# On the Validity of Traditional Vulnerability Scoring Systems for Adversarial Attacks against LLMs

Atmane Ayoub Mansour Bahar * and Ahmad Samer Wazan

Posted Date: 30 December 2024

doi: 10.20944/preprints202412.2419.v1

Keywords: Adversarial Attacks; Large Language Models; Vulnerability Metrics; Risk Assessment; Descriptive Statistics

 doi:10.20944/preprints202412.2419.v1



*Article*

# On the Validity of Traditional Vulnerability Scoring Systems for Adversarial Attacks Against LLMs

**Atmane Ayoub Mansour Bahar [1,*] and Ahmad Samer Wazan [2]**

[1] Research Assistant, Algiers, Algeria
[2] College of Technological Innovation Zayed University, Abu Dhabi, United Arab Emirates
* Correspondence: atmane.mansourbahar@gmail.com

**Abstract: Purpose** - This research investigates the effectiveness of established vulnerability metrics, such as the Common Vulnerability Scoring System (CVSS), in evaluating attacks on Large Language Models (LLMs), with a focus on Adversarial Attacks (AAs). The study explores the influence of both general and specific metric factors in determining vulnerability scores, providing new perspectives on potential enhancements to these metrics. **Approach** - This study adopts a quantitative approach, calculating and comparing the coefficient of variation of vulnerability scores across 56 adversarial attacks on LLMs. The attacks, sourced from various research papers, and obtained through online databases, were evaluated using multiple vulnerability metrics. Scores were determined by averaging the values assessed by three distinct LLMs. **Findings** - The results indicate that existing scoring-systems yield vulnerability scores with minimal variation across different attacks, suggesting that many of the metric factors are inadequate for assessing adversarial attacks on LLMs. This is particularly true for context-specific factors or those with predefined value sets, such as those in CVSS. These findings support the hypothesis that current vulnerability metrics, especially those with rigid values, are limited in evaluating AAs on LLMs, highlighting the need for the development of more flexible, generalized metrics tailored to such attacks. **Value** - This research offers a fresh analysis of the effectiveness and applicability of established vulnerability metrics, particularly in the context of adversarial attacks on Large Language Models, both of which have gained significant attention in recent years. Through extensive testing and calculations, the study underscores the limitations of these metrics and opens up new avenues for improving and refining vulnerability assessment frameworks specifically tailored for LLMs.

**Keywords:** adversarial attacks; large language models; vulnerability metrics; risk assessment; descriptive statistics

## 1. Introduction

Large Language Models (LLMs) have recently become a cornerstone in artificial intelligence (AI) research and application, thanks to their remarkable ability to understand and generate human-like text [10]. LLMs such as GPT [106], BERT [30], and others have achieved widespread adoption in a variety of fields, including Natural Language Processing (NLP), machine translation, and conversational AI, due to their capacity to generalize across diverse tasks [133]. However, this surge in popularity has also exposed LLMs to a myriad of vulnerabilities, becoming an attractive target for various security threats [1,50,79].

One of the most significant threats to LLMs is Adversarial Attacks (AAs) [101,119,164], which are typically designed to fool Machine Learning (ML) models by modifying input data or introducing carefully-crafted inputs that cause the model to behave inappropriately [50,130]. These attacks often remain indistinguishable to humans but significantly impact the model's decision-making process, posing a significant threat to LLMs, as they can compromise the integrity, reliability, and security of applications that rely on these models [14]. One significant example is the Crescendo attack [111]. This sophisticated method manipulates LLMs by gradually escalating a conversation with benign prompts

 doi:10.20944/preprints202412.2419.v1



that evolve into more harmful requests, effectively bypassing safety mechanisms. Therefore, protecting LLMs has become a critical concern for researchers and practitioners alike [66,167].

To effectively secure LLMs against AAs, it is crucial to assess and rank these threats based on their severity and potential impact on the model. For instance, some attacks, like Prompt Injection [85], are easy to execute and widely applicable, making them higher-priority threats. Others, like Backdoor attacks [76], may require greater sophistication but can cause significant long-term damage [51]. This prioritization allows security teams to focus on the most dangerous attacks first for mitigation efforts. Existing vulnerability metrics, such as the Common Vulnerability Scoring System (CVSS) [114] and OWASP Risk Rating [140], are commonly used to evaluate the danger level of attacks on traditional systems, taking into account factors such as attack vector, attack complexity, and impact. However, their applicability to LLMs remains questionable.

Most existing vulnerability metrics are tailored for assessing **technical** vulnerabilities in software or network systems. In contrast, AAs on LLMs often target the model's **decision-making** capabilities and may not result in traditional technical-impacts, such as data breaches or service outages [161]. For example, attacks like Jailbreaks [24], which manipulate the model's outputs to bypass ethical or safety constraints, cannot easily be classified as technical vulnerabilities. These attacks focus on manipulating the model's behavior rather than exploiting system-level weaknesses. In other terms, the context-specific factors used in existing metrics, such as CVSS, do not adequately account for the unique characteristics of LLMs or the nature of AAs. Consequently, they may be ill-suited for assessing the risk posed by these attacks on LLMs.

In this study, we aim to evaluate the suitability of known vulnerability metrics in assessing Adversarial Attacks against LLMs. We hypothesize that: ***'the factors used by traditional metrics may not be fully applicable to attacks on LLMs'***, because many of these factors are not designed to capture the nuances of AAs.

To test this hypothesis, we evaluated **56** different AAs across four widely used vulnerability metrics. Each attack was assessed using three distinct LLMs, and the scores were averaged to provide a final assessment. This multi-faceted approach aims to provide a nuanced understanding of how well current metrics can distinguish between varying levels of threat posed by different adversarial strategies, as relying solely on human judgment for security assessments would require domain experts, and human evaluation could introduce biases.

Our findings indicate that average scores across diverse attacks exhibit **low variability**, suggesting that many of the existing metric factors may not offer fair distinctions among all types of adversarial threats on LLMs. Furthermore, we observe that metrics incorporating more generalized factors tend to yield better differentiation among adversarial attacks, indicating a potential pathway for refining vulnerability assessments tailored for LLMs.

The contributions of this paper are fourfold.

- We provide a taxonomy of the various classification criteria of Adversarial Attacks existing in the literature, showing the logic followed in classifying AAs into multiple types.
- We present a list of 56 AAs specifically targeting LLMs, which serve as our test scenarios.
- We provide a comprehensive evaluation of some vulnerability metrics, in the context of AAs targeting LLMs, using differential statistics to analyse the variations of metric scores across different attacks.
- We suggest that future work should focus on developing more general and LLM-specific vulnerability metrics that can effectively capture the unique characteristics of AAs targeting these models.

This paper is structured in seven parts. We start in Section 2 by detailing the procedures we employed in this study, especially concerning the data collection, vulnerability assessments through LLMs, and mathematical analysis of the results. After that, we present in Section 3 an overview of AAs and their existing classifications. In Section 4, we present a detailed list of AAs on LLMs, and propose a classification based on the danger level in Section 5. Sections 6, 7, and 8 encompasses respectively,

 doi:10.20944/preprints202412.2419.v1



the evaluation of the vulnerability metrics on LLMs, the discussion of the results, and the perspectives for future enhancements.

## 2. Methods

In this section, we outline the methodology adopted to evaluate vulnerabilities in attacks targeting Large Language Models using established metrics such as DREAD [92], CVSS [114], OWASP Risk Rating [140], and Stakeholder-Specific Vulnerability Categorization (SSVC) [128].

Our approach involves three key steps depicted below in Figure 1 : data collection, assessment, and statistical interpretation.

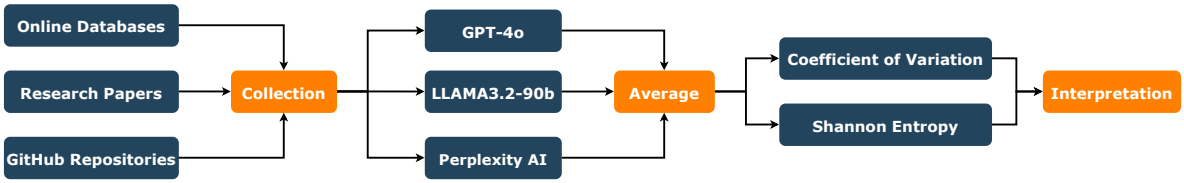



**Figure 1.** Research process

### *2.1. Data collection*

The first step in our methodology was to gather a comprehensive dataset of AAs targeting LLMs. To ensure a thorough and systematic approach, we began by reviewing the literature on these attacks, exploring existing types and classifications. This step provided a broad understanding of the main categories of attacks commonly observed in the context of ML and NLP systems.

Following this foundational review, we focused on identifying recent AAs specifically targeting LLMs. These attacks were grouped into **seven** primary types: Jailbreaks (White-box and Black-box) [150], Prompt Injections [84], Evasion attacks [137], Model-Inference (Membership Inference) attacks [57], Model-Extraction attacks [49], and Poisoning/Trojan/Backdoor attacks [76,88,131]. For each type, we selected **eight** representative attacks, prioritizing those published in recent research or demonstrated in practical scenarios. This effort resulted in a list of 56 attacks, covering a diverse range of threat vectors and methodologies.

To enable a systematic ranking of these attacks based on their potential danger, we decided to assess each attack using **vulnerability metrics**. By applying multiple metrics, we aimed to provide a multi-faceted evaluation of each attack's severity and to ensure that the dataset would serve as a robust basis for further analysis and interpretation.

### *2.2. Score assessments*

To evaluate the severity and danger level of the 56 gathered attacks, we began by identifying widely recognized vulnerability assessment metrics to ensure a comprehensive analysis. After careful consideration, we selected four metrics: DREAD [92], CVSS [114], OWASP Risk Rating [140], and SSVC [128]. These metrics were chosen for their broad adoption and their focus on different factors, enabling a more nuanced understanding of the vulnerabilities. Since the Adversarial Attacks we collected are recent and not yet assessed in the literature, calculating their scores became essential to address this gap.

Manually assessing 56 attacks across four metrics is a daunting task, requiring extensive **expertise** from security analysts, system administrators, and other domain experts. The process involves interpreting complex scenarios, considering varying factors for each metric, and ensuring consistency between all evaluations. Completing such an effort manually could take months or even years, which is impractical given the fast-evolving nature of adversarial threats.

To overcome this challenge and accelerate the process, we leveraged the capabilities of LLMs to perform **semi-automated scoring**. Specifically, we utilized three state-of-the-art models: GPT-4o [97], LLAMA3.2-90b [38], and Perplexity AI [62]. Each model operated independently, assessing the attacks

 doi:10.20944/preprints202412.2419.v1



and vulnerabilities according to the factors defined by the selected metrics. For each scoring factor, we calculated the average score provided by the three LLMs, rounded to the closest unit.

This approach offers several advantages. First, it enables **rapid** assessments. Second, using multiple LLMs increases the robustness of the results by **minimizing biases** or errors from any single model. Furthermore, the models' advanced text-processing capabilities allow them to analyze the **contextual details** of each attack and provide scores that align with the logic of the vulnerability metrics.

A recent work of Chopra et al. [23] proves that LLMs are able to identify and analyze software vulnerabilities; but that they can lead to misinterpretations or oversights in understanding complex vulnerabilities. To address such potential inconsistencies in the assessments, we incorporated a **Human-in-the-Loop (HitL)** verification process. We reviewed the logic and reasoning behind each LLM-provided score to ensure its accuracy and reliability. This step was essential to mitigate any errors or misinterpretations that might arise from the LLMs, especially when handling complex scenarios.

To validate this methodology, we tested it on a set of Common Vulnerabilities and Exposures (CVEs) that already have human validated scores with both CVSS and SSVC [128] to measure the gap, as shown in Table 1. The details of each factor are further explained in Section 5.2.

The results demonstrate that our approach of aggregating the assessments of three LLMS yields scores closely aligned with existing assessments, with few differences related mainly to the advancements of technologies from the first assessment of those vulnerabilities to today. For instance, vulnerabilities such as `'CVE-2015-5374'` and `'CVE-2019-9042'` became less active than before, making their exploitation value with SSVC change from *Active* to *Proof-of-Concept* (refer to Section 5.2.4 for more details).

This experiment also shows that combining the computational efficiency of LLMs with human oversight represents a practical solution for scoring new and unassessed attacks in the absence of readily available experts. This innovative approach not only saves time but also ensures a balanced and consistent evaluation process, enabling a deeper understanding of vulnerabilities and their potential impact.

**Table 1.** Comparison between some existing and LLM-generated CVSS and SSVC values

| CVE-ID | SSVC Values | CVSS (2.0 or 3.0) Values |
|---|---|---|
| **CVE-2014-0751** | **NVD:** E:N/U:L/T:T/P:S | **NVD:** AV:N/AC:L/Au:N/C:P/I:P/A:P |
| | **LLMs:** E:N/U:L/T:P/P:M | **LLMs:** AV:N/AC:L/Au:N/C:P/I:P/A:P |
| **CVE-2015-1014** | **NVD:** E:N/U:L/T:T/P:S | **NVD:** AV:L/AC:L/PR:L/UI:R/S:U/C:H/I:H/A:H |
| | **LLMs:** E:N/U:E/T:T/P:S | **LLMs:** AV:L/AC:L/PR:L/UI:R/S:U/C:H/I:H/A:H |
| **CVE-2015-5374** | **NVD:** E:A/U:L/T:P/P:S | **NVD:** AV:N/AC:L/Au:N/C:N/I:N/A:C |
| | **LLMs:** E:P/U:L/T:P/P:S | **LLMs:** AV:N/AC:L/Au:N/C:N/I:N/A:C |

 doi:10.20944/preprints202412.2419.v1



**Table 1.** *Cont.*

| CVE-ID | SSVC Values | CVSS (2.0 or 3.0) Values |
|---|---|---|
| **CVE-2017-3183** | **NVD:** E:N/U:E/T:T/P:M | **NVD:** AV:N/AC:L/PR:L/UI:N/S:U/C:H/I:H/A:H |
| | **LLMs:** E:N/U:E/T:P/P:M | **LLMs:** AV:N/AC:L/PR:L/UI:N/S:U/C:H/I:H/A:N |
| **CVE-2017-5638** | **NVD:** E:A/U:S/T:T/P:M | **NVD:** AV:N/AC:L/PR:N/UI:N/S:U/C:H/I:H/A:H |
| | **LLMs:** E:A/U:S/T:T/P:S | **LLMs:** AV:N/AC:L/PR:N/UI:N/S:U/C:H/I:H/A:H |
| **CVE-2017-9590** | **NVD:** E:P/U:E/T:T/P:M | **NVD:** AV:N/AC:H/PR:N/UI:N/S:U/C:H/I:N/A:N |
| | **LLMs:** E:P/U:E/T:P/P:M | **LLMs:** AV:N/AC:H/PR:N/UI:N/S:U/C:H/I:N/A:N |
| **CVE-2018-14781** | **NVD:** E:P/U:L/T:P/P:M | **NVD:** AV:A/AC:H/PR:N/UI:N/S:U/C:N/I:H/A:N |
| | **LLMs:** E:P/U:L/T:P/P:M | **LLMs:** AV:A/AC:H/PR:N/UI:N/S:U/C:N/I:H/A:N |
| **CVE-2019-2691** | **NVD:** E:N/U:E/T:P/P:M | **NVD:** AV:N/AC:L/PR:H/UI:N/S:U/C:N/I:N/A:H |
| | **LLMs:** E:N/U:S/T:P/P:M | **LLMs:** AV:N/AC:L/PR:H/UI:N/S:U/C:N/I:N/A:H |
| **CVE-2019-9042** | **NVD:** E:A/U:L/T:T/P:M | **NVD:** AV:N/AC:L/PR:H/UI:N/S:U/C:H/I:H/A:H |
| | **LLMs:** E:P/U:L/T:T/P:M | **LLMs:** AV:N/AC:L/PR:N/UI:N/S:U/C:H/I:H/A:H |

### *2.3. Results interpretations*

Our approach provided a multi-dimensional analysis of Adversarial Attacks against LLMs by leveraging four distinct vulnerability assessment metrics: DREAD, CVSS, OWASP Risk Rating, and SSVC. This comprehensive evaluation allowed us to gain a broad perspective on how these metrics reflect the severity and impact of attacks, as well as their usefulness in ranking and understanding vulnerabilities in the LLM context.

To assess the utility and added value of each factor within the metrics, we analyzed their **variability** across the 56 attacks, grouped by attack type. For the **quantitative metrics** (DREAD and OWASP Risk Rating), we calculated the **coefficient of variation (CV)** for each factor to measure the relative dispersion of scores. For the **qualitative metrics** (CVSS and SSVC), we used **entropy** [118] to quantify the diversity or uniformity of categorical values.

## 3. Adversarial Attacks

The rise of AAs in the field of Machine Learning has posed significant security challenges, especially for Large Language Models. These attacks exploit the vulnerabilities inherent in AI models by manipulating inputs to achieve unintended or harmful outputs. This section provides a detailed exploration of AAs, beginning with their formal definition and an analysis of why they are considered particularly dangerous to LLMs. Then it introduces various types and classifications of AAs, offering insight into the range of attack strategies used to compromise LLMs. Understanding these elements is crucial for designing more robust defenses and enhancing the security of AI-driven systems.

### *3.1. Definition*

Adversarial Attacks are **intentional manipulations** of input data designed to exploit vulnerabilities in ML models [43]. The concept of adversarial examples was first introduced in the domain of Image Recognition by Szegedy [130], and it has since been widely explored across different ML tasks, including NLP [33,105,161]. In the context of LLMs, adversarial inputs are carefully crafted to cause the model to produce incorrect, biased, or harmful **outputs** [68]. Unlike traditional errors, AAs are not random; but are **strategically** designed to exploit the decision boundaries of models by

 doi:10.20944/preprints202412.2419.v1



altering inputs in ways imperceptible to humans and effective against ML models [13]. These attacks can involve minimal changes, such as swapping words, inserting seemingly harmless phrases, or restructuring sentences, that lead to dramatically different responses from the model, often having severe real-world consequences [61,67], particularly in safety-critical applications such as autonomous driving, healthcare diagnostics, and security systems [100]. For example, an Adversarial Attack could lead an autonomous vehicle to misinterpret road signs, resulting in catastrophic accidents [41,165].

On top of that, AAs can come in various forms, each exploiting different aspects of LLMs. These attacks can be broadly categorized based on the attacker's knowledge, the nature of the perturbations, and the model's vulnerability. The following section will explore the different types and classifications of AAs, showing that each type has distinct strategies and potential impacts on LLMs.

### *3.2. Classifications of AAs*

Adversarial Attacks have been classified in various ways in the literature, offering different perspectives on how AAs operate and their potential impact on Machine Learning models. In this section, we have gathered the most common classifications of AAs, based on criterias such as their purpose, target, the attacker's knowledge and strategy, life-cycle stages, CIA[1] triad, and the type of data and control involved. We depict these classifications in Figure 2, and each one will be discussed in detail in the following subsections.

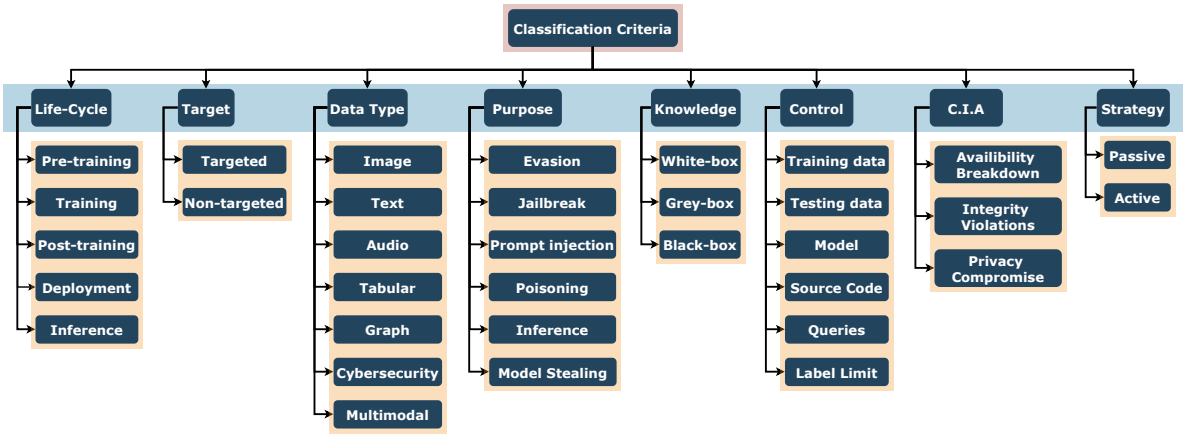



**Figure 2.** Taxonomy of the classification criterias of Adversarial Attacks

#### 3.2.1. Based on the Purpose

One of the most widely used ways to classify AAs is by analyzing their **intended purpose** [9,27,142]. Attacks can be designed either to evade detection by a model or to cause intentional misclassifications, thereby compromising the system's integrity or exploiting its weaknesses. Based on these overarching objectives, several distinct types of AAs have emerged, including: Evasion attacks [76,88,137], Jailbreak attacks [150], Prompt Injections [84], Model Inference attacks [57], Model Extraction (Stealing) attacks [49], and Poisoning/Trojan/Backdoor attacks [76,88,131]. Each type targets different aspects of a ML system, posing unique challenges to the robustness and security of the models.

**Evasion attacks:**

In evasion attacks, adversaries craft inputs that evade detection or mislead the model into making incorrect classifications. For instance, small changes in an image may lead a computer vision model to misclassify it, while adversarial inputs in LLMs can bypass content filters [3,4,21].

[1] Confidentiality, Integrity, and Availability

 doi:10.20944/preprints202412.2419.v1



**Model jailbreaking:**

In jailbreak attacks, the attacker manipulates the model to bypass restrictions or constraints set by the system such as ethical filters. For example, bypassing content filters in a chatbot by providing a carefully crafted prompt that tricks the model into generating restricted outputs [28,29].

**Prompt injections (PI):**

In prompt injections, the attacker provides maliciously designed prompts that cause the model to follow unintended instructions or generate harmful outputs. Unlike jailbreaking, PI typically involves inserting harmful instructions within regular inputs rather than overriding system-level restrictions [141]. An example is injecting hidden instructions within user input to manipulate a language model's behavior in ways not intended by the developers [70,84,87].

**Model inference:**

In Model (or Membership) Inference attacks, adversaries aim to determine whether specific data was part of the training set. By analyzing model outputs, they can infer sensitive or proprietary information from the training data, posing significant privacy risks [107,127].

**Model Extraction:**

Model Extraction (or Stealing) involves probing a black-box model to reconstruct its functionality or recover sensitive data. For example, an attacker could steal a proprietary financial model by systematically querying it and analyzing its responses [121,157].

**Poisoning/Trojan/Backdoors:**

These attacks aims at the integrity of the model during the training phase. The attacker injects malicious data (poisoning) or patterns (trojan, backdoor) into the training set to influence the model's behavior at inference time. For instance, a Poisoning scenario would be introducing mislabeled data to reduce model accuracy [155], and a Trojan scenario would be embedding a hidden trigger in the training data to activate malicious behavior later, such as a trigger that could cause a traffic-light recognition model to classify a red light as a green light in autonomous driving cars [32].

3.2.2. Based on the Target

Adversarial attacks can also be classified based on their target, which refers to whether the attack is aimed at causing a specific or arbitrary misclassification [27].

**Targeted:**

In targeted attacks, the attacker aims to manipulate the model into misclassifying an input into a specific, incorrect class [15]. For example, an attacker might craft an input to make a stop sign consistently classified as a yield sign.

**Non-targeted:**

In non-targeted attacks, the goal is to cause the model to misclassify the input, but the specific incorrect class is irrelevant to the attacker [143]. For instance, an adversarial input could cause a stop sign to be classified as any incorrect traffic sign.

