2023. [Cited: 02 12, 2025.]
[5] Day, Lewin. Chevy Dealer's AI Chatbot Allegedly Sold A New Tahoe For $1. [Online] 2023. [Cited: 02 12, 2025.]
[6] Yagoda, Maria. Airline held liable for its chatbot giving passenger bad advice. [Online] 2024.
[7] OWASP. Top 10 for Large Language Model Applications. 2025.

The list includes the main security risks introduced in GenAI-based applications and shows mitigation and prevention strategies to address them.

Each vulnerability can be exploited in multiple ways and across various components of a GenAI application.

For example, in RAG applications, Prompt Injection, the Top 1 OWASP vulnerability, can be exploited in:

- The query, by a malicious user who manipulates the input.
- Ingested documents, by injecting jailbreaks into the documents used during the retrieval phase.
- The output, by modifying the response returned by tools called during the generation phase.

Additionally, GenAI vulnerabilities can be introduced in the development or distribution phase of AI models.

The models themselves can be poisoned—for example, an attacker or even the model's creator could modify the internal knowledge, replacing true information with falsehoods about selected topics or introducing semantic backdoors[8].

As the number of AI model providers continues to grow, **model poisoning** is set to become of even greater concern in the coming months given the current commercial and geopolitical landscape. Unfortunately, detecting this type of vulnerability is still an active area of research, and current detection capabilities remain limited.

To further complicate defensive strategies, attacks are not limited to model weights or manipulated model behavior during inference. It is also possible to attack users by exploiting vulnerabilities related to the **data format** used to deliver the models. For example, in the past, models were mostly delivered as Python pickle files, allowing the execution of arbitrary code when loaded[9].

With the increasing adoption of GenAI solutions and their integration into **corporate environments**, an unsafe implementation can pose risks to other systems. In 2025, the number of agentic GenAI applications will increase, with some capable of communicating with both internal company resources and external systems (e.g., internal knowledge bases, websites, APIs)[1]. These new capabilities will increase the security concerns regarding GenAI applications. For example, a simple chatbot, with the ability to send parallel HTTP requests could be exploited to launch DOS attacks against internal and external systems, or manipulated, through prompt injection, to actively log user chats with external systems. These attacks are difficult to detect, and even large corporations may not always be able to foresee all the

[8] Watch Out for Your Agents! Investigating Backdoor Threats to LLM-Based Agents. Yang, Wenkai, Xiaohan Bi, Yankai Lin, Sishuo Chen, Jie Zhou, and Xu Sun. s.l. : arXiv, 2024.

[9] Python. Pickle. [Online] [Cited: 02 12, 2025.]

possible implications of integrating new features into their systems. For example, with the introduction of OpenAI's new Operator agent, several GenAI security experts are raising concerns about its security. A key emerging threat is the potential for attackers to introduce malicious visual components in websites, which could manipulate the agent's behavior.

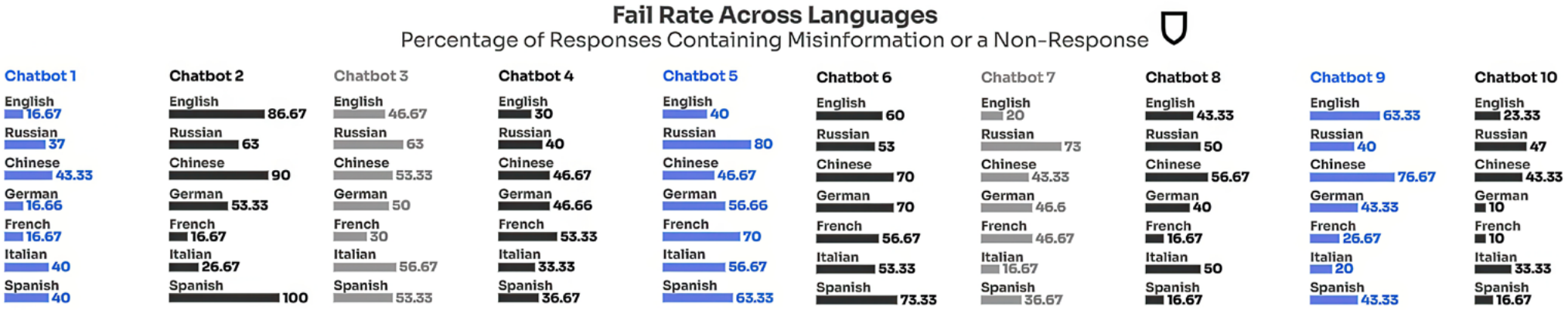
For these reasons, at CRIF, we only introduce new features and new capabilities into our GenAI systems after a thorough verification of the security implications.

2.2 Safety

Although safety constraints are generally introduced during the training of AI models to align them with safety policies, models can still be manipulated to generate instructions that pose hazards. A **hazard** is defined as a "source or situation with a potential for harm: A negative event or negative social development entailing value damage or loss to people"[10]. In real-world AI applications, several incidents have occurred over recent years, and the number is on the rise. An analysis of the AI Incident Dataset (AIID)[11] [12], a public repository tracking incidents that occur in AI applications, revealed that in 2024, the number of incidents increased by nearly 50% compared with 2023.

Furthermore, the distribution of incident type changes from year to year. For example, in 2023, the number of incidents related to misinformation nearly tripled, and NewsGuard's Monthly AI Misinformation Monitor has confirmed the propensity of leading LLMs to spread misinformation across multiple languages[13] [14] - see Figure 2.

Figure 2. Percentage of responses containing misinformation or a non-response. Source: NewsGuard January 2025 AI's Multilingual Failure[14]



[10] Standard model process for addressing ethical concerns during system design. 2024. ISO/IEC/IEEE 24748-7000.

[11] AI Incident Database. [Online] [Cited: 12 02, 2025.]

[12] MIT. AI Risk Repository. [Online] [Cited: 02 12, 2025.]

[13] NewsGuard. [Online] [Cited: 12 02, 2025.]

[14] AI Misinformation Monitor of Leading AI Chatbots Multilingual Edition. NewsGuard. [Online] 2025.

The hazards associated with AI models are evolving over time, and the definition of these hazards in the context of AI applications and related taxonomy is currently being studied by a number of organizations.

For example, with the delivery of the AILuminate v1.0 benchmark, the MLCommons AI Safety Working Group has identified 12 hazards, grouped into three main categories[15]:

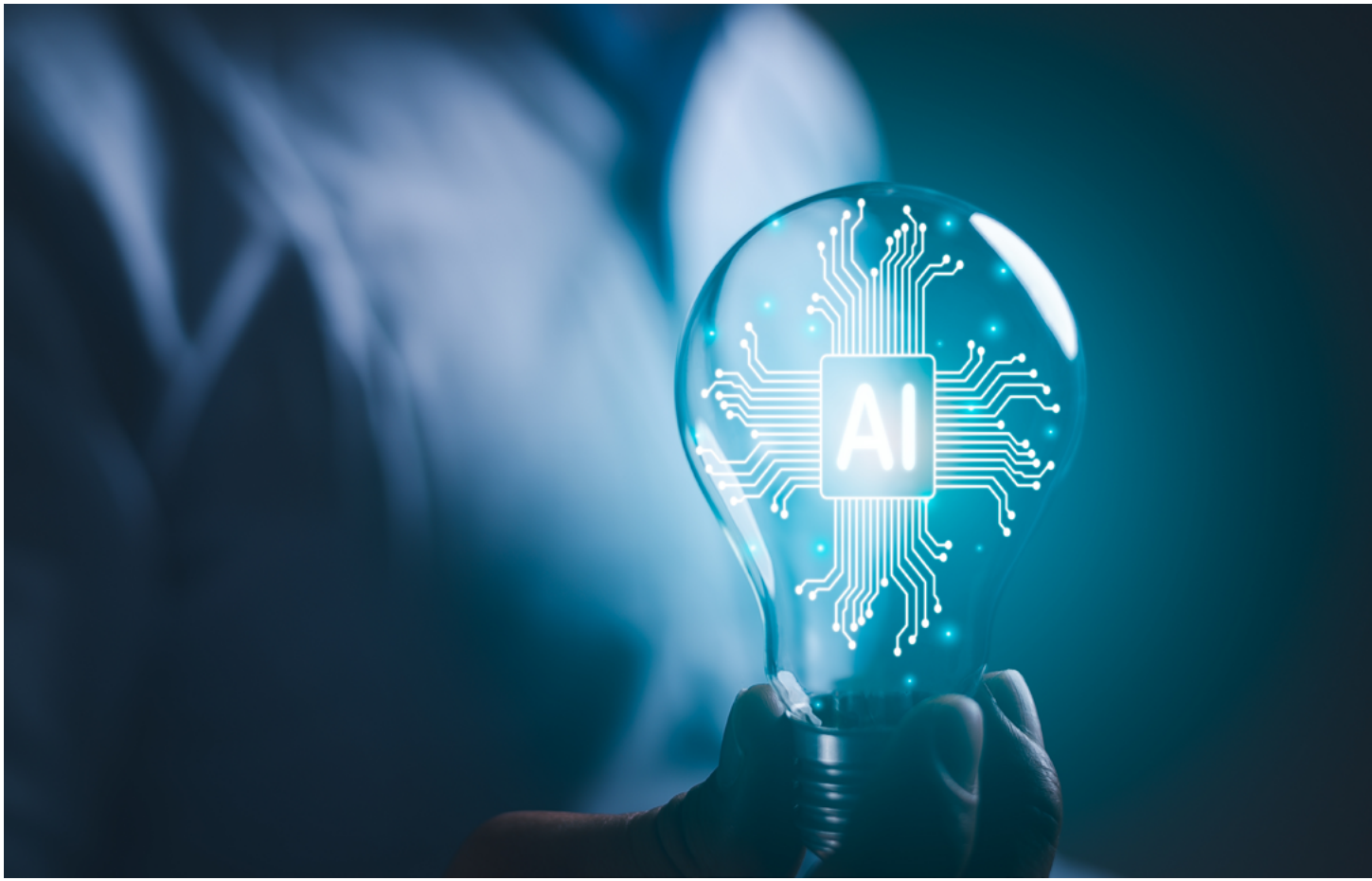
Physical Hazards: Potential to cause physical harm to users or the public

Non-Physical Hazards: Unlikely to cause physical harm but may still be criminal in nature and pose risks to individuals or society

Contextual Hazards: Could cause harm in certain contexts but are innocuous in others

These categories will eventually change over time, and the taxonomy will be extended as the AI landscape expands.

The emergence of new capabilities will potentially lead to a rise in new hazards.



[15] ML Commons AI Luminate. Benchmark for general-purpose AI chat model. [Online] 2025.

Creators of moderation systems—like Meta with Llama Guard[16] and Mistral with the Mistral Moderation API[17]—must continuously update their systems to keep pace with these evolutions. In this regard, Meta, with its Llama Guard 3, has developed a model capable of classifying content across all hazards identified in the AI Safety Benchmark v0.5, released by MLCommons in early 2024[18]. It is expected that in the coming months and years, there will be a stronger alignment between moderation systems and standardized taxonomies. These efforts will enable informed decision making regarding the selection of AI models based on their expected safety characteristics. In this regard, MLCommons has performed a detailed evaluation of several AI models using their recently published AILuminate v1.0 dataset, ranking some of the most well-known publicly available AI systems (e.g., Claude, Gemma, Phi, Gemini, GPT, Llama, and Mistral) against all 12 hazards identified in their taxonomy. More models will be added in the future. See Figure 3.