3.2.3. Based on the Attacker's Knowledge

Another existing classification is categorizing AAs by the amount of knowledge the attacker has about the target model [98,142]. These categories typically include white-box, black-box, and sometimes grey-box attacks, although grey-box is not always explicitly classified.

 doi:10.20944/preprints202412.2419.v1



**White-box:**

In white-box attacks, the attacker has full access to the model's architecture, parameters, and training data, allowing them to exploit the model's gradients for highly effective adversarial examples. For instance, using gradient-based methods, an adversary can precisely manipulate inputs to deceive the model [54,81].

**Black-box:**

In black-box attacks, the attacker has no direct access to the model's internals and can only interact with it by sending queries and observing outputs. Despite this limitation, attackers can use techniques like transfer learning, where adversarial examples generated on a surrogate model are used to attack the target model [60,83].

**Grey-box:**

In grey-box attacks, the attacker has partial knowledge of the model, such as knowing the architecture but lacking access to the exact parameters or training data. These attacks may combine both white-box and black-box techniques to exploit vulnerabilities effectively [90,148].

3.2.4. Based on the Life-Cycle

Adversarial attacks can be categorised also by when they occur in the machine learning pipeline, with some references focusing on the training and deployment phases only [98], and others adding phases such as pre-training, post-training, and inference phase [144].

**Pre-training:**

Pre-training attacks are conducted before the model training begins, often during the data collection phase. For example, poisoned-data injection into a dataset to compromise the model's integrity once training commences [73,80].

**Training:**

Training-phase attacks occur during the actual model training process. A notable example is backdoor injection, where adversaries embed specific triggers in the training data to manipulate the model's behavior later [35,147].

**Post-training:**

Post-training attacks take place immediately after the training process concludes, before the model is deployed. These attacks might involve modifying the model parameters in a way that alters its predictions without detection [104,163].

**Deployment:**

Deployment-phase attacks are executed after the model has been deployed on a hardware device, such as a server or mobile device. An example includes modifying model parameters in memory through techniques like bit-flipping, which can lead to unexpected behaviors [6,20].

**Inference:**

Inference attacks are performed by querying the model with test samples. A specific instance is backdoor activation, where an adversary triggers the model's malicious behaviors by providing inputs that match the previously embedded backdoor conditions [34,69].

3.2.5. Based on the CIA Violation

A fifth classification of AAs is made according to the targeted aspect of the CIA triad, which encompasses confidentiality, integrity, and availability violations [98,112].

 doi:10.20944/preprints202412.2419.v1



**Availability breakdown:**

In availability breakdown attacks, the attacker aims to degrade the model's performance during testing or deployment. This can involve energy-latency attacks that manipulate queries to exhaust system resources, leading to denial of service or reduced responsiveness [8,124].

**Integrity violations:**

Integrity violation attacks target the accuracy and reliability of the model's outputs, resulting in incorrect predictions. For instance, poisoning attacks during training can introduce malicious data, causing the model to produce erroneous results when deployed [47,53].

**Privacy compromise:**

Privacy compromise attacks focus on extracting sensitive information about the model or its training data. Model-extraction attacks exemplify this by allowing an adversary to reconstruct the model's functionalities or retrieve confidential data used during training [18,63].

3.2.6. Based on the Type of Control

Adversarial attacks are classified in other sources based on the type of control the attacker exerts over various elements of the ML model, with some highlighting the control of training and testing data [112], and others [98] proposing more aspects of control, such as the control of the model, source code, and queries, as well as a limited control on the data labels.

**Training data:**

In training data attacks, the attacker manipulates the training dataset by inserting or modifying samples. An example is data poisoning attacks, where malicious inputs are added to influence the model's learning process [138].

**Testing data:**

Testing data attacks involve altering the input samples during the model's deployment phase. Backdoor poisoning attacks serve as an example, where specific triggers are embedded in the testing data to manipulate the model's predictions under certain conditions [113].

**Model:**

Model attacks occur when the attacker gains control over the model's parameters, often by altering the updates applied during training. This can happen in Federated Learning (FL) environments, where malicious model updates are sent to compromise the integrity of the aggregated model [132].

**Source code:**

Source code attacks involve modifying the underlying code of the model, which can include changes to third-party libraries, especially those that are open source. This allows attackers to introduce vulnerabilities directly into the model's functionality [159].

**Queries:**

Query-based attacks allow the attacker to gather information about the model by submitting various inputs and analyzing the outputs. Black-box evasion attacks exemplify this, as adversaries attempt to craft inputs that evade detection while learning about the model's behavior through its responses [42].

**Label limit:**

In label limit attacks, the attacker does not have control over the labels associated with the training data. An example is clean-label poisoning attacks, where the adversary influences the model without altering the labels themselves, making detection more difficult [116].

 doi:10.20944/preprints202412.2419.v1



3.2.7. Based on the Type of Data

An seventh classification of Adversarial attacks is based on the type of data they target, highlighting the diverse methodologies employed across different modalities. Some underline attacks targeting data types as images, text, tabulars, cybersecurity, and even multimodal [98], while other works mention attacks on audio data [16], and graph-based data [25].

**Image:**

In image-based attacks, the attacker crafts adversarial images designed to cause misclassification. An example includes perturbing images to deceive object detectors or image classifiers, leading to incorrect identification [129].

**Text:**

Text attacks involve modifying text inputs to mislead NLP models. For instance, an adversary might introduce typos or antonyms to trick sentiment analysis tools or text classifiers into generating false outputs [46].

**Tabular:**

Tabular data attacks target models that operate on structured data, often seen in applications like finance or healthcare. A common example is poisoning attacks, where malicious entries are inserted into tabular datasets to manipulate model behavior [17].

**Audio:**

Audio-based attacks involve crafting adversarial noise or altering audio inputs to cause misclassification in systems like voice recognition. For example, specific sound patterns can be designed to mislead voice-activated systems, resulting in incorrect command interpretations [78].

**Graphs:**

Graph-based attacks manipulate graph structures and attributes to deceive Graph Neural Networks (GNNs). An attacker might alter edges or node features to induce misclassification or misleading outputs from graph-based models [93].

**Cybersecurity:**

In the cybersecurity domain, AAs target systems like malware detection or intrusion detection systems. An example is poisoning a spam email classifier, where attackers introduce deceptive emails to degrade the model's performance [135].

**Multimodal:**

Multimodal attacks involve exploiting systems that integrate multiple data types. In these cases, attackers might gain insights by submitting queries that encompass different modalities, such as text and image combinations [145].

3.2.8. Based on the Strategy

Last but not least, Adversarial attacks can also be categorized based on the strategy employed by the attacker, distinguishing between passive and active approaches [112].

**Passive:**

In passive attacks, the attacker seeks to gather information about the application or its users without actively interfering with the system's operation. An example is reverse engineering, where an adversary analyzes a black-box classifier to extract its functionalities and gain insights into its behavior [22].

 doi:10.20944/preprints202412.2419.v1



**Active:**

Active attacks are designed to disrupt the normal functioning of an application. The attacker may implement poisoning attacks that introduce malicious inputs, aiming to trigger misclassifications or degrade the model's performance during operation [59].

## 4. Adversarial Attacks on LLMs

In recent years, LLMs have been increasingly targeted by AAs [68,119,154], posing various threats to their reliability, safety, and security. These attacks can take multiple forms and serve distinct purposes, each exploiting different vulnerabilities within the model or its deployment. In this section, we present a comprehensive taxonomy of 56 recent AAs targeting LLMs, following the purpose-based classification of AA (refer to Section 3). We consider 7 types of AAs: White-box Jailbreak attacks, Black-box Jailbreak attack, Prompt Injection, Evasion Attacks, Model Extraction, Model Inference, and Poisoning/Trojan/Backdoor. Each attack type includes 8 prominent examples, which are detailed in the following subsections.

### *4.1. Jailbreak Attacks*

The type of AAs that we begin with are model Jailbreaking attacks, which are designed to bypass safety measures. We consider two approaches in jailbreak attacks according to the **targeted model**: White-box, and Black-box model jailbreaking.

#### 4.1.1. White-box attacks

The first type are White-box Jailbreak attacks, where the attacker has **complete** access to the model's architecture, parameters, and training data. This level of knowledge allows the attacker to design specific inputs that exploit vulnerabilities in the model, often related to the model gradients, in order to bypass its restrictions or safety measures.

#### 4.1.2. Black-box attacks

The second type of attacks are Black-box Jailbreak attack, in which, in contrast to white-box attacks, the attacker has **no access** to the model's internal workings or training data. Instead, the attacker can only interact with the model by providing inputs and observing the outputs, often relying on trial and error to discover effective prompts able to bypass the model's safeguards.
We present in Table 2 a list of recent white-box and black-box jailbreak attacks existing in the literature [125], and if they are Open Source (OS) or not, as each has a different strategy and implementation.

 doi:10.20944/preprints202412.2419.v1



**Table 2.** Examples of jailbreak attacks against LLMs

| Type | Attack | Concept | OS? |
|---|---|---|---|
| W-box | **GCG** [166] | Adding adversarial suffixes using greedy and gradient-based searches | ✓ |
| | **Visual Mod.** [96] | Jailbreaking an LLM using a corresponding Multimodal LLM | ✗ |
| | **PGD** [48] | Jailbreaking attack using Projected Gradient Descent | ✗ |
| | **SCAV** [149] | Guiding Jailbreak attacks against white-box LLMs | ✗ |
| | **Soft Prp.** [115] | Attacking the continuous embedding representation of input tokens | ✓ |
| | **DrAttack** [74] | Decomposition and Reconstruction of prompts for LLM jailbreaking | ✓ |
| | **RADIAL** [36] | Generating instructions based on LLMs' Inherent Response Tendency | ✗ |
| | **ReNeLLM** [31] | Using generalized and nested jailbreak prompts to fool LLMs | ✓ |
| B-box | **PAIR** [19] | Automatic jailbreaking of black box LLMs | ✗ |
| | **Privacy att.** [71] | Extracting people-information memorised by GPT-4o | ✓ |
| | **DAN** [120] | Tricking GPT-4o to break its policies with a role-play | ✓ |
| | **Ad. Att.** [2] | Adding adversarial suffixes using random searches | ✓ |
| | **GCQ** [56] | Enhancing GCG algorithm using best-first search algorithm | ✗ |
| | **PAL** [126] | Token-level attack using gradients from an open-source proxy | ✓ |
| | **IRIS** [108] | Using the same LLM to target itself | ✗ |
| | **Tastle** [146] | Framework of black-box jailbreak for automated red-teaming | ✗ |

### *4.2. Prompt Injection*

The third type of attacks that we illustrate are Prompt injections, where the adversary manipulates the input prompts and queries to deceive the model into producing unintended or harmful outputs. This technique is ranked among the most dangerous attacks against LLMs by OWASP [99]. To illustrate the diverse strategies attackers employ to exploit LLMs with PIs, we have gathered eight different attacks, utilizing both direct injections, where the attacker append a malicious input to a prompt, and indirect injection methods, where the attacker append malicious prompts through file or external inputs. These attacks are presented in Table 3

 doi:10.20944/preprints202412.2419.v1



**Table 3.** Examples of prompt injection attacks against LLMs

| Attack | Concept | OS? |
|---|---|---|
| **Ign. Pp.** [102] | A direct prompt injection technique to mislead the LLM in ignoring instructions | ✓ |
| **Ind. PI** [52] | An indirect prompt injection technique through file input to compromise LLMs | ✓ |
| **Frm. PI** [86] | General framework for formalizing prompt injection in LLMs | ✓ |
| **Mlt. PI** [5] | Using images and sounds for indirect prompt injection in multi-modal LLMs | ✓ |
| **Unv. PI** [82] | An automatic and indirect prompt injection attack | ✓ |
| **Vrt. PI** [152] | Backdooring a prompt injection under a triggered scenario | ✓ |
| **Chat Tmp.** [139] | Creating misleading contexts acceptance elicitation and word anonymization | ✗ |
| **JudgeDeceiver** [122] | Deceiving LLM-as-a-Judge to choose a response among multiple choices | ✗ |

### *4.3. Evasion Attacks*

The forth type of attacks we illustrate are Evasion attacks, in which attackers aim to deceive language models by crafting inputs designed to bypass detection or classification. These attacks often target sentiment analysis and text classification models, seeking to manipulate their outputs through subtle modifications. In Table 4, we have gathered eight different examples and techniques of evasion attacks presented in the literature, some of which employ text perturbations to alter the original input, while others leverage LLMs to generate sophisticated evasion samples against their counterparts.

**Table 4.** Examples of evasion attacks against LLMs

| Attack | Concept | OS? |
|---|---|---|
| **Hot-Flip** [39] | Flipping letters in a word to mislead the LLM to make incorrect classifications | ✗ |
| **PWWS** [109] | Changing some words with their synonyms to mislead text classification tasks | ✓ |
| **Typo-Att.** [103] | Preforming character-level perturbations on a QWERTY keyboard | ✓ |
| **VIPER** [40] | Changing some letters to symbols in harmful words to avoid detection | ✓ |
| **Checklist** [110] | Performing Word-level perturbations using a predifined word checklist | ✓ |
| **BERT-Att.** [72] | Using BERT to generate adversarial samples against other LLMs | ✓ |
| **GBDA** [55] | Gradient-based white box attack using words flipping to mislead text classifiers | ✓ |
| **TF-Att.** [77] | Generating adversarial examples with critical units of sentences using LLMs | ✗ |

### *4.4. Model Extraction*

Model extraction attacks are the fifth type we illustrate in this section. These attacks aim to recreate or steal a language model's functionality by querying it and using the responses to reconstruct the model, this poses a significant threat as they allow adversaries to duplicate proprietary models without access to their internal details. We present below in Table 5, eight examples of Model Extraction attacks, showcasing different methods adversaries use to probe black-box LLMs and either extract training data of the model, or precise personal information of users.

 doi:10.20944/preprints202412.2419.v1



**Table 5.** Examples of model extraction attacks against LLMs

| Attack | Concept | OS? |
|---|---|---|
| **User Extr.** [12] | Extracting personal data of users memorised by LLMs using model queries | ✗ |
| **LLM Tricks** [156] | Tricks to enhance data extraction capabilities on LLMs | ✓ |
| **PII Leakage** [89] | Extraction/Inference attacks for analysing personally identifiable information (PII) | ✓ |
| **ETHICIST** [162] | Data extraction with Loss Smoothed Soft Prompting | ✓ |
| **Scalable Extr.** [95] | Extracting training data from Production LLMs | ✗ |
| **Output2Prompt** [158] | Extracting user prompts by knowing only their outputs | ✓ |
| **PII Compass** [94] | Extracting phone numbers from LLM using black-box queries | ✗ |
| **Alpaca-Vicuna** [64] | Using an LLM to perform data extraction on another LLM | ✗ |

### *4.5. Model Inference*

Model inference (or Membership Inference) are the sixth type of attacks we focus on in this study. These attacks determine whether specific data samples, especially sensitive information, were part of the training set of an LLM. These attacks can compromise the privacy of users or organizations by revealing training data patterns. We gathered in Table 6 eight examples of model inference attacks, which demonstrate how attackers exploit LLMs to infer confidential training data and gain insights into the model's behavior.

**Table 6.** Examples of inference attacks against LLMs

| Attack | Concept | OS? |
|---|---|---|
| **LIRA** [11] | Combining difficulty scores and well-Calibrated Gaussian Likelihood Estimate | ✓ |
| **Ngb. Comp.** [91] | Detecting training data using neighbor text comparison | ✗ |
| **PII Leakage** [89] | Extraction/Inference attacks for analysing PII leakage | ✓ |
| **Data Detect.** [123] | Detecting pretraining samples of an LLM using minimal probabilities | ✓ |
| **ProPILE** [65] | Probing framework to assess the likelihood of a PII in the training set | ✗ |
| **MIA-LLM** [45] | Membership Inference based on Self-calibrated Probabilistic Variation | ✓ |
| **DeCop** [37] | Detecting copyrighted content in training sets using multiple-choice questions | ✓ |
| **ConRecall** [134] | Using Contrastive Decoding to detect LLM's pre-training data | ✓ |

### *4.6. Poisoning/Trojan/Backdoors*

The last attacks on LLM we show are Poisoning, Trojan, and Backdoor attacks, which involve injecting malicious data or hidden triggers during the training phase of an LLM. This can lead to incorrect or dangerous behavior at deployment, allowing attackers to manipulate the model's responses. We have compiled in Table 7 eight examples of these attacks, where adversaries either corrupt the training process with poisoned data, or plant triggers to exploit models during inference, demonstrating the serious risks these methods pose to LLMs.

 doi:10.20944/preprints202412.2419.v1



**Table 7.** Examples of poisoning, trojan, and backdoor attacks against LLMs

| Attack | Concept | OS? |
|---|---|---|
| **TrojLLM** [151] | Inserting Trojans into text prompts in black-box LLM APIs | ✓ |
| **Bst-of-Vnm.** [7] | Attacking RLHF by injecting Poisoned Preference Data | ✗ |
| **CodeBreaker** [153] | Instering Backdoors on code-completion LLMs to sugget vulnerable code | ✓ |
| **Rtv. Poison.** [160] | Misleading LLMs during the RAG process with malicious documents | ✗ |
| **Clinical LLM** [26] | Editing LLMs to reveal serious implications in clinical settings | ✗ |
| **BackdoorLLM** [75] | Comprehensive benchmark for studying backdoor attacks on LLMs | ✓ |
| **CBA** [58] | Composite Backdoor Attacks against LLMs | ✓ |
| **TA²** [136] | Injecting trojan steering vectors into the activation layers of LLMs | ✓ |

## 5. Classification of Adversarial attacks on LLMs based on their danger level

After presenting the existing classifications of AAs and some of the most-recent attacks against LLMs, we propose in this section a new criterion for classifying AAs on LLMs. We present the idea and methodology in the following subsections.

### *5.1. Principle*

Seeing the list of AAs on LLMs presented in Section 4 and how frequent they are, one question that comes across the mind is what attacks should be **mitigated first** to secure LLMs? In order to answer this question, we need to rank the available attacks based on their **danger level** against LLMs in order to know what attacks is a model most-vulnerable to. This can be done by calculating the vulnerability score those of attacks using Vulnerability-assessment metrics [117].

### *5.2. Vulnerability-assessment Metrics*

Vulnerability assessment metrics are critical tools for evaluating and ranking potential security threats based on their severity and likelihood of exploitation. Various methodologies, such as DREAD [92], CVSS [114], OWASP Risk Rating [140], and SSVC [128], provide frameworks for assessing vulnerabilities by considering different factors, including technical attributes, potential impacts, and contextual elements. By systematically analyzing attacks based on their danger, these assessment tools facilitate informed decision-making in an ever-evolving threat landscape, allowing organizations to strengthen their security posture and better protect their assets.

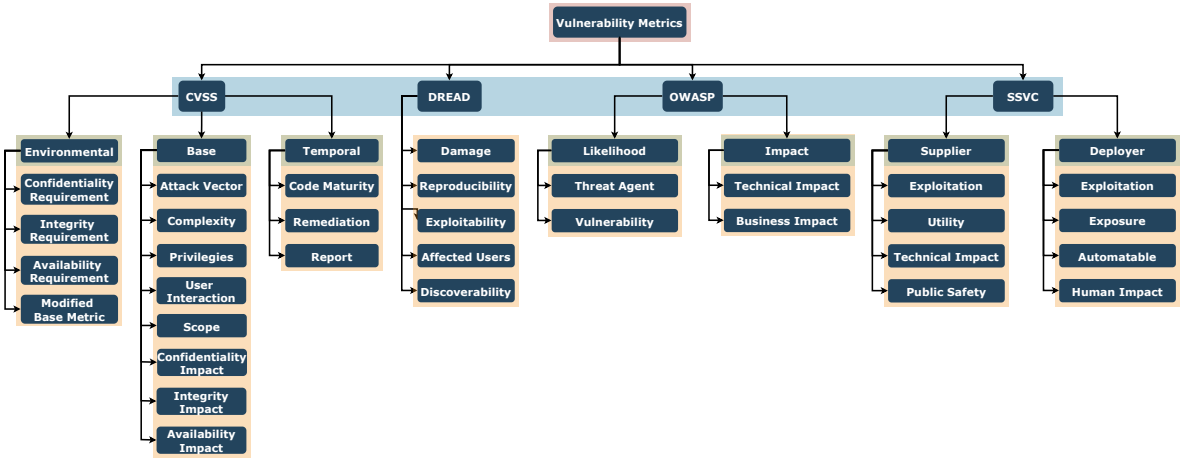



**Figure 3.** Examples of known vulnerability assessment metrics

 doi:10.20944/preprints202412.2419.v1



5.2.1. DREAD [92]

Originally developed by Microsoft, DREAD is a qualitative risk assessment model that ranks, prioritizes, and evaluates the severity of vulnerabilities and potential threats based on five factors: Damage potential (D), Reproducibility (R), Exploitability (E), Affected users (A), and Discoverability (D) of the attack.

Calculations

The vulnerability score is calculated with DREAD as an average score of the five factors, each assessed with a value out of 10. The details of each factor and their values are shown in Table 8 below.

**Table 8.** Metric factors of DREAD [92]

| Factor | Definition | Values |
|---|---|---|
| **Damage Potential (D)** | How much damage can be caused | [1 (Low), 10 (High)] |
| **Reproducibility (R)** | How easy is it to reproduce the attack | |
| **Exploitability (E)** | How easy is it to exploit the vulnerability | |
| **Affected Users (A)** | How many users would be affected | |
| **Discoverability (D)** | How easy is it to discover the vulnerability | |

A value in the range $[1,4[$ is labeled as **'Low'** in the level of criticality, a value in the range $[4,7[$ labeled as **'Medium'** in criticality, and values over 7 are labeled as **'High'** in criticality. The final score is calculated following this equation:

$$Score = (D + R + E + A + D)/5 \quad (1)$$

Limitations

The DREAD model, previously popular for qualitative risk assessment, has several limitations that have reduced its use in favor of more structured frameworks. First of all, its five categories (Damage, Reproducibility, Exploitability, Affected Users, and Discoverability) are highly subjective, leading to inconsistent scoring and prioritization across different assessors and organizations. Moreover, DREAD overlooks contextual factors like the specific environment and business impact, limiting its adaptability for complex needs. Finally, it also fails to account for dynamic threats or mitigation measures, making it less effective for ongoing risk management.

5.2.2. CVSS (Common Vulnerability Scoring System) [114]

Created by the FIRST[2] (Forum of Incident Response and Security Teams), the CVSS is an industry-standard scoring system for rating the severity of software vulnerabilities out of 10. It is encompasses three main metrics:

- **Base Metrics:** Represent the vulnerabilities that are constant over time. It contains factor related to the exploitability of an attack (how easy it is to exploit the vulnerability) like the Attack Vector (AV), Attack Complexity (AC), Privileges Required (PR), User Interaction (UI), and the Scope (S) of the attack. And factors related to the impact of an attack on the CIA triad, such as Confidentiality Impact (C), Integrity Impact (I), and Availability Impact (A).

[2] https://www.first.org/

 doi:10.20944/preprints202412.2419.v1



- **Temporal Metrics (Optional):** Represent the vulnerabilities that might change over time in order to update the base score, it encompasses three factors, Exploit Code Maturity (E), Remediation Level (RL), and Report Confidence (RC).
- **Environmental Metrics (Optional):** Vulnerabilities that are unique to a user environment, such as the Confidentiality Requirements (CR), Integrity Requirement (IR), Availability Requirement (AR), and the modified Base Metrics.

Calculations

The values in CVSS factors are not explicitly numerical; but selected from a specific range of choices, with each qualitative value having a corresponding coefficient. The details of each factor and of the Base Metric and their values according to CVSS version 3.1[3] are presented below in Table 9, and their equivalent decimal values are detailed in Table 10.