Figure 3. Top 6 ranked AI Systems evaluated using the AILuminate v1.0 dataset[15]

AI SYSTEM	GRADE
Claude 3.5 Haiku 20241022 (API)	VERY GOOD
Claude 3.5 Sonnet 20241022 (API)	VERY GOOD
Gemma 29b	VERY GOOD
Phi 3.5 MoE Instruct (API)	VERY GOOD
Gemini 1.5 Pro (API, with option)	GOOD
GPT-4o (API)	GOOD

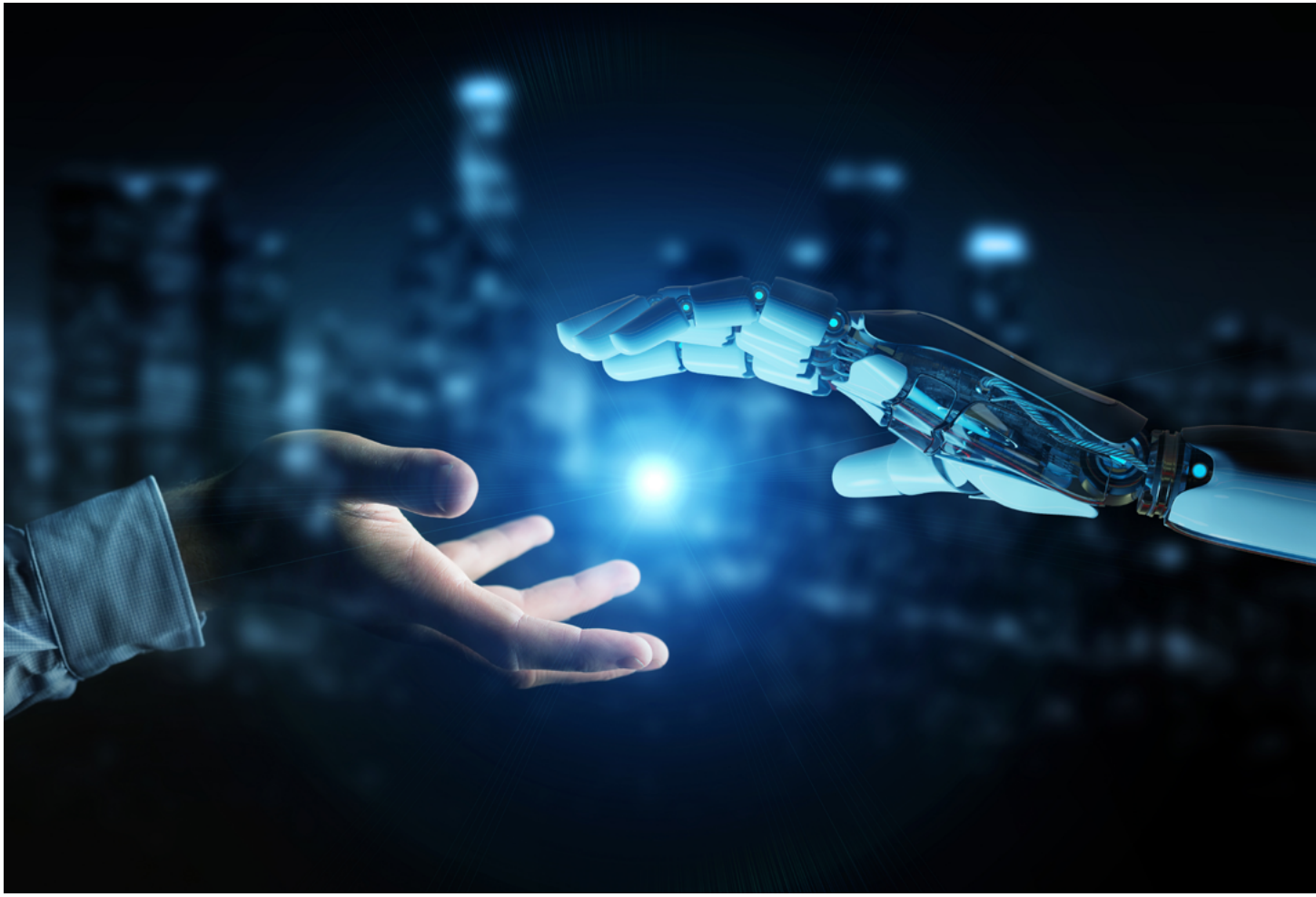
[16] Llama Guard 3 Vision: Safeguarding Human-AI Image Understanding Conversations. Chi, Jianfeng, Ujjwal Karn, Hongyuan Zhan, Eric Smith, Javier Rando, Yiming Zhang, Kate Plawiak, Zacharie Delpierre Coudert, Kartikeya Upasani, and Mahesh Pasupuleti. s.l. : arXiv preprint arXiv:2411.10414, 2024.