**Table 9.** Base metric factors of CVSS 3.1 [44]

| Factor | Definition | Values |
|---|---|---|
| **Attack Vector (AV)** | From where the exploitation is possible | Network (N), Adjacent (A), Local (L), Physical (P) |
| **Attack Complexity (AC)** | How complex is the exploitation | Low (L), High (H) |
| **Privileges Required (PR)** | How much privileges are needed for the exploit | None (N), Low (L), High (H) |
| **User Interaction (UI)** | Is a user interaction required in the compromise | None (N), Required (R) |
| **Scope (S)** | Does the scope of the attack change | Unchanged (U), Changed (C) |
| **Confidentiality Impact (C)** | How much impacted is the confidentiality | None (N), Low (L), High (H) |
| **Integrity Impact (I)** | How much impacted is the integrity | |
| **Availability Impact (A)** | How much impacted is the availability | |

**Table 10.** Numerical values of each CVSS 3.1 factor [44]

| Factor | Value | Decimal Value |
|---|---|---|
| **AV** | Network (N) | 0.85 |
| | Adjacent (A) | 0.62 |
| | Local (L) | 0.55 |
| | Physical (P) | 0.22 |
| **AC** | Low (L) | 0.77 |
| | High (H) | 0.44 |
| **PR** | None (N) | 0.85 |
| | Low (L) | 0.62 (if S = U), 0.68 (if S = C) |
| | High (H) | 0.27 (if S = U), 0.50 (if S = C) |
| **UI** | None (N) | 0.85 |
| | Required (R) | 0.62 |
| **C, I ,A** | None (N) | 0.00 |
| | Low (L) | 0.22 |
| | High (H) | 0.56 |

[3] Although version 4.0 is the most recent, version 3.1 is still the most used in vulnerability assessment.

 doi:10.20944/preprints202412.2419.v1



After assessing a value for each metric, the Base Score of the CVSS is calculated using two different equations depending on the Scope (S), which is either Changed (C) or Unchanged (U). Below are the full details of the equations for both cases:

**If S = U:**

$$\text{Base Score} = \text{roundup}\big(\min\big(\text{Impact}_U + \text{Exploitability}, 10\big) \times 1.08\big) \tag{2}$$

$$\text{Impact}_U = 6.42 \times (1 - (1 - C) \times (1 - I) \times (1 - A)) \tag{3}$$

$$\text{Exploitability} = 8.22 \times \text{AV} \times \text{AC} \times \text{PR} \times \text{UI} \tag{4}$$

**If S = C:**

$$\text{Base Score} = \text{roundup}\big(\min\big(1.08 \times (\text{Impact}_C + \text{Exploitability}), 10\big)\big) \tag{5}$$

$$\text{Impact}_C = 7.52 \times (I - 0.029) - 3.25 \times (I - 0.02)^{15} \tag{6}$$

$$I = 1 - (1 - C) \times (1 - I) \times (1 - A) \tag{7}$$

The exploitability remains the same.

The final Base Score ranges from 0 to 10, with the same criticality assignment as in DREAD, adding to it that a base score of 9 or more is considered a **'Critical'** vulnerability.

Limitations

CVSS is a widely used standard for scoring vulnerabilities but has several limitations that affect its real-world effectiveness. Firstly, it tends to oversimplify calculations by focusing on technical aspects like attack complexity and impacts on confidentiality, integrity, and availability, while neglecting business impact and regulatory considerations. Additionally, the Temporal score of CVSS, intended to reflect changing conditions, relies on manual updates rather than real-time adjustments, making it less responsive to evolving threats. Finally, CVSS can be inconsistent, as different organizations may interpret scoring criteria differently, leading to varying assessments for the same vulnerability.

5.2.3. OWASP Risk Rating [140]

Developed by the Open Web Application Security Project (OWASP)[4], it is a risk assessment methodology that evaluates vulnerabilities and categorizes security risks in web applications by assessing likelihood (based on threat agent and vulnerability characteristics) and impact (considering technical and business factors) to produce an overall risk score.

- **Likelihood:** Calculates the probability of the attack to be exploited based on two components:
  - **Threat Agent (TA):** Quantifies the skill level, motivation, opportunity, and size of the threat-agent population
  - **Vulnerability (V):** Quantifies the ease of discovery, ease of exploit, awareness, awareness of the system administrators, and the intrusion detection level.
- **Impact:** Calculates the impact or loss produced by the attacks, it encompasses two types of impact:
  - **Technical Impact (TI):** Quantifies the impact on Confidentiality, Integrity, and Availability.
  - **Business Impact (BI):** Quantifies the financial damage, reputation damage, non-compliance, and privacy violation

Calculations

The vulnerability score is calculated based on the average score of each component. The values and definitions of each factor of OWASP Risk Rating is presented in Table 11.

[4] https://owasp.org/

 doi:10.20944/preprints202412.2419.v1



**Table 11.** Metric factors of OWASP Risk Rating [140]

| Factor | Definition | Values |
|---|---|---|
| **Skill Level (SL)** | How much expertise is needed | [1 (Low), 10 (High)] |
| **Motivation (M)** | How much motivated is the attacker | |
| **Opportunity (O)** | How easy is it to exploit the vulnerability | |
| **Size of TA (S)** | How many attackers can be there | |
| **Ease of Discovery (ED)** | How easy is it to discover the vulnerability | |
| **Ease of Exploit (EE)** | How easy is it to exploit the vulnerability | |
| **Awareness (A)** | How much aware are the defenders | |
| **Intrusion Detect. (ID)** | How difficult is it to detect the attack | |
| **Confidentiality (LC)** | How much is confidentiality impacted | |
| **Integrity (LI)** | How much is the integrity impacted | |
| **Availability (LAV)** | How much is the availability impacted | |
| **Financial Dmg. (FD)** | How much financial loss can result | |
| **Reputation Dmg. (RD)** | How much the reputation can be harmed | |
| **Non-Compliance (NC)** | How much legal violations can happen | |
| **Privacy Violation (PV)** | How much users' privacy is violated | |

In this metric, a value in the range $[1, 3[$ is considered an attack of **'Low'** criticality. The **'Medium'** criticality range is $[3, 6[$, and the values starting from 6 are labeled as **'High'** in criticality. After assessing all the values, the final score is a multiplication between the score of Likelihood and the score of Impact as shown below:

$$OWASPScore = Likelihood * Impact \tag{8}$$

The score of Likelihood is calculated as the mean of the Threat Agent and the Vulnerability scores:

$$Likelihood = (Score_{TA} + Score_{V})/2 \tag{9}$$

And the score of Impact is calculated as the mean of the Technical and Business impact scores

$$Impact = (Score_{TI} + Score_{BI})/2 \tag{10}$$

Where the score of each component (TA, V, TI, BI) are respectively the average score of their factors:

$$Score_{TA} = (SkillLevel + Motivation + Opportunity + Size_{TA})/4 \tag{11}$$

$$Score_{V} = (Easeof Discovery + Easeof Exploit + Awareness + IntrusionDetection)/4 \tag{12}$$

$$Score_{TI} = (Confidentiality + Integrity + Availability)/3 \tag{13}$$

$$Score_{BI} = (FinancialDmg + ReputationDmg + NonCompliance + PrivacyViolation)/4 \tag{14}$$

 doi:10.20944/preprints202412.2419.v1



The rank of the final OWASP severity-score (Low, Medium, High, Critical) is defined based on the combinations shown in the Table 12. For example, if the $Likelihood = 5/10$ (Medium criticality) and the $Impact = 6/10$ (High criticality), the final score according the matrix is **High**.

**Table 12.** Criticality Matrix of OWASP Risk Rating [140]

| | | | | |
|---|---|---|---|---|
| **Impact** | **High** | **Medium** | **High** | **Critical** |
| | **Medium** | **Low** | **Medium** | **High** |
| | **Low** | **Note** | **Low** | **Medium** |
| | | **Low** | **Medium** | **High** |
| | **Likelihood** | | | |

Limitations

The OWASP Risk Rating methodology, though widely used for web application security assessment, has notable limitations. Its reliance on subjective evaluations of factors like threat agent skill and impact severity can result in inconsistent ratings across different assessors and lead to biased outcomes. Additionally, OWASP Risk Rating lacks specificity for environments like cloud or mobile and does not adapt to rapidly changing threat landscapes, making it less responsive in dynamic security contexts. Finally, having many factors increases the complexity of this metric and its reliance on experts knowledge to assess each factor precisely.

5.2.4. SSVC (Stakeholder-Specific Vulnerability Categorization) [128]

The SSVC is a framework that prioritizes vulnerabilities based on qualitative decision trees tailored to specific stakeholder roles, instead of numerical severity scores. The main two stakeholders represented are:

- **Suppliers:** They decide how urgent it is to develop and release patches for their systems based on reports about potential vulnerabilities. Their decision tree is based on factors such as Exploitation, Technical Impact, Utility, and Safety Impact.
- **Deployers:** They decide when and how to deploy the patches developed by the suppliers. Their decision tree is based on similar factors such as Exploitation, System Exposure, Automation, and Human Impact.

Calculations

In our case, we consider LLMselves as Suppliers trying to assess the potential vulnerabilities impacting their LLM. Table 13 below show the different factors used in evaluating vulnerabilities using SSVC as a supplier.

 doi:10.20944/preprints202412.2419.v1



**Table 13.** Metric factors of SSVC for a supplier [128]

| Factor | Definition | Values |
|---|---|---|
| **Exploitation (E)** | In which state is the exploitation | None (N), Proof-of-Concept (P), Active (A) |
| **Automatable (A)** | Can the attack be automatable | No (N), Yes (Y) |
| **Value Density (V)** | How valuable is the information accessed by the attacker | Diffuse (D), Concentrated (C) |
| **Utility (U)** | How much useful is the exploit for the attacker | Laborious (L), Efficient (E), Super Efficient (S) |
| **Technical Impact (T)** | How much impact does the vulnerability do | Partial (P), Total (T) |
| **Public-Safety Impact (S)** | How much impact has vulnerability on the public | Minimal (M), Significant (S) |

The value of Utility (U) is calculated based in the values if Automatable (A) and Value Density (V) as follows:

The final decision is taken by following the logic described in Figure 4. There four main possible outcomes ranked from the lowest priority to the highest one are: Defer, Scheduled, Out-of-cycle, and Immediate. Each one of them represents the emergency level for developing corresponding patches.

 doi:10.20944/preprints202412.2419.v1



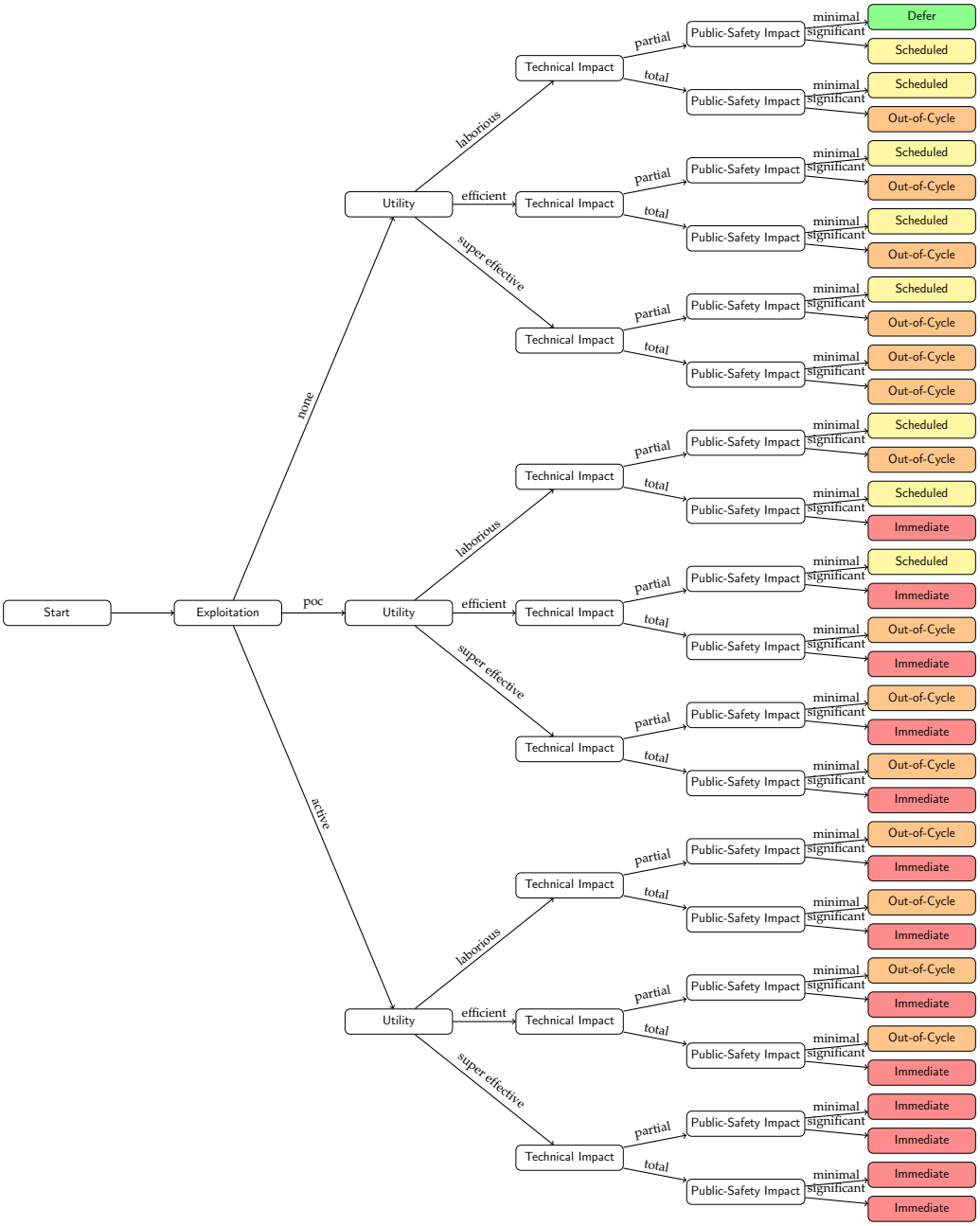



**Figure 4.** Decision Tree for Suppliers in SSVC [128]

Limitations

The SSVC metric has several limitations. It relies heavily on qualitative decision points, which may lead to subjective interpretations and inconsistencies across stakeholders. Additionally, the absence of numerical scoring might limit its integration with existing risk management systems that rely on quantitative data, potentially requiring significant adjustments to current workflows. Lastly, SSVC is tailored for specific stakeholder roles, which may make it be less effective in hybrid roles or complex environments where stakeholders overlap.

## 6. Assessment of AAs on LLMs with Vulnerability Metrics

In this section, we present and interpret the results of assessing the criticality of AAs against LLMs, grouped in seven types: White-box Jailbreak, Black-box Jailbreak, Prompt Injection, Evasion attacks, Model Extraction, Model Inference, and Poisoning/Trojan/Backdoor attacks. The detailed scores of these attacks given by the 3 LLMs (GPT-4o, LLAMA, and Perplexity) and their average are

 doi:10.20944/preprints202412.2419.v1



presented in Appendix A. We represent the results in score-vectors and in spider-graph formats for more interpretability.

Note that for the qualitative factors of CVSS and SSVC, we represent their values **numerically** in the spider graph following this logic:

- For CVSS Factors:
  - If they have four values (eg. AV), they are represented with values from 1 to 4.
  - If have three values (eg. PR, C, I, A), they are represented with values from 1 to 3.
  - If they have two values (eg. AC, UI, S), they are represented with the values 2 and 4.
- For SSVC Factors:
  - If they have three values (eg. E, U), they are represented with values from 1 to 3.
  - If they have two values (eg. A, V, T, P), they are represented with the values 1 and 3.

### *6.1. Assessment of White-box Jailbreak attacks*

We start by evaluating White-box jailbreak attacks, the chosen attacks are the same presented in Section 4.1.1 earlier: (1) GCG [166], (2) Visual Modality [96], (3) PGD [48], (4) SCAV [149], (5) Soft Prompt Threats [115], (6) DrAttack [74], (7) RADIAL [36], (8) ReNeLLM [31].

#### 6.1.1. With DREAD

We start by evaluating the eight White-box jailbreaks attacks using DREAD [92]. Here are below the attack vectors of each attack:

- (1) → (D:8/R:9/E:8/A:8/D:6) = 7.8 (High)
- (2) → (D:6/R:6/E:6/A:6/D:5) = 5.8 (Medium)
- (3) → (D:7/R:7/E:7/A:7/D:5) = 6.6 (Medium)
- (4) → (D:7/R:6/E:5/A:6/D:5) = 5.8 (Medium)
- (5) → (D:8/R:9/E:7/A:7/D:6) = 7.4 (High)
- (6) → (D:8/R:8/E:7/A:8/D:6) = 7.4 (High)
- (7) → (D:7/R:6/E:7/A:6/D:5) = 6.2 (Medium)
- (8) → (D:8/R:9/E:8/A:7/D:6) = 7.6 (High)

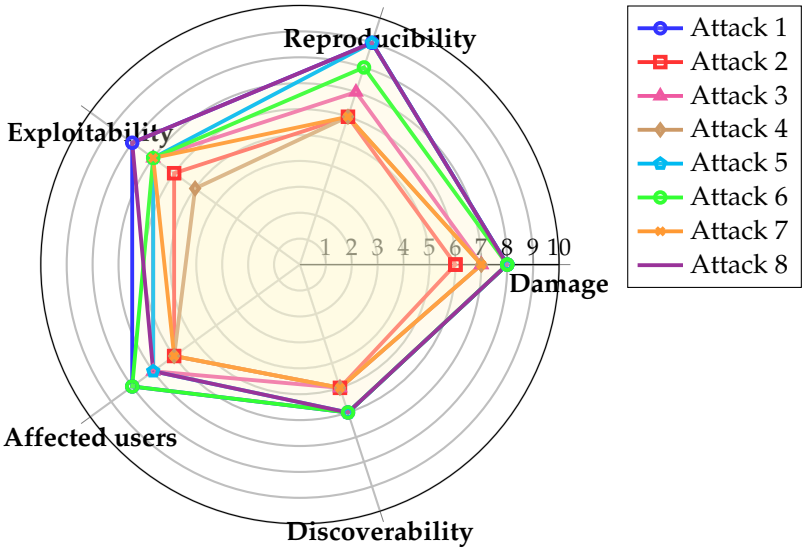
The detailed calculations for these attacks are presented in Table A1, with assessments supervised by a **Human-in-the-Loop (HitL)** to minimize **misconceptions**. For instance, GPT-4o initially scored 8/10 for the Discoverability factor in DREAD for the first attack [166], while LLAMA-3 and Perplexity AI both assigned a score of 6/10. GPT-4o's higher score stemmed from a **misunderstanding** of the factor's meaning, interpreting Discoverability as the level of researcher awareness about the threat rather than the ease with which it can be discovered. After **clarifying** this distinction, GPT-4o revised its score to 5/10, aligning more closely with the intended definition of the metric.

A different issue arose when evaluating the **impact** of attacks. For example, the Damage factor of the second attack [96] was rated 6/10 by GPT-4o and 5/10 by LLAMA-3, reflecting moderate damage due to situational input requirements, such as specific visual input use cases. However, Perplexity AI assigned a higher score of 8/10, citing potential scenarios where the attack could have a significant impact on the targeted system. In this case, the discrepancy was due to differing **interpretations** rather than misunderstandings, making it difficult to standardize the scores. To address this, **averaging** the three scores provided a balanced result, aligning closely with the consensus of GPT-4o and LLAMA-3.

Using this approach, we reduced inconsistencies in the scoring process. The final DREAD scores are illustrated in a spider graph, highlighting that White-box Jailbreak attacks can inflict considerable damage on systems while being relatively easy to reproduce. However, discovering these threats remains a significant challenge.

 doi:10.20944/preprints202412.2419.v1






6.1.2. With CVSS

Then, we evaluate the assessment of these attacks using CVSS [114]. The corresponding CVSS Vectors are shown below:

- (1) → (AV:N/AC:H/PR:N/UI:N/S:C/C:L/I:H/A:N) = 7.5 (High)
- (2) → (AV:N/AC:H/PR:N/UI:N/S:C/C:L/I:H/A:N) = 7.5 (High)
- (3) → (AV:N/AC:H/PR:N/UI:N/S:C/C:L/I:H/A:N) = 7.5 (High)
- (4) → (AV:N/AC:H/PR:N/UI:N/S:C/C:L/I:H/A:N) = 7.1 (High)
- (5) → (AV:N/AC:L/PR:L/UI:N/S:C/C:L/I:H/A:N) = 8.5 (High)
- (6) → (AV:N/AC:L/PR:L/UI:N/S:C/C:L/I:H/A:N) = 8.5 (High)
- (7) → (AV:N/AC:H/PR:N/UI:R/S:C/C:L/I:H/A:N) = 6.9 (Medium)
- (8) → (AV:N/AC:L/PR:N/UI:R/S:C/C:L/I:H/A:N) = 8.2 (High)

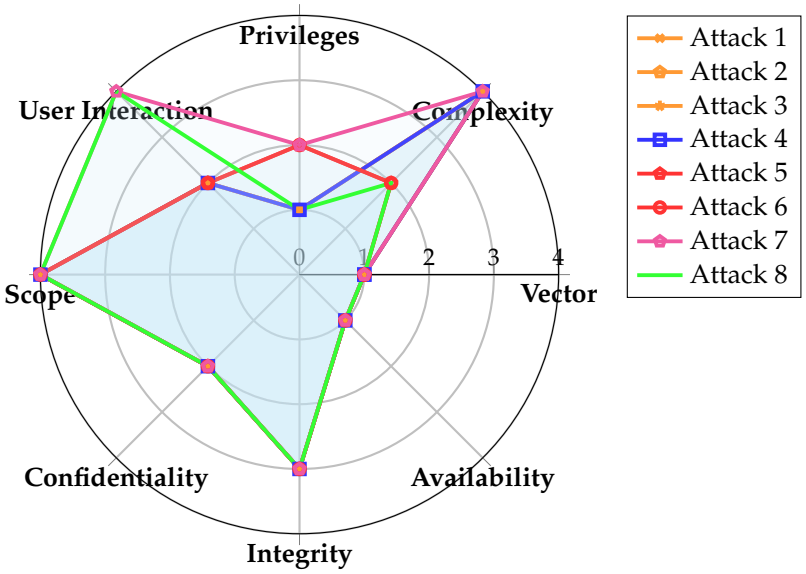
The detailed scores are presented in Table A2. During the analysis, some LLMs encountered challenges in correctly interpreting the characteristics of each attack. For instance, GPT-4o initially concluded that White-box jailbreak attacks only impact the Confidentiality of data, with no effect on Integrity—a conclusion that was refuted by the other two LLMs. It was crucial to identify such **misunderstandings** and **guide** the models to recognize their errors. Rather than providing direct corrections, we prompted GPT-4o with **questions** such as: *Do these attacks target Integrity given that they involve manipulation of gradients and embeddings?* This approach enabled the model to identify and rectify its own mistake while fostering greater **caution** in subsequent assessments.

This process highlights another key advantage of using multiple LLMs: **they provide diverse perspectives and explanations**, which help identify and address unconventional or erroneous analyses. Moreover, there was also other slight divergence in scoring factors such as the Scope and User Interaction; but using an averaging method helps align the final scores to the majority.

After averaging the final values, we visualized the scores using a spider chart for clarity. The CVSS scores reveal that White-box attacks are typically executed through the network, requiring low-to-medium privileges and primarily targeting the Integrity of systems.

 doi:10.20944/preprints202412.2419.v1






6.1.3. With OWASP Risk Rating

A third evaluation of white-box jailbreak attacks is done using OWASP RR [140]. The corresponding vulnerability vectors of each attack is:

- (1) → (SL:7/M:6/O:6/S:6/ED:7/EE:8/A:5/ID:5/LC:5/LI:7/LA:4/FD:7/RD:8/NC:4/PV:4) = 3.6 (High)
- (2) → (SL:7/M:6/O:5/S:5/ED:5/EE:7/A:5/ID:5/LC:5/LI:6/LA:3/FD:6/RD:7/NC:4/PV:4) = 2.8 (Medium)
- (3) → (SL:7/M:7/O:5/S:6/ED:6/EE:7/A:5/ID:5/LC:5/LI:7/LA:3/FD:6/RD:7/NC:6/PV:5) = 3.2 (Medium)
- (4) → (SL:5/M:6/O:4/S:3/ED:4/EE:6/A:5/ID:4/LC:4/LI:6/LA:1/FD:5/RD:6/NC:3/PV:4) = 1.9 (Medium)
- (5) → (SL:6/M:7/O:6/S:5/ED:6/EE:7/A:6/ID:5/LC:5/LI:7/LA:1/FD:6/RD:7/NC:4/PV:4) = 2.8 (High)
- (6) → (SL:7/M:7/O:6/S:6/ED:6/EE:7/A:5/ID:5/LC:5/LI:7/LA:2/FD:6/RD:7/NC:4/PV:5) = 3.2 (High)
- (7) → (SL:6/M:6/O:5/S:4/ED:5/EE:6/A:5/ID:5/LC:4/LI:6/LA:1/FD:5/RD:6/NC:3/PV:3) = 2.1 (Medium)
- (8) → (SL:7/M:7/O:7/S:5/ED:7/EE:7/A:6/ID:5/LC:4/LI:7/LA:1/FD:6/RD:7/NC:4/PV:4) = 3.1 (High)

The scores assigned by each LLM are detailed in Table A3. With its multiple factors, the OWASP Risk Rating provided a more comprehensive analysis of each attack. However, we encountered some **interpretation discrepancies**, particularly with Perplexity AI. This model argued that these attacks have a Medium-to-High impact on Confidentiality—an assessment that differed from its CVSS evaluation of the same attacks. This highlights the inherent **subjectivity** in scoring, as analyzing identical attacks in separate conversations can yield inconsistent results. In contrast, the other two LLMs provided scores consistent with the CVSS evaluation for Confidentiality and Integrity, along with a Low-to-None impact on Availability. Averaging the scores **mitigated** such discrepancies while preserving the unique perspectives offered by each LLM, especially in factors like Non-Compliance and Privacy Violation. Notably, LLAMA-3 failed to detect any impact in these areas, whereas GPT-4o and Perplexity AI highlighted their significance.

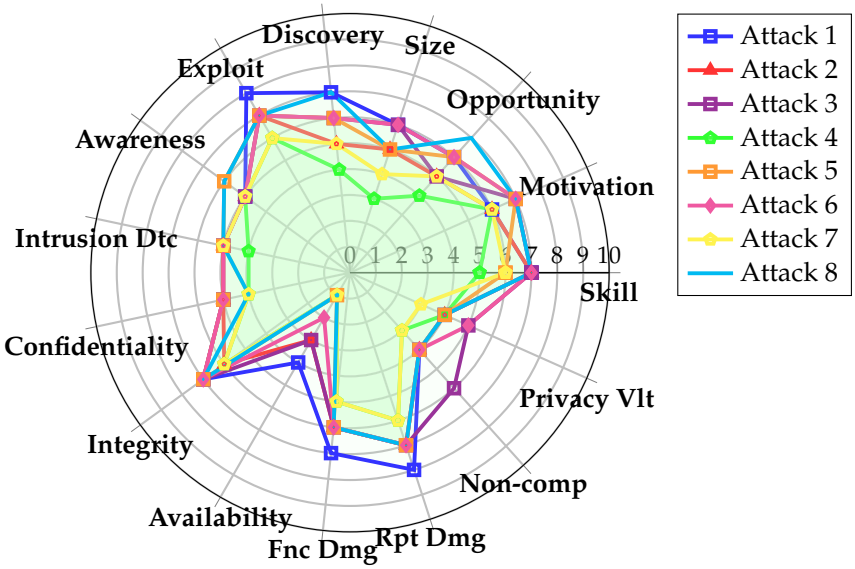
Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 December 2024 doi:10.20944/preprints202412.2419.v1



Another challenge we observed was the tendency of LLMs to rely on **memorized values** when evaluating attacks across multiple factors. For example, GPT-4o initially assigned identical scores to the first two attacks [96,166]. Upon prompting it to provide objective and distinct evaluations, GPT-4o revised its scores, adjusting the ED value from 6/10 to 5/10, the Availability impact from 6/10 to 5/10, and the NC value from 6/10 to 7/10. It justified these changes by acknowledging **similarities** between the attacks while ensuring the scores reflected nuanced differences.

The averaged scores are visualized below for clarity. These results align with the CVSS evaluation in terms of technical impact and ease of exploitation, while also shedding light on the reputational damage that could arise if such attacks are exploited. Moreover, they emphasize that White-box attacks have a medium impact on Non-Compliance and Privacy Violation.

**Assessment of White-box model jailbreaking attacks with OWASP RR**




6.1.4. With SSVC

Finally, we evaluate these attacks using SSVC [128]. The corresponding vectors, as a supplier, are shown below:

- (1) → (E:P/A:Y/V:C/U:S/T:T/P:S) = Immediate (Very High)
- (2) → (E:P/A:N/V:C/U:E/T:T/P:S) = Immediate (Very High)
- (3) → (E:P/A:N/V:C/U:E/T:T/P:S) = Immediate (Very High)
- (4) → (E:P/A:N/V:C/U:E/T:P/P:M) = Scheduled (Medium)
- (5) → (E:P/A:Y/V:C/U:S/T:T/P:S) = Immediate (Very High)
- (6) → (E:P/A:Y/V:C/U:S/T:T/P:S) = Immediate (Very High)
- (7) → (E:P/A:N/V:C/U:E/T:P/P:M) = Scheduled (Medium)
- (8) → (E:P/A:Y/V:C/U:S/T:T/P:S) = Immediate (Very High)

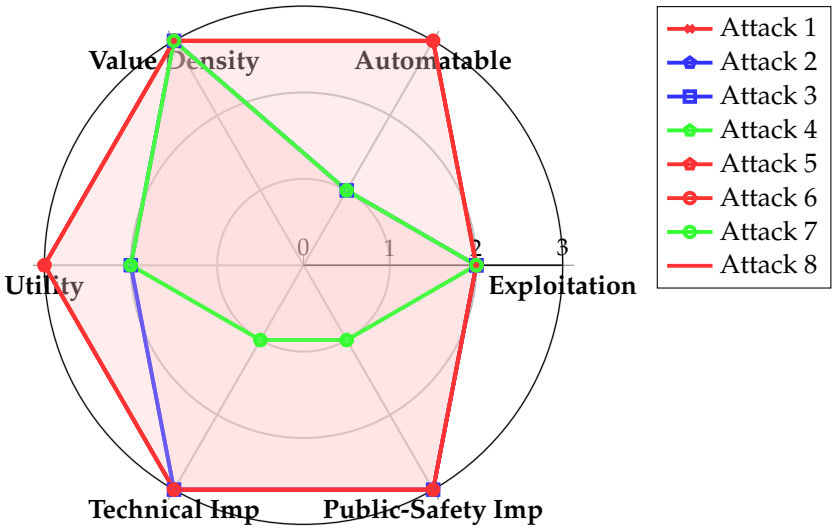
Table A4 presents the detailed SSVC assessment scores provided by each LLM. As SSVC is relatively **straightforward** to apply, the LLMs performed the evaluations without significant issues. The primary role of the HitL in this context was to **interpret** the rationale behind the values assigned by the LLMs, particularly for the Exploitation factor. For instance, when evaluating the second White-box jailbreak attack [96], GPT-4o determined there was no PoC for the attack, as its implementation was not publicly available, and accordingly assigned it a "None" value. In contrast, LLAMA-3 and Perplexity AI offered a different perspective. Both argued that the paper provided sufficient detail about the attack, making it possible to reproduce with some effort. Consequently, they concluded that a PoC exists. With the majority of models agreeing, the average score reflected their viewpoint, recognizing the presence of a PoC.

 doi:10.20944/preprints202412.2419.v1



The final SSVC scores are visualized below in a spider chart. These results indicate that White-box jailbreak attacks can be automated and highly rewarding, underscoring their significant risks to both technical systems and public safety.

**Assessment of White-box model jailbreaking attacks with SSVC**




*6.2. Assessment of Black-box Jailbreak attacks*

We evaluate now Black-box jailbreak attacks, the eight attacks are the same presented in Section 4.1.2 earlier: (1) Privacy attack on GPT-4o [71], (2) PAIR [19], (3) DAN [120], (4) Simple Adaptive Attack [2], (5) PAL [126], (6) GCQ [56], (7) IRIS [108], (8) Tastle [146].

6.2.1. With DREAD

We start with the evaluation using DREAD [92]. Below are the DREAD vectors of each of the eight Black-box Jailbreak attacks:

- (1) → (D:8/R:7/E:7/A:8/D:5) = 7 (High)
- (2) → (D:8/R:8/E:8/A:7/D:5) = 7.2 (High)
- (3) → (D:8/R:8/E:7/A:7/D:6) = 7.2 (High)
- (4) → (D:9/R:8/E:8/A:8/D:6) = 7.8 (High)
- (5) → (D:8/R:9/E:8/A:7/D:5) = 7.4 (High)
- (6) → (D:8/R:6/E:7/A:7/D:5) = 6.6 (Medium)
- (7) → (D:9/R:8/E:9/A:8/D:5) = 7.8 (High)
- (8) → (D:8/R:7/E:7/A:7/D:5) = 6.8 (Medium)

The details are presented in Table A5. This time, no misunderstandings occurred, as the corrections made during the DREAD assessment of White-box attacks were already in place. However, some divergences in attack analysis still arose. For instance, the Exploitability factor of the first attack [71] was rated 6/10 by GPT-4o, which noted that the attack requires specific query patterns but is still manageable to execute. In contrast, LLAMA-3 and Perplexity AI assigned a score of 8/10, arguing that the implementation details provided in the paper make the attack easily exploitable.

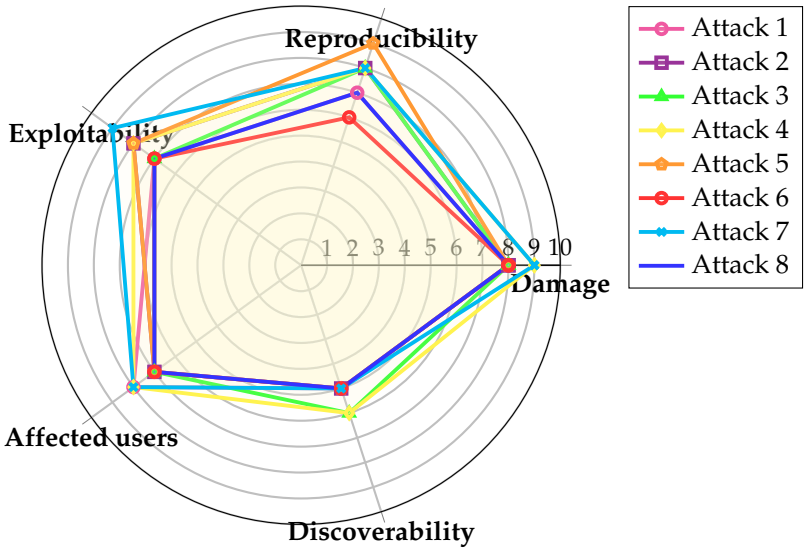
Another challenge was the potential **memorization** of values. For example, LLAMA-3 gave identical scores for the fourth, fifth, and seventh attacks [2,108,126], justifying this by highlighting the **similar characteristics** of these attacks. While this explanation is plausible, as the scores were consistent with those of the other LLMs, averaging the scores across all models helped mitigate these analytical inconsistencies by favoring the majority consensus.

 doi:10.20944/preprints202412.2419.v1



After averaging the scores, we visualized the results in a spider chart for clarity. The DREAD scores indicate that Black-box Jailbreak attacks, like their White-box counterparts, can inflict significant damage while being highly reproducible and exploitable, yet challenging to detect.

**Assessment of Black-box model jailbreaking attacks with DREAD**




6.2.2. With CVSS

In this second assessment of Black-box Jailbreak, we evaluate the attacks using CVSS [114]. The corresponding CVSS Vectors are shown below:

- (1) → (AV:N/AC:L/PR:N/UI:N/S:U/C:L/I:L/A:N) = 6.5 (Medium)
- (2) → (AV:N/AC:L/PR:N/UI:N/S:U/C:L/I:H/A:N) = 8.2 (High)
- (3) → (AV:N/AC:L/PR:N/UI:N/S:U/C:L/I:L/A:N) = 6.5 (Medium)
- (4) → (AV:N/AC:L/PR:N/UI:N/S:C/C:L/I:N/A:N) = 7.2 (High)
- (5) → (AV:N/AC:L/PR:N/UI:N/S:C/C:L/I:N/A:N) = 7.2 (High)
- (6) → (AV:N/AC:H/PR:N/UI:N/S:U/C:L/I:N/A:N) = 5.4 (Medium)
- (7) → (AV:N/AC:L/PR:L/UI:N/S:U/C:L/I:H/A:N) = 7.1 (High)
- (8) → (AV:N/AC:L/PR:N/UI:N/S:U/C:L/I:H/A:N) = 8.2 (High)

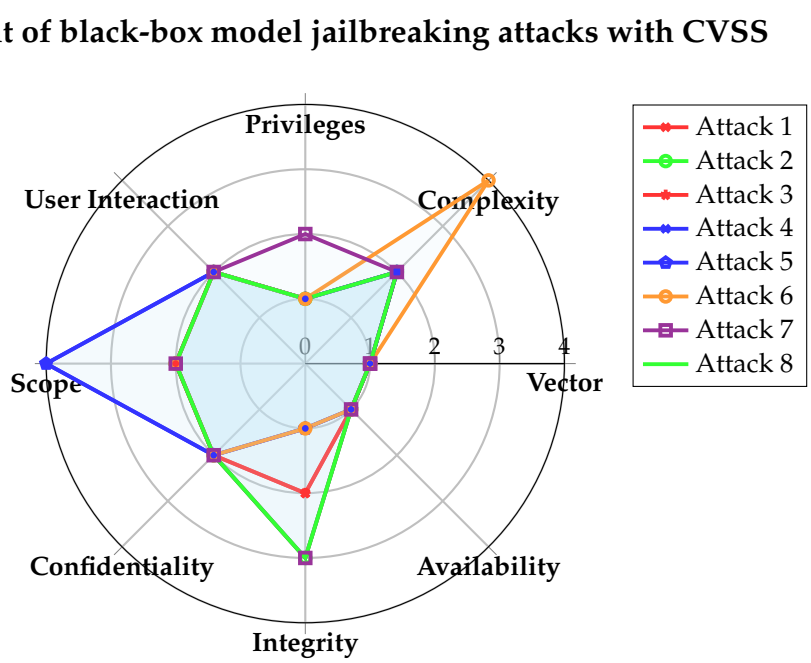
Table A6 outlines the detailed CVSS scores for the Black-box Jailbreak attacks. As observed with previous assessments, the three LLMs displayed some divergence in evaluating the technical impact of each attack. However, averaging the scores allowed us to establish a balanced consensus that moderated the variations in their evaluations.

One notable issue arose with LLAMA-3 in **interpreting** the User Interaction factor, which assesses whether a user other than the attacker must interact with the system for the attack to succeed. In the case of Black-box jailbreaks, where most attacks are executed remotely, no additional user interaction is required—a point accurately identified by GPT-4o and Perplexity AI. However, LLAMA-3 initially marked the UI factor as "Required," justifying this based on the attacker's interaction with the system. The HitL **clarified** through prompts that the UI factor refers specifically to interactions by users other than the attacker. Following this explanation, LLAMA-3 adjusted its evaluation, aligning with the "None" rating given by the other LLMs.

After averaging the scores, the final results are visualized below in a spider chart. The CVSS scores highlight that Black-box jailbreak attacks are easier to reproduce compared to White-box jailbreaks, require no privileges, and have a low-to-moderate impact on both integrity and confidentiality.

 doi:10.20944/preprints202412.2419.v1






6.2.3. With OWASP Risk Rating

A third evaluation of is done with OWASP Risk Rating [140]. The corresponding vulnerability vectors of each attack is:

- (1) → (SL:6/M:8/O:8/S:6/ED:6/EE:7/A:5/ID:6/LC:8/LI:2/LA:1/FD:6/RD:8/NC:5/PV:7) = 3.3 (High)
- (2) → (SL:5/M:8/O:8/S:6/ED:6/EE:7/A:6/ID:6/LC:8/LI:1/LA:1/FD:6/RD:8/NC:4/PV:6) = 3 (High)
- (3) → (SL:4/M:8/O:8/S:6/ED:7/EE:8/A:7/ID:7/LC:9/LI:1/LA:1/FD:7/RD:8/NC:5/PV:9) = 3.8 (High)
- (4) → (SL:7/M:8/O:7/S:5/ED:6/EE:8/A:6/ID:8/LC:8/LI:1/LA:1/FD:6/RD:8/NC:5/PV:8) = 3.5 (High)
- (5) → (SL:6/M:8/O:8/S:5/ED:6/EE:8/A:5/ID:7/LC:7/LI:1/LA:1/FD:5/RD:7/NC:5/PV:7) = 3 (High)
- (6) → (SL:6/M:8/O:8/S:6/ED:6/EE:8/A:6/ID:7/LC:7/LI:2/LA:1/FD:6/RD:7/NC:5/PV:7) = 3 (High)
- (7) → (SL:6/M:9/O:8/S:5/ED:6/EE:8/A:5/ID:7/LC:8/LI:3/LA:1/FD:6/RD:9/NC:7/PV:8) = 3.8 (High)
- (8) → (SL:6/M:8/O:7/S:5/ED:6/EE:8/A:5/ID:7/LC:8/LI:3/LA:1/FD:6/RD:8/NC:5/PV:8) = 3.4 (High)

Table A7 presents the detailed OWASP RR assessments conducted using three LLMs. Unlike previous evaluations, no significant errors were observed in the scoring provided by the models. However, some divergence was noted in specific factors. For example, when assessing the Opportunity factor for the sixth attack [56], GPT-4o and LLAMA-3 scored it 8/10 and 9/10, respectively, arguing that these attacks target online LLMs, thereby increasing the availability of opportunities for exploitation. In contrast, Perplexity AI assigned a score of 6/10, reasoning that the attacks are not immediately apparent or straightforward to execute, resulting in a medium-to-high Opportunity rating. To maintain **neutrality** and **objectivity**, we chose not to modify or influence these values, allowing the models' perspectives to remain intact. Averaging the scores enabled a **balanced** consideration of all three points of view.

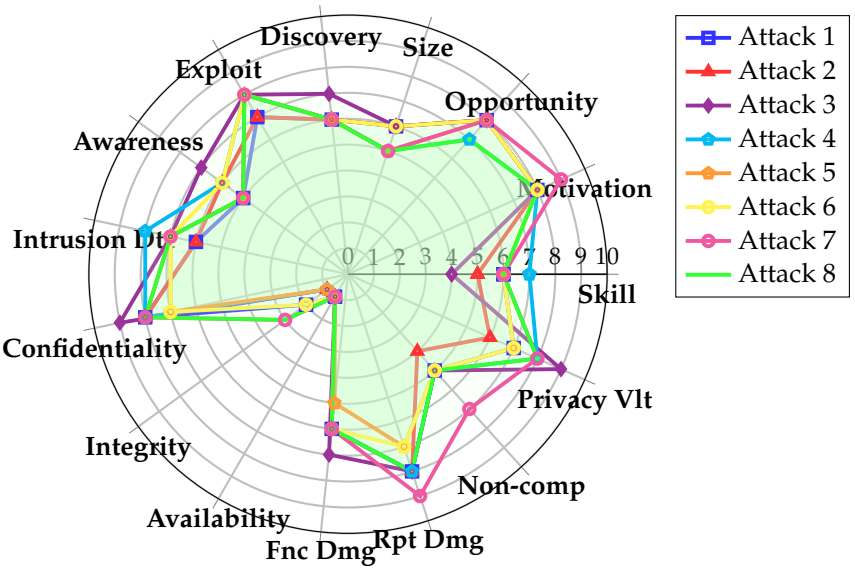
The final scores are visualized below in the spider chart. The results indicate that Black-box jailbreak attacks have a significant impact on the confidentiality of data, as they can extract sensitive

 doi:10.20944/preprints202412.2419.v1



information from the models. In contrast, their impact on integrity is minimal, and they have no impact on availability. The OWASP RR metric further highlights the severe implications these attacks have on privacy violations and the reputation of the targeted organization.

**Assessment of Black-box model jailbreaking attacks with OWASP RR**




6.2.4. With SSVC

The forth evaluation is performed using SSVC [128] in a supplier role. The corresponding vulnerability vectors are detailed below:

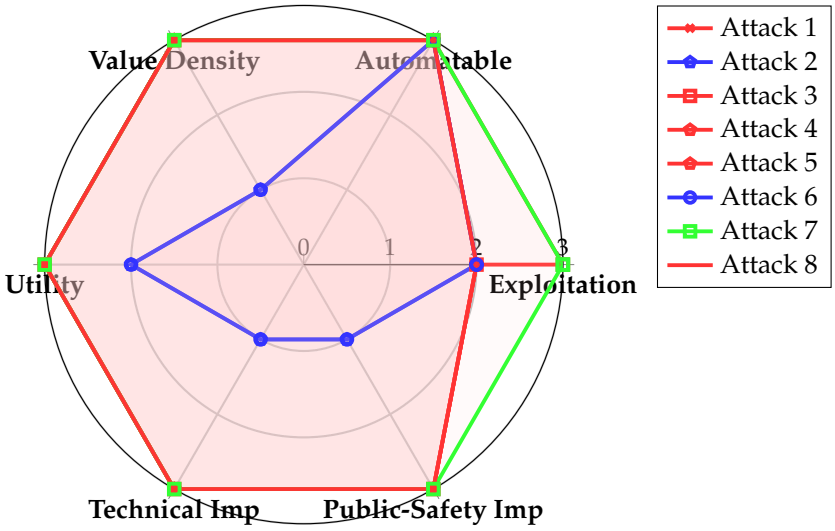
- (1) → (E:P/A:Y/V:C/U:S/T:T/P:S) = Immediate (Very High)
- (2) → (E:P/A:Y/V:D/U:E/T:P/P:M) = Scheduled (Medium)
- (3) → (E:A/A:Y/V:C/U:S/T:T/P:S) = Immediate (Very High)
- (4) → (E:P/A:Y/V:C/U:S/T:T/P:S) = Immediate (Very High)
- (5) → (E:P/A:Y/V:C/U:S/T:T/P:S) = Immediate (Very High)
- (6) → (E:P/A:Y/V:D/U:E/T:P/P:M) = Scheduled (Medium)
- (7) → (E:A/A:Y/V:C/U:S/T:T/P:S) = Immediate (Very High)
- (8) → (E:P/A:Y/V:C/U:S/T:T/P:S) = Immediate (Very High)

Table A8 presents the SSVC scores assigned by the three LLMs. The primary challenge encountered during this assessment was the ability of the LLMs to remain **up-to-date**. Specifically, some attacks might have been actively exploited in the past but are now less prevalent. For example, in the case of the third and fourth attacks [2,120], some LLMs classified these as "Active," while others evaluated them at the "Proof-of-Concept" stage. Determining which LLM is correct in such scenarios is challenging. To address this, we prompted the LLMs to confirm their assessments by asking clarifying questions such as: *"Are there proofs of recent active exploitations of these attacks?"* This approach led to adjustments in certain scores. For instance, LLAMA-3 revised its assessment for the third attack from "Active" to "Proof-of-Concept," explaining that while the attack was previously active, there is no current evidence of active exploitation.

The final scores are visualized below in the spider chart. The results indicate that the SSVC scores align closely with those of DREAD, demonstrating that these Black-box jailbreak attacks are highly dangerous and straightforward to exploit, regardless of whether they target the CIA triad or financial aspects.

 doi:10.20944/preprints202412.2419.v1






For the subsequent assessments, we will present only the results, as the justifications follow the same reasoning outlined for the White-box and Black-box Jailbreak attacks.

### *6.3. Assessment of Prompt Injection attacks*

The third assessment is that of PI attacks, we evaluate the attacks described earlier in Section 4.2: (1) Ignore Previous Prompt [102], (2) Indirect Instruction Injection [52] (3) Formalised Prompt Injection [86], (4) Injection through file input [5], (5) Universal Prompt Injection [82], (6) Virtual Prompt Injection [152], (7) Chat History Tampering [139], (8) JudgeDeceiverAttack [122].