[17] Mistral AI. Mistral Moderation. [Online] 2024. [Cited: 02 12, 2025.]

[18] Introducing v0. 5 of the AI Safety Benchmark From MLCommons. Vidgen B, Agrawal A, Ahmed AM, Akinwande V, Al-Nuaimi N, Alfaraj N, Alhajjar E, Aroyo L, Bavalatti T, Bartolo M, Blili-Hamelin B. s.l. : arXiv, 2024.

While it is essential to mitigate the risks of AI models that generate unsafe content, it is equally important to understand the **biases** they acquire during training. For example, a model might refuse to answer a question about a specific topic, yet still have biases related to it. Recent studies show that AI models can have political preferences, but how these biases can impact the decision-making process in other tasks is still being studied. Although some of these models would refuse to suggest political choices, these biases are still embedded in them[19] [20] [21] [22].

These characteristics motivated us to analyze GenAI systems, considering not only their ability to refuse to answer harmful questions but also their biases toward specific topics, in particular those related to the sectors we operate in.



[19] The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation. Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. s.l. : arXiv preprint arXiv:2301.01768, 2023.

[20] Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in Large Language Models. Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Rose Kirk, Hinrich Schütze, and Dirk Hovy. s.l. : arXiv preprint arXiv:2402.16786, 2024.

[21] Hidden Persuaders: LLMs' Political Leaning and Their Influence on Voters. Potter, Yujin, Shiyang Lai, Junsol Kim, James Evans, and Dawn Song. s.l. : arXiv preprint arXiv:2410.24190, 2024.

[22] How would ChatGPT vote in a federal election? A study exploring algorithmic political bias in artificial intelligence. Sullivan-Paul, Michaela. s.l. : Ph.D. thesis, School of Public Policy, University of Tokyo, 2023.

2.3 Accuracy

Evaluating the accuracy of AI solutions is challenging, as it typically involves collecting **ground truth data** and defining **metrics** to evaluate the systems. The promising performance of GenAI systems in **zero-shot learning activities** has driven their development in **data-scarce contexts**, where creating ground truth data is often costly and time-consuming. Unfortunately, as the industry accelerates the adoption of GenAI solutions, it is common to see their application with limited efforts to assess accuracy, even in contexts where acquiring ground truth data is feasible.

Furthermore, GenAI is usually applied to tasks that involve generating **free-form texts** (e.g., chatbots), which are complex to evaluate using standard metrics, given that a correct response can be written in multiple semantically equivalent ways.

In this regard, AI models can facilitate this operation in two different ways:

- **Comparing free-form text ground truths with the outputs generated by GenAI systems**[23]
- **Constructing ground truth data**[24]

With the recent advancements in **AI reasoning capabilities**, using AI systems as expert evaluators is a compelling alternative to traditional human evaluators. While promising, ensuring high-quality content generation remains challenging and requires careful consideration. Nonetheless, these approaches are already being used in multiple applications, ranging from generating high-quality synthetic data used for the pretraining of GenAI models to evaluating Q&A systems.