#### 6.3.1. With DREAD

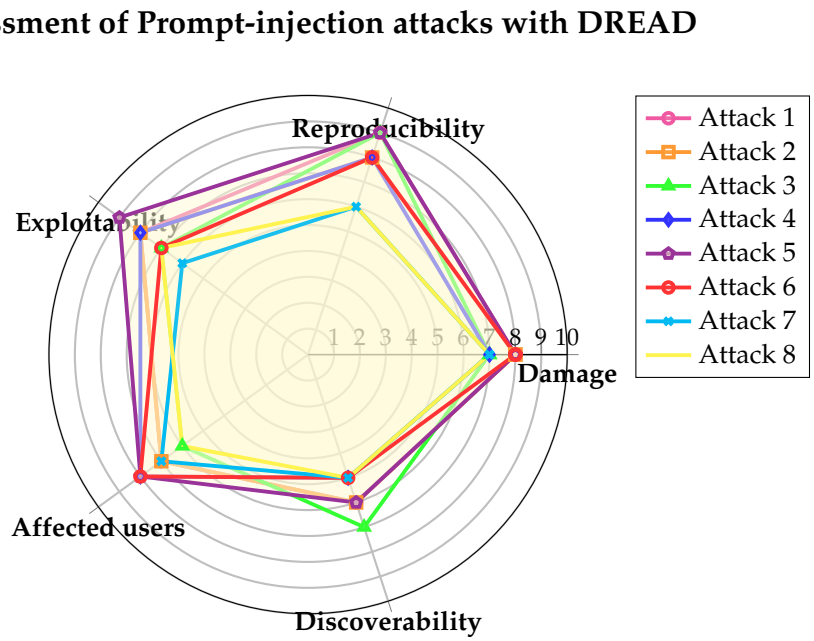
As done before, we start by evaluating the eight prompt injection attacks using DREAD [92], and we find the corresponding vulnerability vectors as follows:

- (1) → (D:8/R:9/E:8/A:7/D:6) = 7.6 (High)
- (2) → (D:8/R:8/E:8/A:7/D:6) = 7.4 (High)
- (3) → (D:7/R:9/E:7/A:6/D:7) = 7.2 (High)
- (4) → (D:7/R:8/E:8/A:8/D:5) = 7.2 (High)
- (5) → (D:8/R:9/E:9/A:8/D:6) = 8 (High)
- (6) → (D:8/R:8/E:7/A:8/D:5) = 7.2 (High)
- (7) → (D:7/R:6/E:6/A:7/D:5) = 6.2 (Medium)
- (8) → (D:7/R:6/E:7/A:6/D:5) = 6.2 (Medium)

The detailed scores are shown in Table A9, with the final results visualized in the Spider-chart below. The DREAD analysis reveals that Prompt-Injection attacks cause significant damage to systems and impact a wide range of users, but they are comparatively harder to exploit and reproduce than Jailbreak attacks.

 doi:10.20944/preprints202412.2419.v1






6.3.2. With CVSS

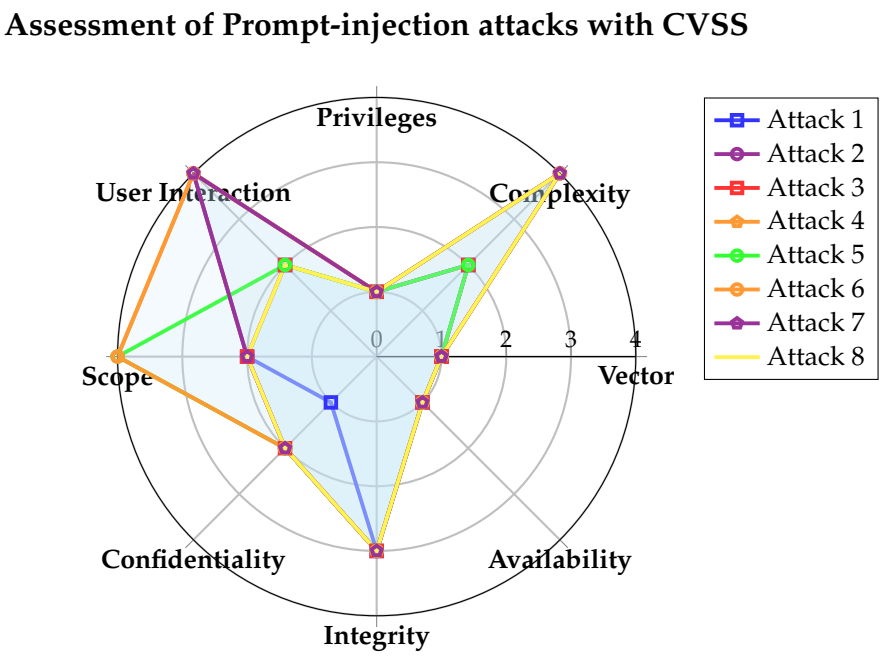
The second assessment of PI attacks is done with CVSS [114]. The corresponding CVSS Vectors are shown below:

- (1) → (AV:N/AC:L/PR:N/UI:N/S:U/C:N/I:H/A:N) = 7.5 (High)
- (2) → (AV:N/AC:H/PR:N/UI:R/S:U/C:L/I:H/A:N) = 5.9 (Medium)
- (3) → (AV:N/AC:L/PR:N/UI:N/S:U/C:L/I:H/A:N) = 8.2 (High)
- (4) → (AV:N/AC:H/PR:N/UI:R/S:C/C:L/I:H/A:N) = 6.9 (Medium)
- (5) → (AV:N/AC:L/PR:N/UI:N/S:C/C:L/I:H/A:N) = 9.3 (Critical)
- (6) → (AV:N/AC:H/PR:N/UI:R/S:C/C:L/I:H/A:N) = 6.9 (Medium)
- (7) → (AV:N/AC:H/PR:N/UI:R/S:U/C:L/I:H/A:N) = 5.9 (Medium)
- (8) → (AV:N/AC:H/PR:N/UI:N/S:U/C:L/I:H/A:N) = 6.5 (Medium)

The detailed CVSS results are presented in Table A10, with the final scores visualized in the Spider-chart below. The analysis indicates that Prompt-Injection attacks share similarities with Jailbreak attacks, as they are primarily executed remotely through the network. However, they are slightly more complex to perform than Jailbreak attacks. These attacks predominantly target system integrity, have a lesser impact on confidentiality, and do not affect availability.

 doi:10.20944/preprints202412.2419.v1






6.3.3. With OWASP Risk Rating

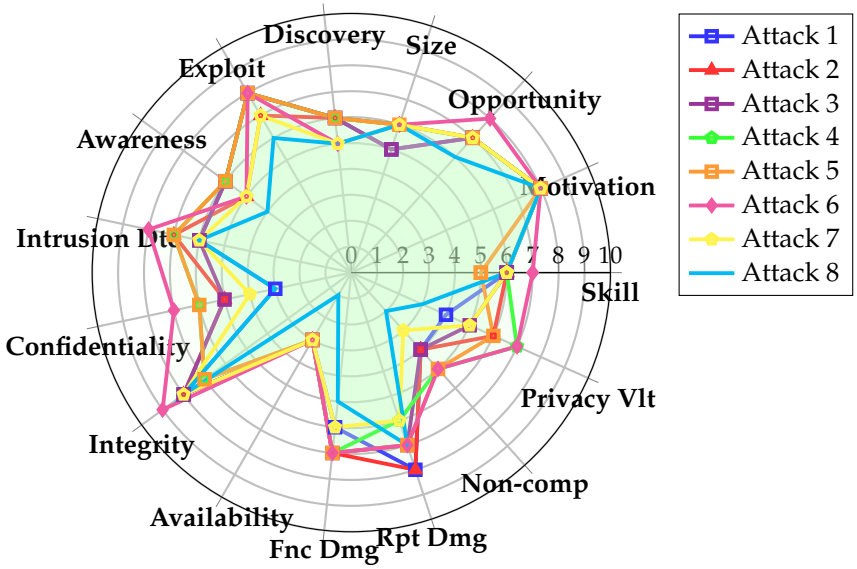
Another evaluation of Prompt Injection attacks is done with OWASP Risk Rating [140]. The corresponding vulnerability vectors of each attack is:

- (1) → (SL:6/M:8/O:7/S:5/ED:6/EE:8/A:6/ID:6/LC:3/LI:8/LA:3/FD:6/RD:8/NC:4/PV:4) = 3.3 (High)
- (2) → (SL:6/M:8/O:7/S:6/ED:6/EE:7/A:5/ID:7/LC:5/LI:8/LA:3/FD:7/RD:8/NC:4/PV:6) = 3.8 (High)
- (3) → (SL:6/M:8/O:7/S:5/ED:6/EE:8/A:6/ID:6/LC:5/LI:8/LA:3/FD:7/RD:8/NC:4/PV:5) = 3.7 (High)
- (4) → (SL:6/M:8/O:7/S:6/ED:6/EE:8/A:6/ID:7/LC:6/LI:7/LA:3/FD:7/RD:8/NC:5/PV:7) = 4.1 (Critical)
- (5) → (SL:5/M:8/O:7/S:6/ED:6/EE:8/A:6/ID:7/LC:6/LI:7/LA:3/FD:7/RD:8/NC:5/PV:6) = 3.8 (High)
- (6) → (SL:7/M:8/O:8/S:6/ED:5/EE:8/A:5/ID:8/LC:7/LI:9/LA:3/FD:7/RD:8/NC:5/PV:7) = 4.4 (Critical)
- (7) → (SL:6/M:8/O:7/S:6/ED:5/EE:7/A:5/ID:6/LC:4/LI:8/LA:3/FD:6/RD:7/NC:3/PV:5) = 2.6 (Medium)
- (8) → (SL:6/M:7/O:6/S:6/ED:5/EE:6/A:4/ID:6/LC:3/LI:8/LA:1/FD:5/RD:7/NC:2/PV:3) = 1.9 (Medium)

The assessments conducted using three LLMs are detailed in Table A11, and the final scores are depicted in the Spider-chart below for enhanced visualization. The OWASP RR results corroborate that Prompt-Injection attacks exert a greater impact on integrity than on confidentiality and availability. Additionally, they highlight the significant influence these attacks have on privacy violations and reputation damage, which are critical factors beyond the technical scope. Notably, the assessments also reveal a general lack of awareness among public users regarding these specific threats.

 doi:10.20944/preprints202412.2419.v1






6.3.4. With SSVC

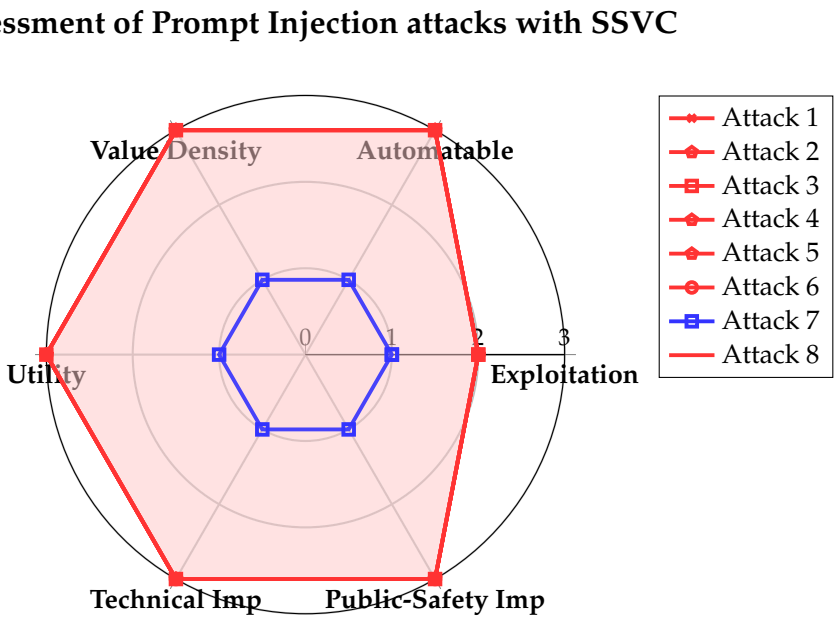
A last evaluation is performed using SSVC [128] as done before, the results of the assessments are presented below:

- (1) → (E:P/A:Y/V:C/U:S/T:T/P:S) = Immediate (Very High)
- (2) → (E:P/A:Y/V:C/U:S/T:T/P:S) = Immediate (Very High)
- (3) → (E:P/A:Y/V:C/U:S/T:T/P:S) = Immediate (Very High)
- (4) → (E:P/A:Y/V:C/U:S/T:T/P:S) = Immediate (Very High)
- (5) → (E:P/A:Y/V:C/U:S/T:T/P:S) = Immediate (Very High)
- (6) → (E:P/A:Y/V:C/U:S/T:T/P:S) = Immediate (Very High)
- (7) → (E:N/A:N/V:D/U:L/T:P/P:M) = Defer (Low)
- (8) → (E:P/A:Y/V:C/U:S/T:T/P:S) = Immediate (Very High)

Table A12 provides the detailed SSVC assessments conducted with the three LLMs. These scores offer additional insights beyond those captured by other metrics, emphasizing that Prompt Injection attacks are highly automatable, posing significant risks to both technical systems and public safety. The averaged results are visualized in the Spider-chart below for enhanced clarity.

 doi:10.20944/preprints202412.2419.v1






*6.4. Assessment of Evasion attacks*

The forth experiment is evaluating eight Evasion attacks described in Section 4.3: (1) Hot Flip [39], (2) PWWS [109], (3) TypoAttack [103], (4) VIPER [40], (5) CheckList [110], (6) BertAttack [72], (7) GBDA [55], (8) TF-Attack [77].

6.4.1. With DREAD

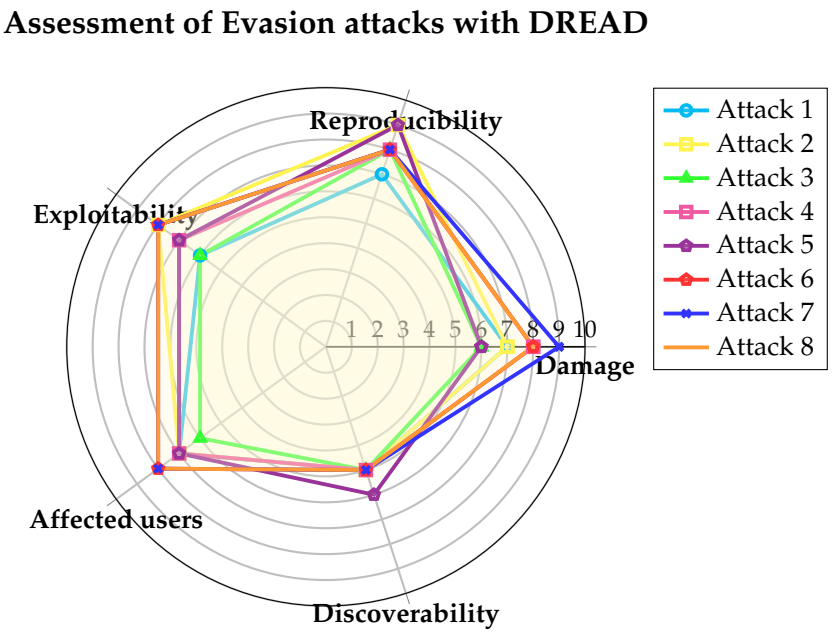
The first evaluation is done using DREAD [92]. The corresponding vulnerability vectors as follows:

- (1) → (D:7/R:7/E:6/A:7/D:5) = 6.4 (Medium)
- (2) → (D:7/R:9/E:8/A:7/D:5) = 7.2 (High)
- (3) → (D:6/R:8/E:6/A:6/D:5) = 6.2 (Medium)
- (4) → (D:8/R:8/E:7/A:7/D:5) = 7 (High)
- (5) → (D:6/R:9/E:7/A:7/D:6) = 7 (High)
- (6) → (D:8/R:8/E:8/A:8/D:5) = 7.4 (High)
- (7) → (D:9/R:8/E:8/A:8/D:5) = 7.6 (High)
- (8) → (D:8/R:8/E:8/A:8/D:5) = 7.4 (High)

Table A13 presents the detailed assessments conducted with the three LLMs, with the final scores visualized in the Spider-chart below. The DREAD evaluation reveals that evasion attacks generally cause medium-to-high damage and are highly reproducible, easily exploitable, and difficult to detect, while having the potential to impact a wide range of users. This underscores the critical need to mitigate such attacks.

 doi:10.20944/preprints202412.2419.v1






6.4.2. With CVSS

The second assessment of Evasion attacks is done with CVSS [114]. The corresponding CVSS Vectors are shown below:

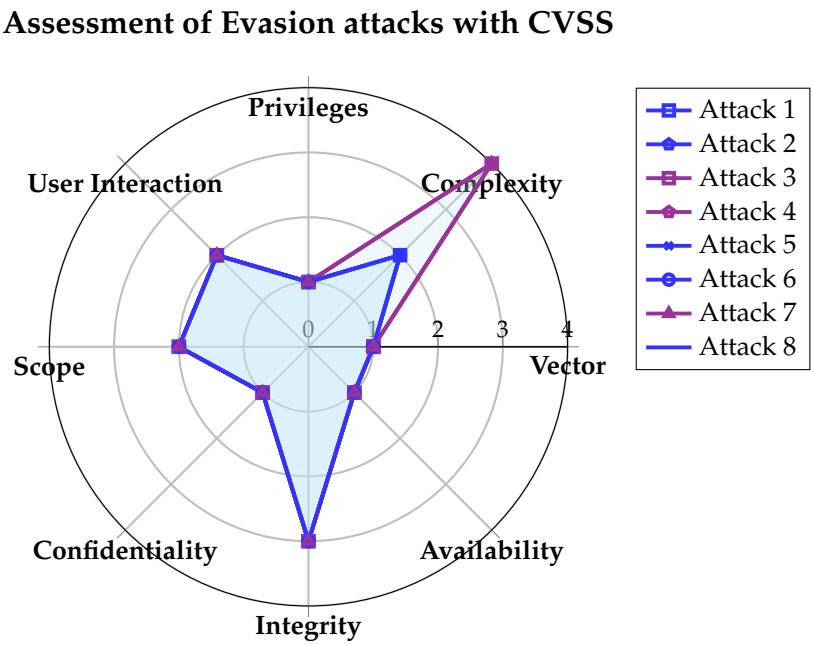
- (1) → (AV:N/AC:L/PR:N/UI:N/S:U/C:N/I:H/A:N) = 7.5 (High)
- (2) → (AV:N/AC:L/PR:N/UI:N/S:U/C:N/I:H/A:N) = 7.5 (High)
- (3) → (AV:N/AC:H/PR:N/UI:N/S:U/C:N/I:H/A:N) = 5.9 (Medium)
- (4) → (AV:N/AC:H/PR:N/UI:N/S:U/C:N/I:H/A:N) = 5.9 (Medium)
- (5) → (AV:N/AC:L/PR:N/UI:N/S:U/C:N/I:H/A:N) = 7.5 (High)
- (6) → (AV:N/AC:L/PR:N/UI:N/S:U/C:N/I:H/A:N) = 7.5 (High)
- (7) → (AV:N/AC:H/PR:N/UI:N/S:U/C:N/I:H/A:N) = 5.9 (Medium)
- (8) → (AV:N/AC:L/PR:N/UI:N/S:U/C:N/I:H/A:N) = 7.5 (High)

Table A14 displays the scores provided by the three LLMs along with their average. For enhanced clarity and ease of interpretation, the score vectors are visualized in the Spider-chart below.

The CVSS evaluations reveal a consistent scoring pattern for evasion attacks, emphasizing their typical characteristics. These attacks are often performed over a network, require minimal complexity, and do not necessitate privileges or user interaction. While they have no impact on data Confidentiality or Availability, they can significantly affect Integrity.

 doi:10.20944/preprints202412.2419.v1






6.4.3. With OWASP Risk Rating

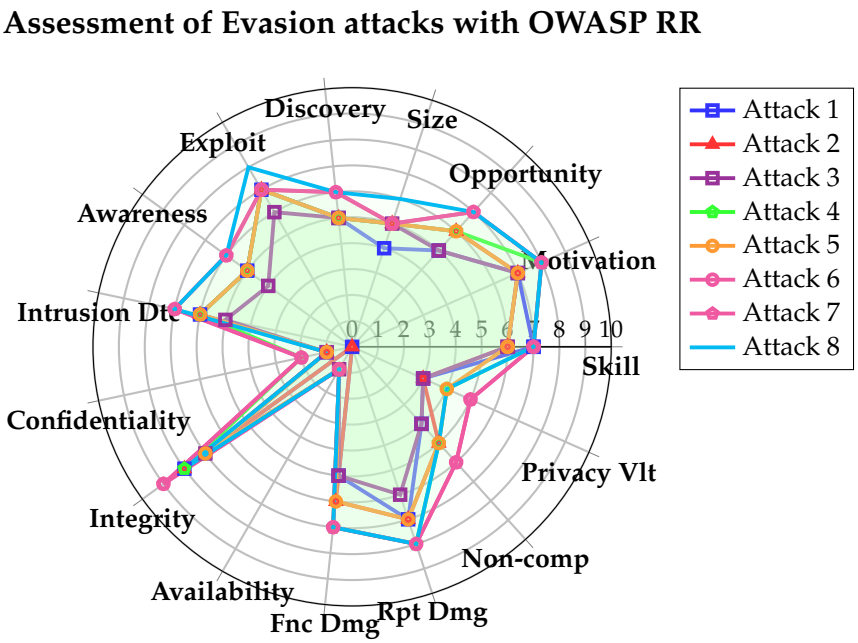
Another evaluation of Evasion attacks is done with OWASP Risk Rating [140]. The corresponding vulnerability vectors of each attack is:

- (1) → (SL:7/M:7/O:5/S:4/ED:5/EE:7/A:5/ID:6/LC:1/LI:8/LA:0/FD:5/RD:7/NC:4/PV:3) = 2.2 (Medium)
- (2) → (SL:6/M:7/O:6/S:5/ED:5/EE:7/A:5/ID:6/LC:1/LI:8/LA:0/FD:6/RD:7/NC:5/PV:3) = 2.4 (Medium)
- (3) → (SL:6/M:7/O:5/S:5/ED:5/EE:6/A:4/ID:5/LC:1/LI:7/LA:1/FD:5/RD:6/NC:4/PV:3) = 2 (Medium)
- (4) → (SL:7/M:8/O:6/S:5/ED:5/EE:7/A:5/ID:6/LC:2/LI:8/LA:1/FD:7/RD:8/NC:5/PV:4) = 3 (High)
- (5) → (SL:6/M:7/O:6/S:5/ED:5/EE:7/A:5/ID:6/LC:1/LI:7/LA:1/FD:6/RD:7/NC:5/PV:4) = 2.5 (Medium)
- (6) → (SL:7/M:8/O:7/S:5/ED:6/EE:7/A:6/ID:7/LC:2/LI:9/LA:1/FD:7/RD:8/NC:6/PV:5) = 3.5 (High)
- (7) → (SL:7/M:8/O:7/S:5/ED:6/EE:7/A:6/ID:7/LC:2/LI:9/LA:1/FD:7/RD:8/NC:6/PV:5) = 3.5 (High)
- (8) → (SL:7/M:8/O:7/S:6/ED:6/EE:8/A:6/ID:7/LC:1/LI:8/LA:1/FD:7/RD:8/NC:5/PV:4) = 3.2 (High)

The detailed OWASP RR scoring is outlined in Table A15, offering insights consistent with those from the CVSS assessments. It highlights that evasion attacks demand only a moderate level of skill and motivation to be executed, are easily exploitable, and are relatively unknown to defenders, making them challenging to detect and mitigate. These attacks pose a significant threat to data integrity while remaining harmless to Confidentiality and Availability. Additionally, OWASP RR sheds light on the substantial financial and reputational impact these attacks can impose on targeted organizations.

 doi:10.20944/preprints202412.2419.v1






6.4.4. With SSVC

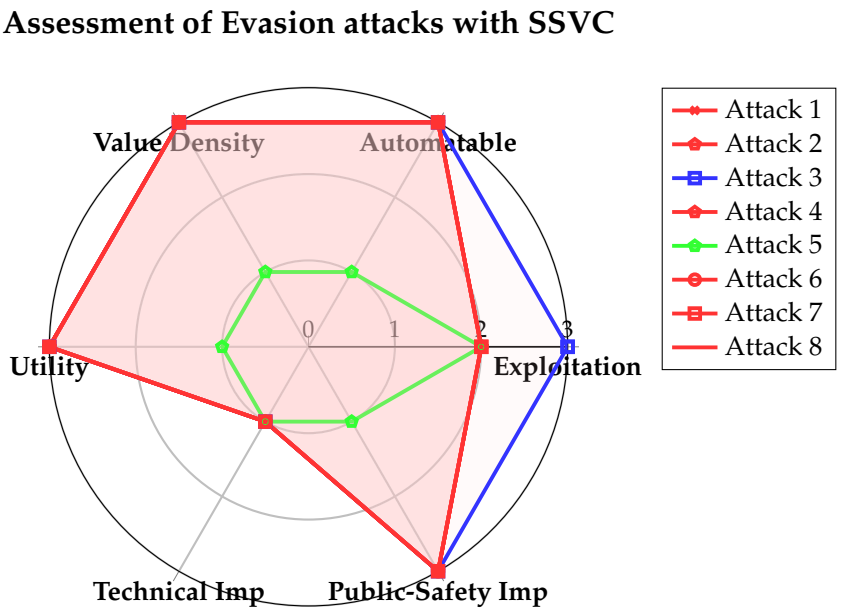
We continue the evaluation of Evasion attacks with SSVC [128] as a last metric, the results of the assessments are shown and detailed below:

- (1) → (E:P/A:Y/V:C/U:S/T:P/P:S) = Immediate (Very High)
- (2) → (E:P/A:Y/V:C/U:S/T:P/P:S) = Immediate (Very High)
- (3) → (E:A/A:Y/V:C/U:S/T:P/P:S) = Immediate (Very High)
- (4) → (E:P/A:Y/V:C/U:S/T:P/P:S) = Immediate (Very High)
- (5) → (E:P/A:N/V:D/U:L/T:P/P:M) = Scheduled (Medium)
- (6) → (E:P/A:Y/V:C/U:S/T:P/P:S) = Immediate (Very High)
- (7) → (E:P/A:Y/V:C/U:S/T:P/P:S) = Immediate (Very High)
- (8) → (E:P/A:Y/V:C/U:S/T:P/P:S) = Immediate (Very High)

The scores are visualized below in a Spider-chart, with the detailed assessments provided in Table A16. Notably, the SSVC results align with those of DREAD and OWASP RR, emphasizing that evasion attacks are frequently exploited by attackers. Additionally, SSVC highlights that these attacks are highly automatable and rewarding, making them particularly valuable to adversaries. However, it suggests that while evasion attacks pose minimal technical threats to organizations, their primary danger lies in their significant potential to compromise public safety, especially in scenarios involving object detection and classification.

 doi:10.20944/preprints202412.2419.v1






*6.5. Assessment of Model Extraction attacks*

Model Extraction are the fifth attacks we evaluate with the five vulnerability metrics. The attacks were presented earlier in Section 4.4 and are respectively: (1) User Data Extraction [12], (2) LLM Tricks [156], (3) Analysing PII Leakage [89], (4) ETHICIST [162], (5) Scalable Extraction [95], (6) Output2Prompt [158], (7) PII-Compass [94], (8) Alpaca VS Vicuna [64].

6.5.1. With DREAD

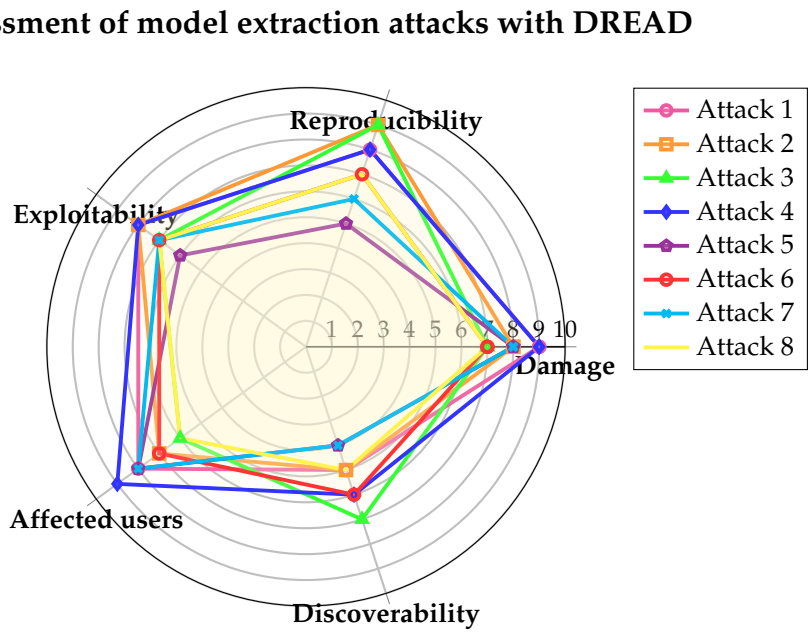
We start evaluating Extraction attacks using DREAD [92]. The corresponding vulnerability vectors as follows:

- (1) → (D:9/R:8/E:8/A:8/D:5) = 7.6 (High)
- (2) → (D:8/R:9/E:8/A:7/D:5) = 7.4 (High)
- (3) → (D:9/R:8/E:8/A:9/D:6) = 8 (High)
- (4) → (D:8/R:8/E:7/A:7/D:5) = 7 (High)
- (5) → (D:8/R:5/E:6/A:8/D:4) = 6.2 (Medium)
- (6) → (D:7/R:7/E:7/A:7/D:6) = 6.8 (Medium)
- (7) → (D:8/R:6/E:7/A:8/D:4) = 6.8 (Medium)
- (8) → (D:7/R:7/E:7/A:6/D:5) = 6.4 (Medium)

The detailed scores for this fifth type of attack are shown in Table A17, with the score vectors visualized in a Spider-chart below for clarity. The DREAD assessment reveals that the exploitability and discoverability of model extraction attacks vary depending on the specific implementation. However, all these attacks share a high level of danger to systems due to their potential to cause significant damage. Additionally, the analysis highlights that such attacks can directly or indirectly affect multiple users, while remaining relatively challenging to detect.

 doi:10.20944/preprints202412.2419.v1






6.5.2. With CVSS

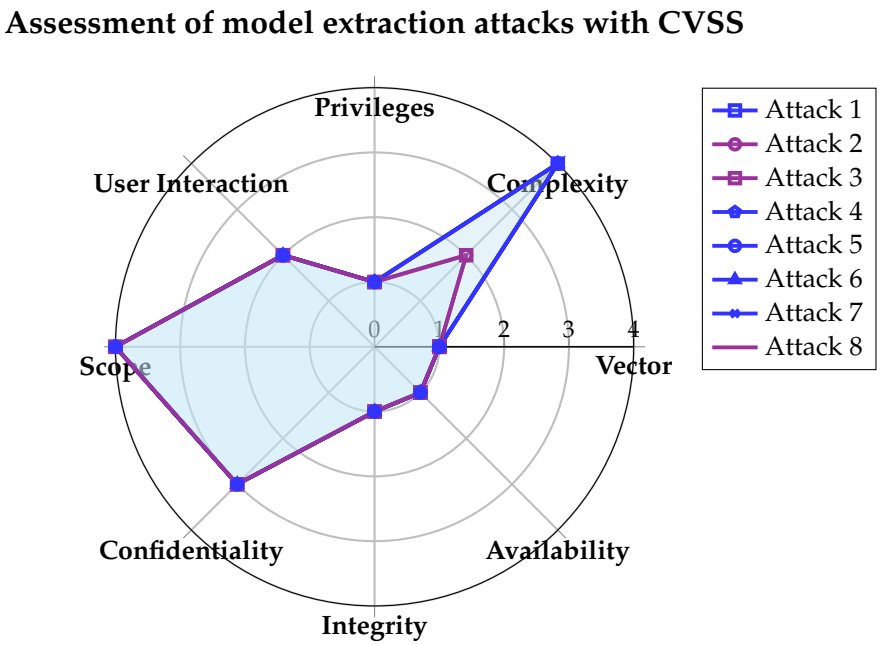
The second assessment of Model Extraction attacks is done with CVSS [114]. The corresponding Vectors are:

- (1) → (AV:N/AC:H/PR:N/UI:N/S:C/C:H/I:N/A:N) = 6.8 (Medium)
- (2) → (AV:N/AC:L/PR:N/UI:N/S:C/C:H/I:N/A:N) = 8.6 (High)
- (3) → (AV:N/AC:L/PR:N/UI:N/S:C/C:H/I:N/A:N) = 8.6 (High)
- (4) → (AV:N/AC:H/PR:N/UI:N/S:C/C:H/I:N/A:N) = 6.8 (Medium)
- (5) → (AV:N/AC:H/PR:N/UI:N/S:C/C:H/I:N/A:N) = 6.8 (Medium)
- (6) → (AV:N/AC:H/PR:N/UI:N/S:C/C:H/I:N/A:N) = 6.8 (Medium)
- (7) → (AV:N/AC:H/PR:N/UI:N/S:C/C:H/I:N/A:N) = 6.8 (Medium)
- (8) → (AV:N/AC:L/PR:N/UI:N/S:C/C:H/I:N/A:N) = 8.6 (High)

The CVSS scores assigned by each LLM are detailed in Table A18, with the final vectors visualized in the Spider-chart below. Compared to DREAD, CVSS provides more granular insights into the nature of the damage caused by these attacks, particularly their impact on confidentiality. Additionally, the assessment highlights that these attacks typically do not require specific privileges or user interaction for execution. However, their scope can vary depending on the type of data extracted, making them broader in target range than previous attack types.

 doi:10.20944/preprints202412.2419.v1






6.5.3. With OWASP Risk Rating

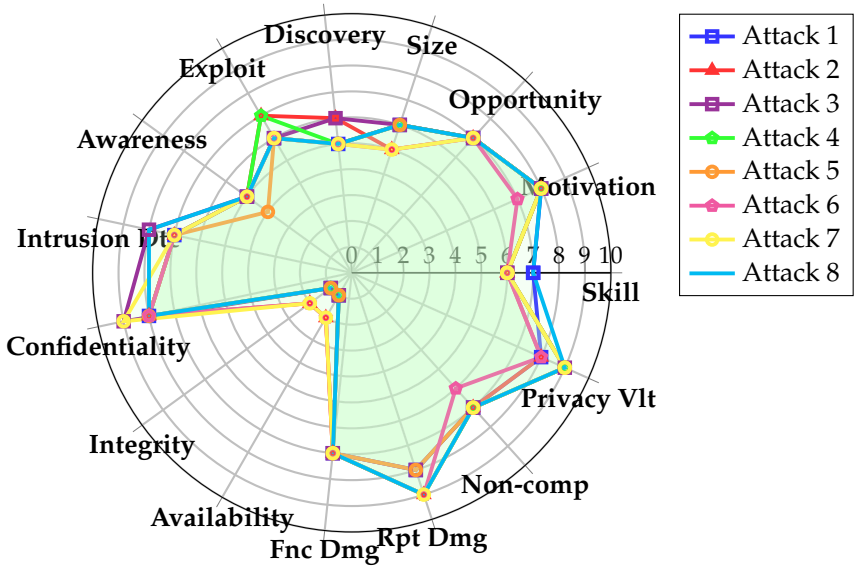
A third evaluation of Model Extraction attacks is done using OWASP RR [140]. The corresponding vulnerability vectors of each attack is:

- (1) → (SL:7/M:8/O:7/S:6/ED:5/EE:6/A:5/ID:7/LC:8/LI:1/LA:1/FD:7/RD:8/NC:7/PV:8) = 3.5 (High)
- (2) → (SL:6/M:8/O:7/S:5/ED:6/EE:7/A:5/ID:7/LC:8/LI:2/LA:2/FD:7/RD:9/NC:7/PV:8) = 3.8 (High)
- (3) → (SL:6/M:8/O:7/S:6/ED:6/EE:6/A:5/ID:8/LC:9/LI:1/LA:1/FD:7/RD:8/NC:7/PV:9) = 3.7 (High)
- (4) → (SL:6/M:8/O:7/S:6/ED:5/EE:7/A:5/ID:7/LC:8/LI:1/LA:1/FD:7/RD:9/NC:7/PV:9) = 3.6 (High)
- (5) → (SL:6/M:8/O:7/S:6/ED:5/EE:6/A:4/ID:7/LC:8/LI:1/LA:1/FD:7/RD:8/NC:7/PV:9) = 3.4 (High)
- (6) → (SL:6/M:7/O:7/S:5/ED:5/EE:6/A:5/ID:7/LC:8/LI:2/LA:2/FD:7/RD:9/NC:6/PV:8) = 3.5 (High)
- (7) → (SL:6/M:8/O:7/S:5/ED:5/EE:6/A:5/ID:7/LC:9/LI:2/LA:2/FD:7/RD:9/NC:7/PV:9) = 3.8 (Critical)
- (8) → (SL:7/M:8/O:7/S:6/ED:5/EE:6/A:5/ID:8/LC:8/LI:1/LA:1/FD:7/RD:9/NC:7/PV:9) = 3.7 (High)

The detailed scores assigned by the LLMs are presented in Table A19, with the final scores of each attack shown in the chart below for better visualization of their assessments. The OWASP RR scores align with the findings from DREAD and CVSS, offering additional insights. This metric reveals that Model Extraction attacks require only a moderate level of skill and motivation to be performed, and are easily exploitable. Notably, system administrators and defenders often lack awareness of these attacks and their potential risks, particularly their significant impact on data confidentiality. Additionally, these attacks pose a substantial threat to an organization's finances and reputation, while also leading to privacy violations that can result in increased audit challenges.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 December 2024 doi:10.20944/preprints202412.2419.v1






6.5.4. With SSVC

The forth and last evaluation of Model-Extraction attacks is done with SSVC [128] as previously, the results of the assessments are shown and detailed below:

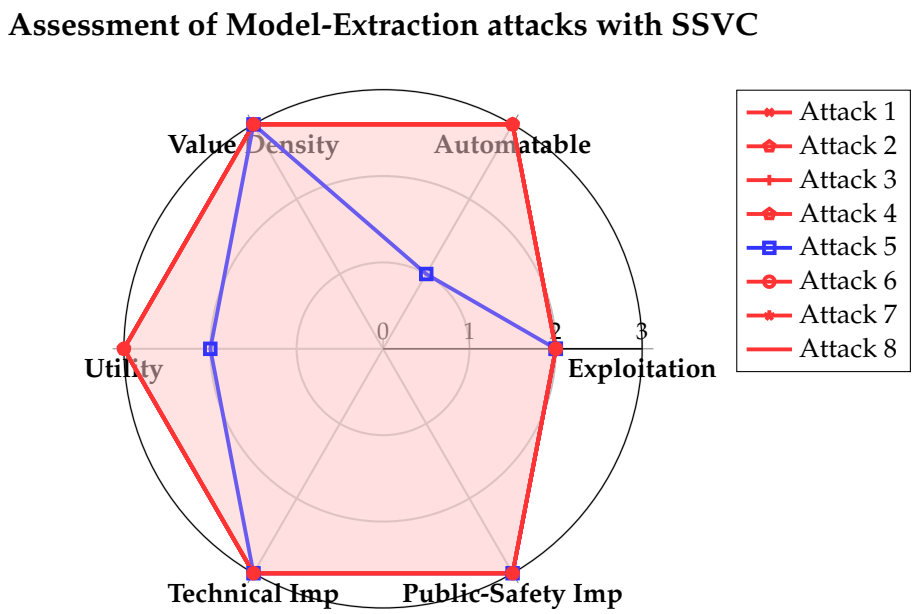
- (1) → (E:P/A:Y/V:C/U:S/T:T/P:S) = Immediate (Very High)
- (2) → (E:P/A:Y/V:C/U:S/T:T/P:S) = Immediate (Very High)
- (3) → (E:P/A:Y/V:C/U:S/T:T/P:S) = Immediate (Very High)
- (4) → (E:P/A:Y/V:C/U:S/T:T/P:S) = Immediate (Very High)
- (5) → (E:P/A:N/V:C/U:E/T:T/P:S) = Immediate (Very High)
- (6) → (E:P/A:Y/V:C/U:S/T:T/P:S) = Immediate (Very High)
- (7) → (E:P/A:Y/V:C/U:S/T:T/P:S) = Immediate (Very High)
- (8) → (E:P/A:Y/V:C/U:S/T:T/P:S) = Immediate (Very High)

Table A20 presents the individual scores provided by each LLM along with their average, which is visualized below in a spider chart for a clearer understanding of the characteristics of Model-Extraction attacks.

The SSVC assessments highlight the ease of automation and high rewards associated with these attacks, making them particularly effective for adversaries. Notably, this metric emphasizes the significant impact these attacks can have on both the technical aspects of organizations and public safety, particularly by jeopardizing the privacy of user data.

 doi:10.20944/preprints202412.2419.v1






*6.6. Assessment of Model Inference attacks*

The next type of attacks we assess are Model Inference attacks, presented in Section 4.5: (1) LIRA [11], (2) Detecting Pretraining Data [123], (3) Neighborhood Comparison [91], (4) ProPILE [65], (5) Analysing PII Leakage [89], (6) Conrecall [134], (7) MIA-LLM [45], (8) DeCop [37].

6.6.1. With DREAD

We start evaluating Model Inference attacks using DREAD [92]. The corresponding vulnerability vectors as follows:

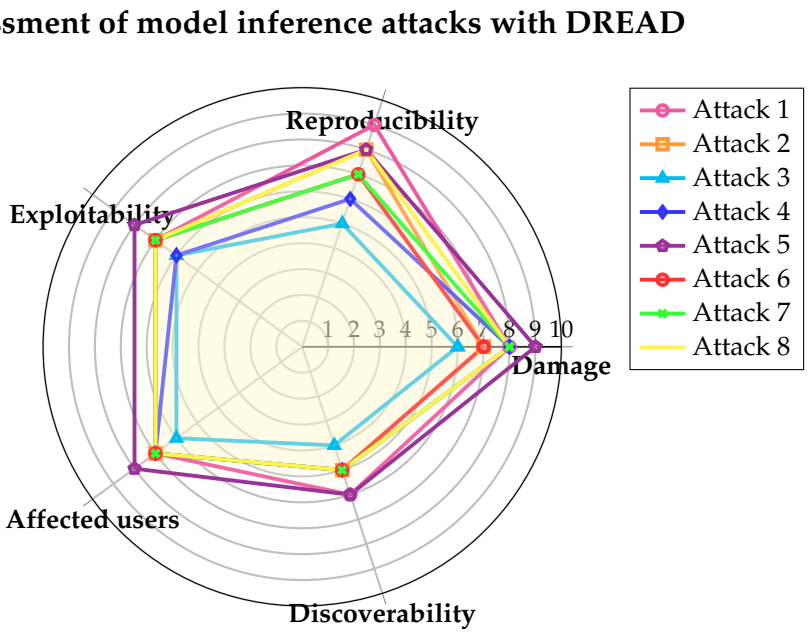
- (1) → (D:8/R:9/E:7/A:7/D:6) = 7.4 (High)
- (2) → (D:7/R:8/E:7/A:7/D:5) = 6.8 (Medium)
- (3) → (D:6/R:5/E:6/A:6/D:5) = 5.6 (Medium)
- (4) → (D:8/R:6/E:6/A:7/D:5) = 6.4 (Medium)
- (5) → (D:9/R:8/E:8/A:9/D:6) = 8 (High)
- (6) → (D:7/R:7/E:7/A:7/D:5) = 6.6 (Medium)
- (7) → (D:8/R:7/E:7/A:7/D:5) = 6.8 (Medium)
- (8) → (D:8/R:8/E:7/A:7/D:5) = 6.4 (Medium)

Table A21 provides the detailed scores for this sixth type of attack, as assessed by the three LLMs along with their average. For better visualization and interpretation, these scores are represented below in a spider chart.

The DREAD assessment reveals that Model Inference attacks share many characteristics with Model Extraction attacks. Both pose significant damage to systems and organizations, are highly reproducible and exploitable, and can impact a large number of users either directly or indirectly. They are also moderately challenging to discover. The primary distinction lies in their danger levels—Model Extraction attacks are slightly more harmful as they enable the extraction of models or user data, whereas Model Inference attacks are more specific, allowing adversaries to determine whether certain data was part of the training set.

 doi:10.20944/preprints202412.2419.v1






6.6.2. With CVSS

The second assessment of Model Inference attacks is done with CVSS [114]. The corresponding Vectors are:

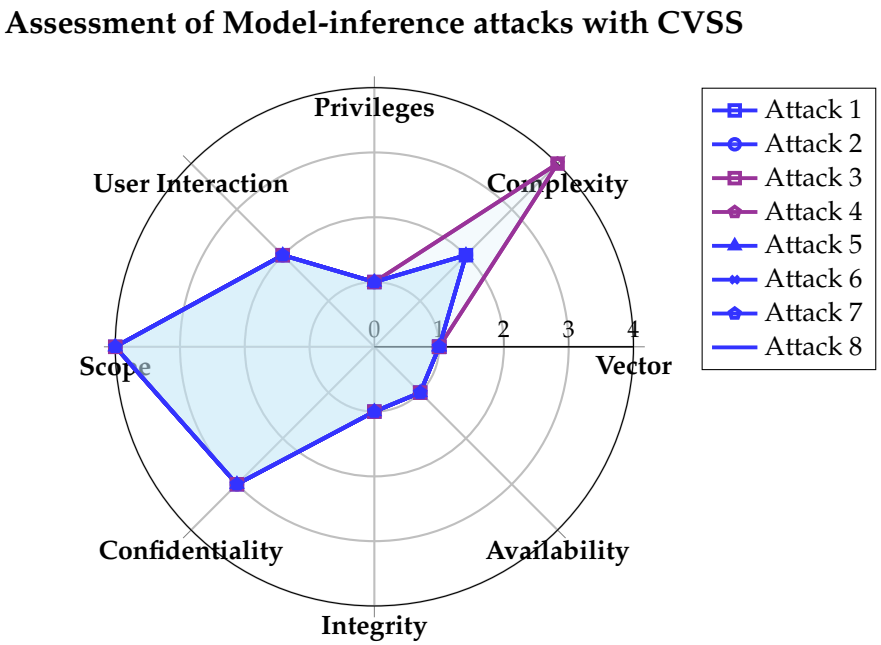
- (1) → (AV:N/AC:L/PR:N/UI:N/S:C/C:H/I:N/A:N) = 8.6 (High)
- (2) → (AV:N/AC:L/PR:N/UI:N/S:C/C:H/I:N/A:N) = 8.6 (High)
- (3) → (AV:N/AC:H/PR:N/UI:N/S:C/C:H/I:N/A:N) = 6.8 (Medium)
- (4) → (AV:N/AC:H/PR:N/UI:N/S:C/C:H/I:N/A:N) = 6.8 (Medium)
- (5) → (AV:N/AC:L/PR:N/UI:N/S:C/C:H/I:N/A:N) = 8.6 (High)
- (6) → (AV:N/AC:L/PR:N/UI:N/S:C/C:H/I:N/A:N) = 8.6 (High)
- (7) → (AV:N/AC:L/PR:N/UI:N/S:C/C:H/I:N/A:N) = 8.6 (High)
- (8) → (AV:N/AC:L/PR:N/UI:N/S:C/C:H/I:N/A:N) = 8.6 (High)

The complete list of scores is detailed in Table A22, with the final averaged scores visualized below in the spider chart. The chart closely resembles that of Model Extraction attacks, differing primarily in that Model Inference attacks are less complex to execute.