This is only possible if **multiple reasoning** and **self-assessment steps** are applied consistently during the generation phase. Otherwise, low-quality data may be generated, making its use as ground truth impractical. Although extremely useful and scalable, synthetic generation must not be seen as a replacement for human-made labels, but rather as a complement. The main advantage of synthetic data is its scalability, while human-made labels remain essential for correctly evaluating a GenAI application.

For example, at CRIF we have adopted a hybrid approach where human and synthetic data is used in a coordinated way. In this process, human-generated data contributes to the synthetic data generation process, helping to align outputs with the expected user behavior.

[23] A Survey on LLM-as-a-Judge. Gu J, Jiang X, Shi Z, Tan H, Zhai X, Xu C, Li W, Shen Y, Ma S, Liu H, Wang Y. s.l. : arXiv preprint arXiv:2411.15594, 2024.

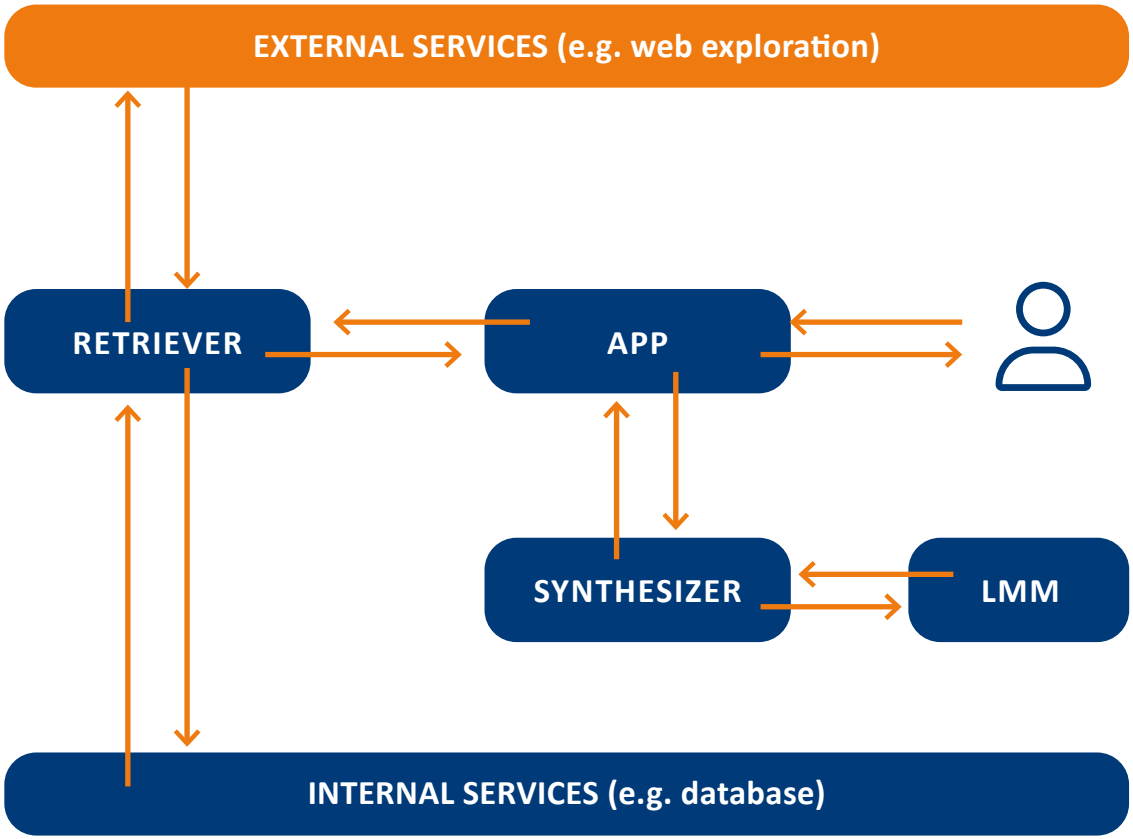
[24] Automated Evaluation of Retrieval-Augmented Language Models with Task-Specific Exam Generation. Guinet, Gauthier, Behrooz Omidvar-Tehrani, Anoop Deoras, and Laurent Callot. s.l. : arXiv preprint arXiv:2405.13622, 2024.

3 The CRIF way

CRIF has extensive experience in developing solutions for regulated sectors, where the application of strict security and privacy policies is essential. We have extended these capabilities into our GenAI solutions, such as the AI Playground, which enables the execution of customer-driven use cases, and the AI Factory, a state-of-the-art backend system that manages, wraps, and enhances the interaction between LLMs and the AI Playground.

We address security concerns in different ways, starting from a **strict supply chain verification** to the **implementation of safety and security layers** that protect GenAI applications from multiple types of attacks, depending on how GenAI is used within the process. We implement both generic and custom solutions to intercept direct attacks on GenAI systems aimed at manipulating the system to generate unsafe or out-of-scope content. Additionally, when GenAI systems require the integration of external knowledge bases, we implement security layers to mitigate the risk of indirect attacks.

Figure 4. Schematic view of a simple RAG application, with its main components.



Let's look at an example of a generic RAG application integrated with internal services (e.g., knowledge stored in a database) and external services (e.g., web exploration through HTTP requests), from which it can extract contextual information during interaction with the user, as shown in Figure 4. Let's also assume that the RAG application includes, in its initial instructions, a mandate to restrict its actions solely to obtaining information related to the application domain.

Even in this simple example, several risks factors must be taken into consideration. First of all, as in any application of this type, data segregation of the internal database may be necessary. The data retrieved from the internal database must be checked before using it to generate the answer to a query. The data could, for example, contain jailbreaks attempts and serve as a channel for an indirect prompt injection attack on the application.

A similar attack could be performed from web exploration. For example, a website could contain jailbreaks inside its HTML code, tailored to attack GenAI apps that analyze website content. Such an attack could be performed by an external entity, not necessarily by the website's creator. Finally, the user could act maliciously, trying to jailbreak the application to perform out-of-scope actions.

Jailbreaking a system like this can have a significant impact not only on the application, but also on internal/external entities. For example, the jailbreak, whether introduced directly or indirectly, could be used to ask the app to perform some specific malicious actions. If the application has access to the company's internal network, it could be used as a proxy to access internal services that are not meant to be accessed, or as a proxy to mask the malicious user's IP address to launch attacks on external services. Additionally, the app, could be used as a proxy to access the underlying model, without going through a valid subscription process.

Furthermore, the model itself could have been poisoned with hidden backdoors, ready to be triggered with specific token sequences. These backdoors could allow the attacker to easily manipulate the model's behavior, making all the malicious activities described above even simpler.

This simple example highlights the importance of protecting GenAI applications, and this can be done in several ways. In RAG applications of this type, for example, we add jailbreak classifiers—among other safeguards—at every step involving the interactions between the user and the app, as well as between the app and other services (internal/external). Additionally, unless explicitly required by the use case, we limit the services accessible by the app to a list of verified, in-scope resources.

To mitigate the risk of model poisoning, we strictly select **widely recognized LLMs** that are provided by **trusted entities**. In some cases, this can limit the applicability of some models, but since no robust backdoor detection mechanism is available at the moment, and may never be, this is the compromise we are making to balance **security** and **quality**.

Additionally, the history of interactions between the user and the system can be logged and used to identify malicious actions. For example, multiple attempts to jailbreak the system.

Our GenAI solutions are **tested against a set of closed test benchmarks**, which are constantly updated based on the evolution of the most critical GenAI vulnerabilities. We implement the **mitigation and prevention policies** recommended by OWASP and rigorously verify their correct implementation through red-teaming activities. Testing is performed both by human operators, who try to manipulate our systems away from their intended behavior, and through the creation of synthetic data to generate large-scale benign and malign user interactions.