Notably, these attacks can be carried out remotely via the network without requiring any user privileges or interaction. They have a significant impact on confidentiality and exhibit a variable scope, as the extracted information can be used to target other systems or users.

 doi:10.20944/preprints202412.2419.v1






6.6.3. With OWASP Risk Rating

A third evaluation of Model Inference attacks is done using OWASP RR [140]. The corresponding vulnerability vectors of each attack is:

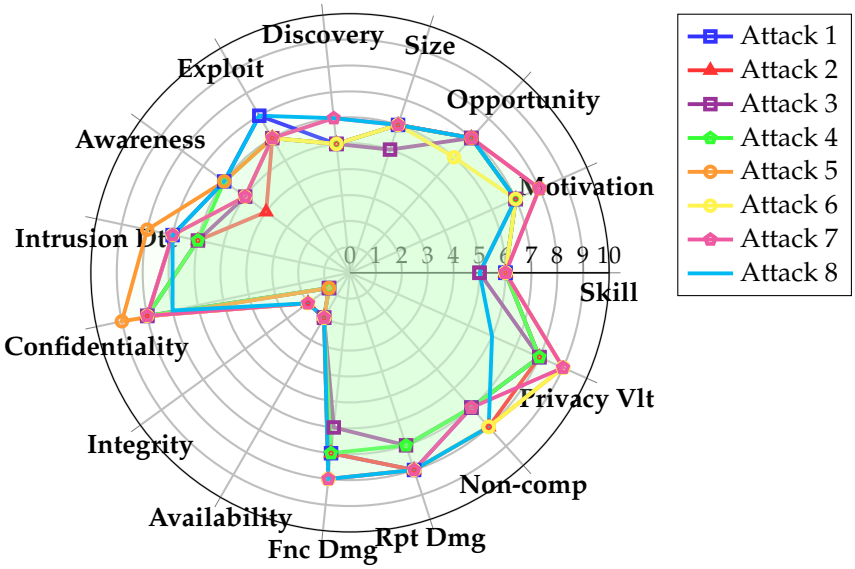
- (1) → (SL:6/M:7/O:7/S:6/ED:5/EE:7/A:6/ID:7/LC:8/LI:1/LA:2/FD:7/RD:8/NC:7/PV:8) = 3.5 (High)
- (2) → (SL:6/M:7/O:7/S:6/ED:5/EE:6/A:4/ID:6/LC:8/LI:2/LA:2/FD:7/RD:8/NC:8/PV:8) = 3.5 (Medium)
- (3) → (SL:5/M:7/O:7/S:5/ED:5/EE:6/A:5/ID:6/LC:8/LI:1/LA:2/FD:6/RD:7/NC:7/PV:8) = 3.1 (Medium)
- (4) → (SL:6/M:7/O:7/S:6/ED:5/EE:6/A:6/ID:6/LC:8/LI:1/LA:2/FD:7/RD:7/NC:7/PV:8) = 3.4 (Medium)
- (5) → (SL:6/M:8/O:7/S:6/ED:5/EE:6/A:6/ID:8/LC:9/LI:1/LA:2/FD:8/RD:8/NC:8/PV:9) = 4 (Critical)
- (6) → (SL:6/M:7/O:6/S:6/ED:5/EE:6/A:5/ID:7/LC:8/LI:2/LA:2/FD:8/RD:8/NC:8/PV:9) = 3.7 (Critical)
- (7) → (SL:6/M:8/O:7/S:6/ED:6/EE:6/A:5/ID:7/LC:8/LI:2/LA:2/FD:8/RD:8/NC:7/PV:9) = 3.8 (Critical)
- (8) → (SL:5/M:7/O:7/S:6/ED:6/EE:7/A:6/ID:7/LC:7/LI:2/LA:2/FD:8/RD:8/NC:8/PV:6) = 4.2 (Critical)

The detailed assessments are provided in Table A23, with their averages visualized in the chart below.

The OWASP RR scores align closely with those of CVSS, reaffirming that Model Inference attacks primarily impact confidentiality. These attacks are easy to discover and exploit but are challenging for defenders to detect due to limited awareness of their risks. Additionally, the assessments emphasize that Model Inference attacks significantly violate privacy while being less complex than Model Extraction attacks. Instead of extracting various data from a training set, Model Inference attacks determine whether specific data was included in the training process.

 doi:10.20944/preprints202412.2419.v1






6.6.4. With SSVC

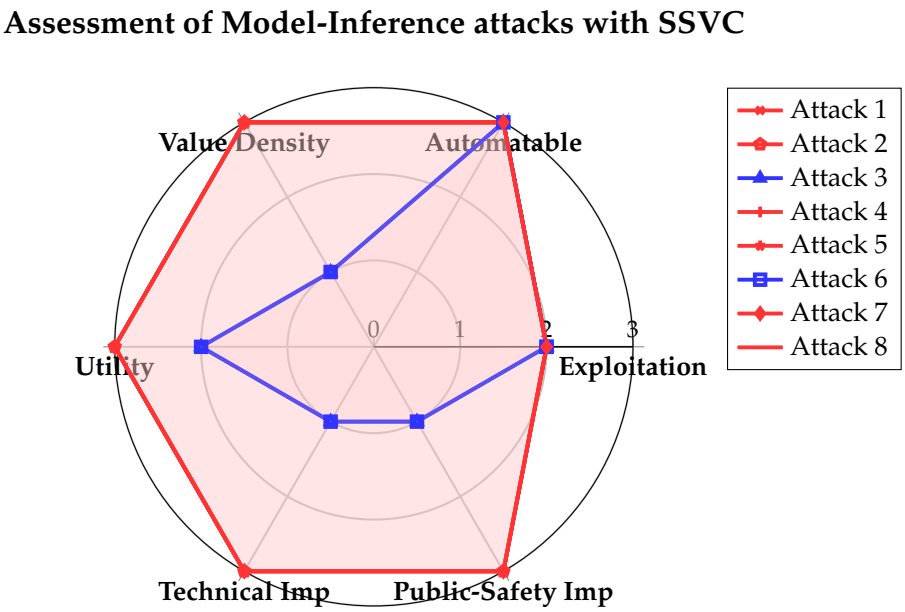
The last evaluation of Model-Inference attacks is performed using SSVC [128], and the results are shown and detailed below:

- (1) → (E:P/A:Y/V:C/U:S/T:T/P:S) = Immediate (Very High)
- (2) → (E:P/A:Y/V:C/U:S/T:T/P:S) = Immediate (Very High)
- (3) → (E:P/A:N/V:D/U:E/T:P/P:M) = Scheduled (Medium)
- (4) → (E:P/A:Y/V:C/U:S/T:T/P:S) = Immediate (Very High)
- (5) → (E:P/A:Y/V:C/U:S/T:T/P:S) = Immediate (Very High)
- (6) → (E:P/A:Y/V:D/U:E/T:P/P:M) = Scheduled (Medium)
- (7) → (E:P/A:Y/V:C/U:S/T:T/P:S) = Immediate (Very High)
- (8) → (E:P/A:Y/V:C/U:S/T:T/P:S) = Immediate (Very High)

The detailed SSVC scores are presented in Table A24, with the final averages visualized in the chart below. Similar to previous attacks, SSVC aligns with the insights provided by other metrics. However, it uniquely highlights that Model Inference attacks are highly automatable and rewarding, making them particularly effective for adversaries. These attacks pose a significant impact not only on the technical aspects of systems but also on user safety and data privacy.

 doi:10.20944/preprints202412.2419.v1






*6.7. Assessment of Poisoning/Trojan/Backdoor attacks*

The last type of attacks we assess are Poisoning, Trojan, and Backdoor attacks, already presented in Section 4.6: (1) TrojLLM [151], (2) Best-of-Venom [7], (3) CodeBreaker [153], (4) Retrieval Poisoning [160], (5) Clinical LLMs [26], (6) BackdoorLLM [75], (7) CBA [58], (8) TA² [136].

6.7.1. With DREAD

We start evaluating Poisoning, Trojan, and Backdoor attacks using DREAD [92]. The corresponding vulnerability vectors as follows:

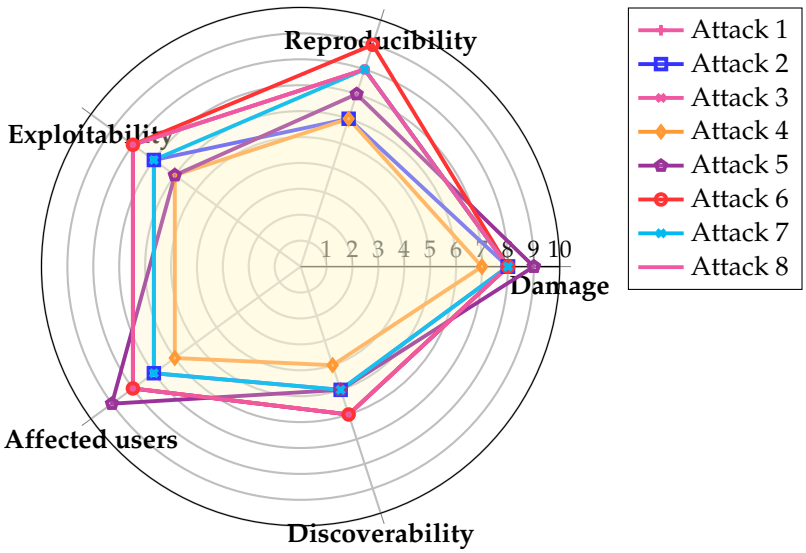
- (1) → (D:8/R:8/E:8/A:8/D:6) = 7.6 (High)
- (2) → (D:8/R:6/E:7/A:7/D:5) = 6.6 (Medium)
- (3) → (D:8/R:8/E:8/A:8/D:6) = 7.6 (High)
- (4) → (D:7/R:6/E:6/A:6/D:4) = 5.8 (Medium)
- (5) → (D:9/R:7/E:6/A:9/D:5) = 7.2 (High)
- (6) → (D:8/R:9/E:8/A:8/D:6) = 7.8 (High)
- (7) → (D:8/R:8/E:7/A:7/D:5) = 7 (High)
- (8) → (D:8/R:8/E:8/A:8/D:6) = 7.6 (High)

The detailed scores for this final type of attack are provided in Table A25, with their averages visualized in the chart below.

The DREAD analysis reveals that Poisoning, Trojan, and Backdoor attacks generally cause significant damage—often surpassing other attack types like Jailbreak or Evasion. Their reproducibility, exploitability, and the number of affected users vary depending on the specific attack but typically range from medium to high. However, these attacks are notably difficult to detect, making them particularly dangerous to LLMs. Addressing these vulnerabilities presents a significant challenge for system administrators.

 doi:10.20944/preprints202412.2419.v1






6.7.2. With CVSS

The second assessment of these attacks is done with CVSS [114]. The corresponding Vectors are:

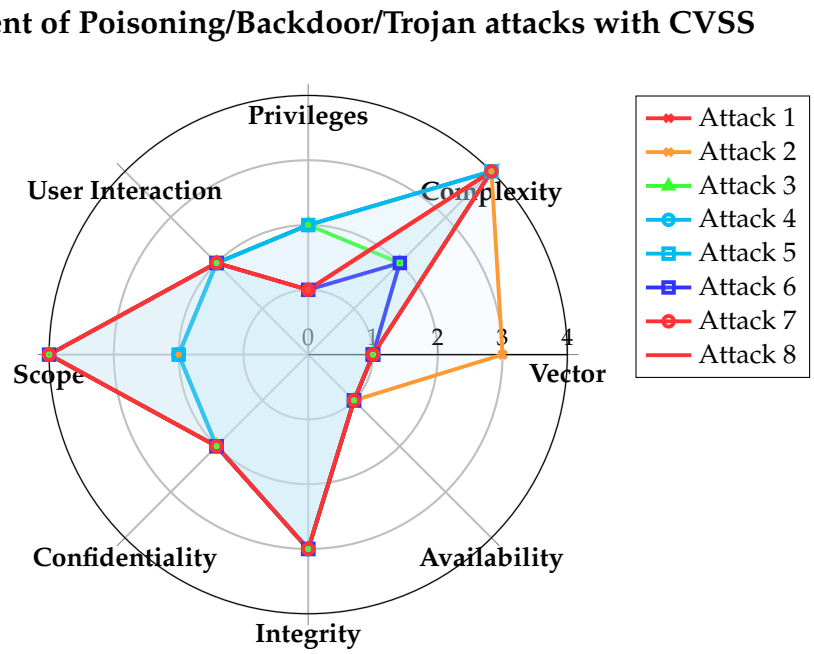
- (1) → (AV:N/AC:H/PR:N/UI:N/S:C/C:L/I:H/A:N) = 7.5 (High)
- (2) → (AV:L/AC:H/PR:L/UI:N/S:U/C:L/I:H/A:N) = 5.3 (Medium)
- (3) → (AV:N/AC:L/PR:L/UI:N/S:C/C:L/I:H/A:N) = 8.5 (High)
- (4) → (AV:N/AC:H/PR:L/UI:N/S:U/C:L/I:H/A:N) = 5.9 (Medium)
- (5) → (AV:N/AC:H/PR:L/UI:N/S:U/C:L/I:H/A:N) = 5.9 (Medium)
- (6) → (AV:N/AC:L/PR:N/UI:N/S:C/C:L/I:H/A:N) = 9.3 (Critical)
- (7) → (AV:N/AC:H/PR:N/UI:N/S:C/C:L/I:H/A:N) = 7.5 (High)
- (8) → (AV:N/AC:H/PR:N/UI:N/S:C/C:L/I:H/A:N) = 7.5 (High)

These vectors are presented below in a Spider-chart for better visualization, the detailed scores are shown in Table A26.

The CVSS assessment reveals that these attacks are generally complex to execute but predominantly impact two technical domains: they exert a high impact on integrity and a medium impact on confidentiality. Furthermore, their execution often does not require elevated privileges, depending on the attack's complexity. This highlights the significant potential danger posed by these types of attacks.

 doi:10.20944/preprints202412.2419.v1






6.7.3. With OWASP Risk Rating

A third evaluation of Poisoning, Trojan, and Backdoor attacks is done using OWASP RR [140]. The corresponding vulnerability vectors of each attack is:

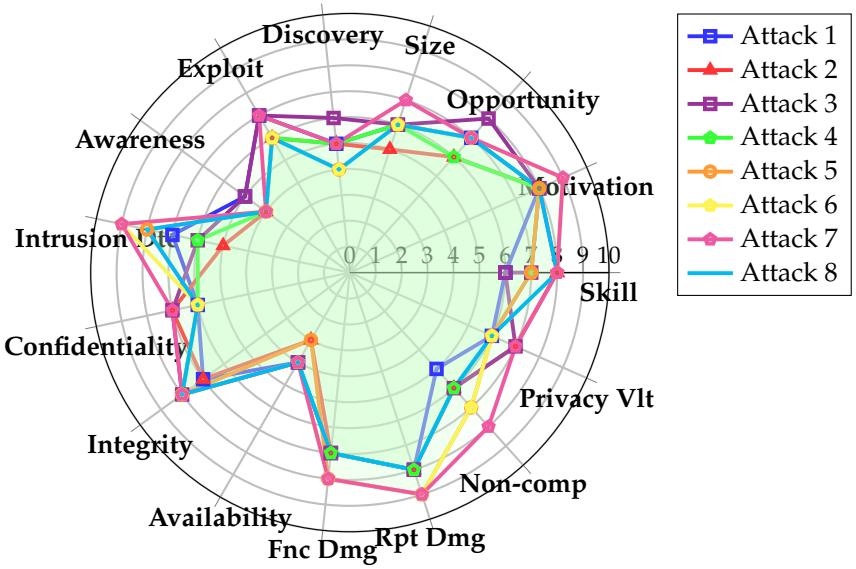
- (1) → (SL:6/M:8/O:7/S:6/ED:5/EE:7/A:5/ID:7/LC:6/LI:7/LA:4/FD:7/RD:8/NC:5/PV:6) = 3.9 (Critical)
- (2) → (SL:8/M:8/O:6/S:5/ED:5/EE:6/A:4/ID:5/LC:7/LI:7/LA:3/FD:7/RD:8/NC:6/PV:7) = 3.7 (High)
- (3) → (SL:7/M:8/O:8/S:6/ED:6/EE:7/A:5/ID:6/LC:7/LI:8/LA:4/FD:7/RD:8/NC:6/PV:7) = 4.4 (Critical)
- (4) → (SL:7/M:8/O:6/S:6/ED:5/EE:6/A:4/ID:6/LC:6/LI:8/LA:4/FD:7/RD:8/NC:6/PV:6) = 4 (Critical)
- (5) → (SL:7/M:8/O:7/S:6/ED:4/EE:6/A:4/ID:8/LC:6/LI:8/LA:3/FD:8/RD:9/NC:7/PV:6) = 4.2 (Critical)
- (6) → (SL:8/M:9/O:7/S:6/ED:4/EE:6/A:4/ID:9/LC:6/LI:8/LA:4/FD:8/RD:9/NC:7/PV:6) = 4 (Critical)
- (7) → (SL:8/M:9/O:7/S:7/ED:5/EE:7/A:4/ID:9/LC:7/LI:8/LA:4/FD:8/RD:9/NC:8/PV:7) = 4.9 (Critical)
- (8) → (SL:8/M:8/O:7/S:6/ED:4/EE:6/A:4/ID:8/LC:6/LI:8/LA:4/FD:7/RD:8/NC:6/PV:6) = 4.1 (Critical)

The calculated values are detailed in Table A27 and visualized in the Spider-chart below. This chart spans a wider area compared to others, emphasizing the significant danger posed by Poisoning, Trojan, and Backdoor attacks. OWASP RR indicates that these attacks require advanced skills and high motivation to execute, unlike simpler attacks such as Model-Inference. Additionally, a pronounced lack of awareness among system administrators complicates their detection. These attacks often have severe impacts on integrity, financial stability, and organizational reputation, as well as contributing to non-compliance and privacy violations, making them some of the most impactful threats in the assessment.

 doi:10.20944/preprints202412.2419.v1



**Assessment of Poisoning/Trojan/Backdoor attacks with OWASP RR**




6.7.4. With SSVC

The final evaluation is performed using SSVC [128] as done with previous attacks. The results of this assessments are detailed below:

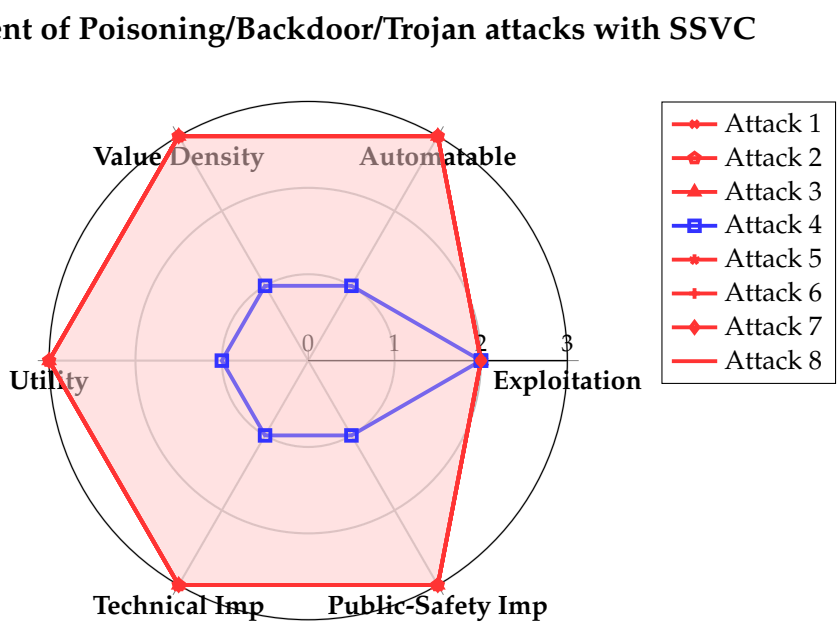
- (1) → (E:P/A:Y/V:C/U:S/T:T/P:S) = Immediate (Very High)
- (2) → (E:P/A:Y/V:C/U:S/T:T/P:S) = Immediate (Very High)
- (3) → (E:P/A:Y/V:C/U:S/T:T/P:S) = Immediate (Very High)
- (4) → (E:P/A:N/V:D/U:L/T:P/P:M) = Scheduled (Medium)
- (5) → (E:P/A:Y/V:C/U:S/T:T/P:S) = Immediate (Very High)
- (6) → (E:P/A:Y/V:C/U:S/T:T/P:S) = Immediate (Very High)
- (7) → (E:P/A:Y/V:C/U:S/T:T/P:S) = Immediate (Very High)
- (8) → (E:P/A:Y/V:C/U:S/T:T/P:S) = Immediate (Very High)

Lastly, the SSVC scores are visualized in the chart below, with detailed assessments provided in Table A28.

The chart aligns closely with those of previous attack types, indicating that these attacks are easily automatable, which increases their exploitation and effectiveness for adversaries. Furthermore, this analysis corroborates findings from other metrics, confirming that Poisoning, Trojan, and Backdoor attacks have a substantial impact on both the technical aspects of a system and the financial stability and safety of its users.

 doi:10.20944/preprints202412.2419.v1






## 7. Discussion

In this section, we present and analyze the results of our assessments of various attacks on large language models using the four vulnerability assessment metrics: DREAD, CVSS, OWASP Risk Rating, and SSVC. By examining the variations in metric values across all evaluated attacks, we aim to identify patterns, inconsistencies, and strengths in each framework. This analysis will provide insights into how effectively these metrics capture the severity and impact of adversarial attacks on LLMs. Ultimately, we will assess the overall utility and reliability of these metrics in evaluating attacks specific to LLMs, offering recommendations on their applicability and potential areas for improvement.

### *7.1. Evaluation of DREAD*

To assess the relevance of the DREAD scoring model, we analyzed the Coefficient of Variation (COV%) for each of its factors, as detailed in Table 14.

Our findings reveal that the factors exhibit varying levels of variability, though most are relatively low. The **Damage** factor shows minimal variation, with COV% below 10% for five of the seven attack classes and slightly higher values for Evasion (13.45%) and Model Inference (10.87%) attacks. Across all 56 attacks, the Damage scores are consistently close, predominantly ranging between 7, 8, and 9 on a scale of 10. Similarly, the **Discoverability** factor demonstrates low variability, with COV% near 10% across all classes, and scores typically falling between 5 and 6. The same pattern is observed for the **Exploitability** and **Affected Users** factors, both of which maintain intra-class COV% around 10%.

This limited variability suggests that these four factors provide insufficient differentiation between adversarial attacks on LLMs. Their inability to distinguish effectively among attack classes renders them **unsuitable** for ranking the relative danger or impact of these attacks.