Given that we operate in an international context, we work with **data in multiple languages**. To deal with this, we have implemented systems to prevent the exploitation of low-resource languages as a means of bypassing safety measures, which are primarily implemented to detect English-based attacks[25] [26]. For example, as a mitigation strategy, our pipelines include the detection of the user language, and if it is not in the list of those languages used during training of the model, we discard the analysis of the content. Furthermore, we created a pipeline like the one used in[27] for attack generation and we included the possibility to generate attacks in multiple languages.

To mitigate the **cold start problem**, we have invested in developing several pipelines that automatically generate **high-quality ground truth datasets** to verify the accuracy of our GenAI systems throughout their lifecycle. Additionally, these pipelines can be integrated with **human-labeled ground truth datasets** to generate more tailored data.
We consider a **variety of metrics** in the evaluation of our solutions, depending on the use case. For example, for RAG solutions, we consider the relevancy and faithfulness of the answers, as well as their adherence to predefined guidelines (e.g., style, tone, …). This is achieved using an **in-house approach**, with extensive use of an **LLM in an LLM-as-a-judge setup**, similar to the one

[25] Multilingual Jailbreak Challenges in Large Language Models. Deng, Yue, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. 2023 : arXiv preprint arXiv:2310.06474.
[26] LLMs Lost in Translation: M-ALERT uncovers Cross-Linguistic Safety Gaps. Friedrich, Felix, Simone Tedeschi, Patrick Schramowski, Manuel Brack, Roberto Navigli, Huu Nguyen, Bo Li, and Kristian Kersting. s.l. : arXiv preprint arXiv:2412.15035, 2024.
[27] Prompt Security. ps-fuzz. [Online]

described in[28], but with the introduction of **additional reasoning steps** to reduce the probability of unreliable answers. Furthermore, we have developed **custom prompts** to enable the integration of human-labeled ground truth data into the generation process.

Our solutions are implemented considering all the axes described in the previous sections. For example, in the initial model selection phase, we balance reasoning capabilities (e.g., MMLU, MATH-500, GPQA, AIME, …)[29] [30] with the propensity to generate unsafe content (e.g., AILuminate). Specifically, we evaluate the safety of our solutions in accordance with the MLCommons Safety taxonomy and associated recommendations.

Considering the characteristics of the sector we operate in, we have placed particular emphasis on the **analysis of contextual hazards**, and, in particular, on the Specialized Advice subcategory, as we want to avoid our GenAI systems providing specialized advice on critical subjects.

The extensive analysis we have conducted allows us to make an informed selection of the AI models that power our GenAI solutions and the definition of the components of their architectures. Moreover, the analysis allows us to consider the impact of adopting one AI model over another, taking into consideration its costs, performance across our defined evaluation axes, throughput, and latency.

[28] Expect the Unexpected: FailSafe Long Context QA for Finance. Kamble, Kiran, Melisa Russak, Dmytro Mozolevskyi, Muayad Ali, Mateusz Russak, and Waseem AlShikh. s.l. : arXiv preprint arXiv:2502.06329, 2025.
[29] Measuring Massive Multitask Language Understanding. Hendrycks, Dan, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. s.l. : arXiv preprint arXiv:2009.03300, 2020.
[30] GPQA: A Graduate-Level Google-Proof Q&A Benchmark. Rein, David, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. s.l. : arXiv preprint arXiv:2311.12022, 2023.

4 Conclusions

The evolution of the GenAI landscape is offering multiple opportunities in many sectors and applications. There are multiple risks and more will emerge. Adoption must consider several factors that are challenging to measure, and only robust frameworks can mitigate these risks. At CRIF, we have many years' experience in the analysis of sensitive data and in the adoption of safety measures and bias-reduction approaches, even before the recent opportunities offered by GenAI. This experience helps us define and extend our existing safety and security measures within the current GenAI landscape.

crif.com