The **Reproducibility** factor, in contrast, shows greater variability, although inconsistently across attack classes. For example, White-box Jailbreaks and Prompt-Injection attacks exhibit higher COV% values (17.64% and 21.06%, respectively), indicating that attack complexity significantly influences reproducibility. However, this trend is not observed for Black-box Jailbreaks and Evasion attacks, which are generally easier to execute. As a result, while Reproducibility shows potential for distinguishing attacks, its inconsistent variability diminishes its overall utility, especially within **individual** attack-classes.

 doi:10.20944/preprints202412.2419.v1



**Table 14.** Variations of DREAD assessments

**(a)** White-box jailbreak

| N° | D | R | E | A | D |
|---|---|---|---|---|---|
| **1** | 8 (H) | 9 (H) | 8 (H) | 8 (H) | 6 (M) |
| **2** | 6 (M) | 6 (M) | 6 (M) | 6 (M) | 5 (M) |
| **3** | 7 (H) | 7 (H) | 7 (H) | 7 (H) | 5 (M) |
| **4** | 7 (H) | 6 (M) | 5 (M) | 6 (M) | 5 (M) |
| **5** | 8 (H) | 9 (H) | 7 (H) | 7 (H) | 6 (M) |
| **6** | 8 (H) | 8 (H) | 7 (H) | 8 (H) | 6 (M) |
| **7** | 7 (H) | 6 (M) | 7 (H) | 6 (M) | 5 (M) |
| **8** | 8 (H) | 9 (H) | 8 (H) | 7 (H) | 6 (M) |
| $\bar{x}$ | 7.38 | 7.5 | 6.75 | 6.88 | 5.5 |
| $\sigma$ | 0.69 | 1.32 | 0.94 | 0.78 | 0.5 |
| *COV* | **9.44%** | **17.64%** | **14.34%** | **11.35%** | **9.09%** |

**(b)** Black-box jailbreak

| N° | D | R | E | A | D |
|---|---|---|---|---|---|
| **1** | 8 (H) | 7 (H) | 7 (H) | 8 (H) | 5 (M) |
| **2** | 8 (H) | 8 (H) | 8 (H) | 7 (H) | 5 (M) |
| **3** | 8 (H) | 8 (H) | 7 (H) | 7 (H) | 6 (M) |
| **4** | 9 (H) | 8 (H) | 8 (H) | 8 (H) | 6 (M) |
| **5** | 8 (H) | 9 (H) | 8 (H) | 7 (H) | 5 (M) |
| **6** | 8 (H) | 6 (M) | 7 (H) | 7 (H) | 5 (M) |
| **7** | 9 (H) | 8 (H) | 9 (H) | 8 (H) | 5 (M) |
| **8** | 8 (H) | 7 (H) | 7 (H) | 7 (H) | 5 (M) |
| $\bar{x}$ | 8.25 | 7.63 | 7.63 | 7.38 | 5.25 |
| $\sigma$ | 0.43 | 0.85 | 0.69 | 0.48 | 0.43 |
| *COV* | **5.25%** | **11.24%** | **9.13%** | **6.56%** | **8.25%** |

**(c)** Prompt-injection attacks

| N° | D | R | E | A | D |
|---|---|---|---|---|---|
| **1** | 8 (H) | 9 (H) | 8 (H) | 7 (H) | 6 (M) |
| **2** | 8 (H) | 8 (H) | 8 (H) | 7 (H) | 6 (M) |
| **3** | 7 (H) | 9 (H) | 7 (H) | 6 (M) | 7 (H) |
| **4** | 7 (H) | 8 (H) | 8 (H) | 8 (H) | 5 (M) |
| **5** | 8 (H) | 9 (H) | 9 (H) | 8 (H) | 6 (M) |
| **6** | 8 (H) | 8 (H) | 7 (H) | 8 (H) | 5 (M) |
| **7** | 7 (H) | 6 (M) | 6 (M) | 7 (H) | 5 (M) |
| **8** | 7 (H) | 6 (M) | 7 (H) | 6 (M) | 5 (M) |
| $\bar{x}$ | 7.5 | 7.88 | 7.5 | 7.13 | 5.63 |
| $\sigma$ | 0.5 | 1.66 | 0.87 | 0.78 | 0.69 |
| *COV* | **6.67%** | **21.06%** | **11.6%** | **10.9%** | **12.2%** |

**(d)** Evasion attacks

| N° | D | R | E | A | D |
|---|---|---|---|---|---|
| **1** | 7 (H) | 7 (H) | 6 (M) | 7 (H) | 5 (M) |
| **2** | 7 (H) | 9 (H) | 8 (H) | 7 (H) | 5 (M) |
| **3** | 6 (M) | 8 (H) | 6 (M) | 6 (M) | 5 (M) |
| **4** | 8 (H) | 8 (H) | 7 (H) | 7 (H) | 5 (M) |
| **5** | 6 (M) | 9 (H) | 7 (H) | 7 (H) | 6 (M) |
| **6** | 8 (H) | 8 (H) | 8 (H) | 8 (H) | 5 (M) |
| **7** | 9 (H) | 8 (H) | 8 (H) | 8 (H) | 5 (M) |
| **8** | 8 (H) | 8 (H) | 8 (H) | 8 (H) | 5 (M) |
| $\bar{x}$ | 7.38 | 8.13 | 7.25 | 7.25 | 5.13 |
| $\sigma$ | 0.99 | 0.60 | 0.83 | 0.67 | 0.34 |
| *COV* | **13.45%** | **7.38%** | **11.44%** | **9.12%** | **6.45%** |

**(e)** Model-extraction attacks

| N° | D | R | E | A | D |
|---|---|---|---|---|---|
| **1** | 9 (H) | 8 (H) | 8 (H) | 8 (H) | 5 (M) |
| **2** | 8 (H) | 9 (H) | 8 (H) | 7 (H) | 5 (M) |
| **3** | 9 (H) | 8 (H) | 8 (H) | 9 (H) | 6 (M) |
| **4** | 8 (H) | 8 (H) | 7 (H) | 7 (H) | 5 (M) |
| **5** | 8 (H) | 5 (M) | 6 (M) | 8 (H) | 4 (M) |
| **6** | 7 (H) | 7 (H) | 7 (H) | 7 (H) | 6 (M) |
| **7** | 8 (H) | 6 (M) | 7 (H) | 8 (H) | 5 (M) |
| **8** | 7 (H) | 7 (H) | 7 (H) | 6 (M) | 5 (M) |
| $\bar{x}$ | 8 | 7.25 | 7.25 | 7.5 | 5.13 |
| $\sigma$ | 0.71 | 1.09 | 0.67 | 0.75 | 0.60 |
| *COV* | **8.88%** | **15.03%** | **9.12%** | **10.00%** | **11.72%** |

**(f)** Model-inference attacks

| N° | D | R | E | A | D |
|---|---|---|---|---|---|
| **1** | 8 (H) | 9 (H) | 7 (H) | 7 (H) | 6 (M) |
| **2** | 7 (H) | 8 (H) | 7 (H) | 7 (H) | 5 (M) |
| **3** | 6 (M) | 5 (M) | 6 (M) | 6 (M) | 5 (M) |
| **4** | 8 (H) | 6 (M) | 6 (M) | 7 (H) | 5 (M) |
| **5** | 9 (H) | 8 (H) | 8 (H) | 9 (H) | 6 (M) |
| **6** | 7 (H) | 7 (H) | 7 (H) | 7 (H) | 5 (M) |
| **7** | 8 (H) | 7 (H) | 7 (H) | 7 (H) | 5 (M) |
| **8** | 8 (H) | 8 (H) | 7 (H) | 7 (H) | 5 (M) |
| $\bar{x}$ | 7.63 | 7.25 | 6.88 | 7.13 | 5.25 |
| $\sigma$ | 0.83 | 1.09 | 0.64 | 0.83 | 0.47 |
| *COV* | **10.87%** | **15.03%** | **9.30%** | **11.65%** | **8.95%** |

**(g)** Poisoning/Trojan/Backdoor attacks

| N° | D | R | E | A | D |
|---|---|---|---|---|---|
| **1** | 8 (H) | 8 (H) | 8 (H) | 8 (H) | 6 (M) |
| **2** | 8 (H) | 6 (M) | 7 (H) | 7 (H) | 5 (M) |
| **3** | 8 (H) | 8 (H) | 8 (H) | 8 (H) | 6 (M) |
| **4** | 7 (H) | 6 (M) | 6 (M) | 6 (M) | 4 (M) |
| **5** | 9 (H) | 7 (H) | 6 (M) | 9 (H) | 5 (M) |
| **6** | 8 (H) | 9 (H) | 8 (H) | 8 (H) | 6 (M) |
| **7** | 8 (H) | 8 (H) | 7 (H) | 7 (H) | 5 (M) |
| **8** | 8 (H) | 8 (H) | 8 (H) | 8 (H) | 6 (M) |
| $\bar{x}$ | 8 | 7.5 | 7.25 | 7.62 | 5.38 |
| $\sigma$ | 0.5 | 1.00 | 0.83 | 0.86 | 0.70 |
| *COV* | **6.25%** | **13.34%** | **11.45%** | **11.28%** | **13.03%** |

### *7.2. Evaluation of CVSS assessments*

The CVSS framework offers qualitative assessments, making Entropy (H) a more suitable measure of variability than the COV%. The results of this analysis are summarized in Table 15.

 doi:10.20944/preprints202412.2419.v1



**Table 15.** Variations of CVSS assessments

**(a)** White-box jailbreak

| N° | AV | AC | PR | UI | S | C | I | A |
|---|---|---|---|---|---|---|---|---|
| **1** | N | H | N | N | C | L | H | N |
| **2** | N | H | N | N | C | L | H | N |
| **3** | N | H | N | N | C | L | H | N |
| **4** | N | H | L | N | C | L | H | N |
| **5** | N | L | L | N | C | L | H | N |
| **6** | N | L | L | N | C | L | H | N |
| **7** | N | H | N | R | C | L | H | N |
| **8** | N | L | N | R | C | L | H | N |
| *M* | N | H | N | N | C | L | H | N |
| $p_i$ | 1 | 5/8 | 5/8 | 6/8 | 1 | 1 | 1 | 1 |
| *H* | **0.00** | **0.95** | **0.95** | **0.81** | **0.00** | **0.00** | **0.00** | **0.00** |

**(b)** Black-box jailbreak

| N° | AV | AC | PR | UI | S | C | I | A |
|---|---|---|---|---|---|---|---|---|
| **1** | N | L | N | N | U | L | L | N |
| **2** | N | L | N | N | U | L | H | N |
| **3** | N | L | N | N | U | L | L | N |
| **4** | N | L | N | N | C | L | N | N |
| **5** | N | L | N | N | C | L | N | N |
| **6** | N | H | N | N | U | L | L | N |
| **7** | N | L | L | N | U | L | H | N |
| **8** | N | L | N | N | U | L | H | N |
| *M* | N | L | N | N | U | L | L,H | N |
| $p_i$ | 1 | 7/8 | 7/8 | 1 | 6/8 | 1 | 3/8 | 1 |
| *H* | **0.00** | **0.54** | **0.54** | **0.00** | **0.81** | **0.00** | **1.56** | **0.00** |

**(c)** Prompt-injection attacks

| N° | AV | AC | PR | UI | S | C | I | A |
|---|---|---|---|---|---|---|---|---|
| **1** | N | L | N | N | U | N | H | N |
| **2** | N | H | N | R | U | L | H | N |
| **3** | N | L | N | N | U | L | H | N |
| **4** | N | H | N | R | C | L | H | N |
| **5** | N | L | N | N | C | L | H | N |
| **6** | N | H | N | R | C | L | H | N |
| **7** | N | H | N | R | U | L | H | N |
| **8** | N | H | N | N | U | L | H | N |
| *M* | N | H | N | N,R | U | L | H | N |
| $p_i$ | 1 | 5/8 | 1 | 4/8 | 5/8 | 7/8 | 1 | 1 |
| *H* | **0.00** | **0.95** | **0.00** | **1.00** | **0.95** | **0.54** | **0.00** | **0.00** |

**(d)** Evasion attacks

| N° | AV | AC | PR | UI | S | C | I | A |
|---|---|---|---|---|---|---|---|---|
| **1** | N | L | N | N | U | N | H | N |
| **2** | N | L | N | N | U | N | H | N |
| **3** | N | H | N | N | U | N | H | N |
| **4** | N | H | N | N | U | N | H | N |
| **5** | N | L | N | N | U | N | H | N |
| **6** | N | L | N | N | U | N | H | N |
| **7** | N | H | N | N | U | N | H | N |
| **8** | N | L | N | N | U | N | H | N |
| *M* | N | L | N | N | U | N | H | N |
| $p_i$ | 1 | 5/8 | 1 | 1 | 1 | 1 | 1 | 1 |
| *H* | **0.00** | **0.95** | **0.00** | **0.00** | **0.00** | **0.00** | **0.00** | **0.00** |

**(e)** Model-extraction attacks

| N° | AV | AC | PR | UI | S | C | I | A |
|---|---|---|---|---|---|---|---|---|
| **1** | N | H | N | N | C | H | N | N |
| **2** | N | L | N | N | C | H | N | N |
| **3** | N | L | N | N | C | H | N | N |
| **4** | N | H | N | N | C | H | N | N |
| **5** | N | H | N | N | C | H | N | N |
| **6** | N | H | N | N | C | H | N | N |
| **7** | N | H | N | N | C | H | N | N |
| **8** | N | L | N | N | C | H | N | N |
| *M* | N | H | N | N | C | H | N | N |
| $p_i$ | 1 | 5/8 | 1 | 1 | 1 | 1 | 1 | 1 |
| *H* | **0.00** | **0.95** | **0.00** | **0.00** | **0.00** | **0.00** | **0.00** | **0.00** |

**(f)** Model-inference attacks

| N° | AV | AC | PR | UI | S | C | I | A |
|---|---|---|---|---|---|---|---|---|
| **1** | N | L | N | N | C | H | N | N |
| **2** | N | L | N | N | C | H | N | N |
| **3** | N | H | N | N | C | H | N | N |
| **4** | N | H | N | N | C | H | N | N |
| **5** | N | L | N | N | C | H | N | N |
| **6** | N | L | N | N | C | H | N | N |
| **7** | N | L | N | N | C | H | N | N |
| **8** | N | L | N | N | C | H | N | N |
| *M* | N | L | N | N | C | H | N | N |
| $p_i$ | 1 | 6/8 | 1 | 1 | 1 | 1 | 1 | 1 |
| *H* | **0.00** | **0.81** | **0.00** | **0.00** | **0.00** | **0.00** | **0.00** | **0.00** |

**(g)** Poisoning/Trojan/Backdoor attacks

| N° | AV | AC | PR | UI | S | C | I | A |
|---|---|---|---|---|---|---|---|---|
| **1** | N | H | N | N | C | L | H | N |
| **2** | L | H | L | N | U | L | H | N |
| **3** | N | L | L | N | C | L | H | N |
| **4** | N | H | L | N | U | L | H | N |
| **5** | N | H | L | N | U | L | H | N |
| **6** | N | L | N | N | C | L | H | N |
| **7** | N | H | N | N | C | L | H | N |
| **8** | N | H | N | N | C | L | H | N |
| *M* | N | H | N,L | N | C | L | H | N |
| $p_i$ | 7/8 | 6/8 | 4/8 | 1 | 5/8 | 1 | 1 | 1 |
| *H* | **0.54** | **0.81** | **1.00** | **0.00** | **0.95** | **0.00** | **0.00** | **0.00** |

Similar to observations from DREAD, several CVSS factors exhibit minimal or no variability across the attacks. For instance, the **Attack Vector** consistently takes the value "Network" for 55 out of the 56 attacks, reflecting the predominance of network-based adversarial attacks targeting online LLMs. This lack of differentiation renders the factor **unsuitable** for assessing the diversity of attack mechanisms against LLMs.

Likewise, the **Privileges Required** and **User Interaction** factors show low variability. The Privileges Required factor is typically "None," except for White-box and Black-box Jailbreak attacks. Similarly, User Interaction is also "None" for most attacks, apart from White-box Jailbreak and Prompt Injection attacks. This suggests these factors are only **relevant to specific types of attacks** but fail to provide meaningful insights across broader categories.

 doi:10.20944/preprints202412.2419.v1



The **Confidentiality**, **Integrity**, and **Availability** (CIA) Impact factors also demonstrate significant limitations. Each type of AAs typically targets a specific aspect of the CIA triad, leaving the other factors unused. For example, **Model Extraction** attacks heavily impact **Confidentiality** while leaving Integrity and Availability unaffected, resulting in **null entropy** for the latter factors. Similarly, attacks such as **Poisoning**, **Trojan**, and **Backdoor** primarily target **Integrity**, leaving Confidentiality and Availability unchanged. While these factors vary across attack types, they remain **static within individual attack categories**, limiting their ability to differentiate attacks at a granular level.

The **Scope** factor follows a similar trend, showing null entropy in four of the seven attack classes (White-box Jailbreak, Evasion, Model Extraction, and Model Inference). Even within its variability, it often remains uniform within a class, such as being consistently "Changed" for all Model Inference attacks or "Unchanged" for all Evasion attacks.

This highlights the limitation of **specific** qualitative-factors in being suitable in some cases and unsuitable in others.

Among the factors, only **Attack Complexity** shows relatively higher entropy, with most attacks presenting two to three different values from the mode. This variability reflects the differing levels of expertise required to execute various attacks, making this factor **appropriate** for assessing attack difficulty. However, it could benefit from further refinement to enhance its precision.

### *7.3. Evaluation of OWASP Risk Rating assessments*

Now we assess the utility of OWAS Risk Rating factors using their Coefficient of Variation calculated in Table 16.

We start with the **Skill Level** factor, its median values for the seven attack classes generally fall between 6 and 7, with variations of less than 10% in most cases. Similar patterns are observed for the **Motivation** and **Opportunity** factors, where Motivation scores are predominantly between 7 and 8, and Opportunity scores range from 6 to 7, both exhibiting very low variability within each class. The same holds true for the **Size of Threat Agent** factor, where the median consistently falls between 5 and 6, with a COV below 9.2% across six of the seven classes.

These consistent results can be attributed to the **shared characteristics** of AAs against LLMs: attackers typically possess medium-to-high skill levels, show strong motivation, have significant opportunities due to the accessibility of LLMs, and represent a medium-sized threat agent, as these attacks are common but often conducted by individuals. This uniformity in attributes leads to **repeated** values across the 56 attacks, limiting the ability of these metrics to **differentiate** between attacks effectively.

A similar trend is observed for the **Ease of Discovery**, **Ease of Exploit**, **Awareness** of defenders, and **Intrusion Detection** Capabilities factors. The median values for these factors remain consistently around 5 and 6 across the seven classes, with COVs between 8% and 9%. This lack of variability within attack types reduces the informativeness of these factors.

The **Confidentiality**, **Integrity**, and **Availability** factors show a similar limitation, as observed with CVSS metrics. Depending on the attack type, at least one of these factors is often **not relevant** and scores minimal values. However, due to OWASP RR's broader scoring scale (values out of 10), these factors exhibit relatively higher COV percentages compared to CVSS, providing slightly more variability.

The **Financial** and **Reputation Damages** factors are critical for assessing attacks on LLMs, given the potential for data breaches and information leaks that can erode customer trust. These factors consistently score medium-to-high values across all attack types. However, their low COV percentages within the same attack type make it challenging to rank attacks fairly based on these criteria.

For **Non-Compliance** and **Privacy Violation**, the results indicate that these factors are relevant **only for specific attack types**, such as Model Extraction, Model Inference, and Poisoning/Trojan/Backdoor attacks. Other types, like White-box Jailbreaks and Model Evasion, exhibit low-to-medium impacts on Non-Compliance, making these factors valuable for specific contexts but less applicable across all attack types.

 doi:10.20944/preprints202412.2419.v1



**Table 16.** Variations of OWASP RR assessments

**(a)** White-box jailbreak

| N° | SL | M | O | S | ED | EE | A | ID | LC | LI | LA | FD | RD | NC | PV |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 7 (H) | 6 (H) | 6 (H) | 6 (H) | 7 (H) | 8 (H) | 5 (M) | 5 (M) | 5 (M) | 7 (H) | 4 (M) | 7 (H) | 8 (H) | 4 (M) | 4 (M) |
| **2** | 7 (H) | 6 (H) | 5 (M) | 5 (M) | 5 (M) | 7 (H) | 5 (M) | 5 (M) | 5 (M) | 6 (H) | 3 (M) | 6 (H) | 7 (H) | 4 (M) | 4 (M) |
| **3** | 7 (H) | 7 (H) | 5 (M) | 6 (H) | 6 (H) | 7 (H) | 5 (M) | 5 (M) | 5 (M) | 7 (H) | 3 (M) | 6 (H) | 7 (H) | 6 (H) | 5 (M) |
| **4** | 5 (M) | 6 (H) | 4 (M) | 3 (M) | 4 (M) | 6 (H) | 5 (M) | 4 (M) | 4 (M) | 6 (H) | 1 (L) | 5 (H) | 6 (H) | 3 (M) | 4 (M) |
| **5** | 6 (H) | 7 (H) | 6 (H) | 5 (M) | 6 (H) | 7 (H) | 6 (H) | 5 (M) | 5 (M) | 7 (H) | 1 (L) | 6 (H) | 7 (H) | 4 (M) | 4 (M) |
| **6** | 7 (H) | 7 (H) | 6 (H) | 6 (H) | 6 (H) | 7 (H) | 5 (M) | 5 (M) | 5 (M) | 7 (H) | 2 (L) | 6 (H) | 7 (H) | 4 (M) | 5 (M) |
| **7** | 6 (H) | 6 (H) | 5 (M) | 4 (M) | 5 (M) | 6 (H) | 5 (M) | 5 (M) | 4 (M) | 6 (H) | 1 (L) | 5 (M) | 6 (H) | 3 (M) | 3 (M) |
| **8** | 7 (H) | 7 (H) | 7 (H) | 5 (M) | 7 (H) | 7 (H) | 6 (H) | 5 (M) | 4 (M) | 7 (H) | 1 (L) | 6 (H) | 7 (H) | 4 (M) | 4 (M) |
| $\bar{x}$ | 6.5 | 6.5 | 5.63 | 5.13 | 6.00 | 6.88 | 5.25 | 4.88 | 4.62 | 6.62 | 2.00 | 5.88 | 6.75 | 4.00 | 4.12 |
| $\sigma$ | 0.71 | 0.50 | 0.75 | 0.78 | 0.89 | 0.60 | 0.43 | 0.33 | 0.48 | 0.48 | 1.12 | 0.60 | 0.66 | 0.87 | 0.60 |
| *COV* | **10.92%** | **7.69%** | **13.34%** | **15.23%** | **14.83%** | **8.72%** | **8.25%** | **6.78%** | **10.47%** | **7.31%** | **55.90%** | **10.20%** | **9.80%** | **21.65%** | **14.53%** |

**(b)** Black-box jailbreak

| N° | SL | M | O | S | ED | EE | A | ID | LC | LI | LA | FD | RD | NC | PV |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 6 (H) | 8 (H) | 8 (H) | 6 (H) | 6 (H) | 7 (H) | 5 (M) | 6 (H) | 8 (H) | 2 (L) | 1 (L) | 6 (H) | 8 (H) | 5 (M) | 7 (H) |
| **2** | 5 (M) | 8 (H) | 8 (H) | 6 (H) | 6 (H) | 7 (H) | 6 (H) | 6 (H) | 8 (H) | 1 (L) | 1 (L) | 6 (H) | 8 (H) | 4 (M) | 6 (H) |
| **3** | 4 (M) | 8 (H) | 8 (H) | 6 (H) | 7 (H) | 8 (H) | 7 (H) | 7 (H) | 9 (H) | 1 (L) | 1 (L) | 7 (H) | 8 (H) | 5 (M) | 9 (H) |
| **4** | 7 (H) | 8 (H) | 7 (H) | 5 (M) | 6 (H) | 8 (H) | 6 (H) | 8 (H) | 8 (H) | 1 (L) | 1 (L) | 6 (H) | 8 (H) | 5 (M) | 8 (H) |
| **5** | 6 (H) | 8 (H) | 8 (H) | 5 (M) | 6 (H) | 8 (H) | 5 (M) | 7 (H) | 7 (H) | 1 (L) | 1 (L) | 5 (M) | 7 (H) | 5 (M) | 7 (H) |
| **6** | 6 (H) | 8 (H) | 8 (H) | 6 (H) | 6 (H) | 8 (H) | 6 (H) | 7 (H) | 7 (H) | 2 (L) | 1 (L) | 6 (H) | 7 (H) | 5 (M) | 7 (H) |
| **7** | 6 (H) | 9 (H) | 8 (H) | 5 (M) | 6 (H) | 8 (H) | 5 (M) | 7 (H) | 8 (H) | 3 (M) | 1 (L) | 6 (H) | 9 (H) | 7 (H) | 8 (H) |
| **8** | 6 (H) | 8 (H) | 7 (H) | 5 (M) | 6 (H) | 8 (H) | 5 (M) | 7 (H) | 8 (H) | 3 (M) | 1 (L) | 6 (H) | 8 (H) | 5 (M) | 8 (H) |
| $\bar{x}$ | 5.75 | 8.12 | 7.75 | 5.50 | 6.12 | 7.75 | 5.75 | 7.00 | 8.00 | 1.62 | 1.00 | 6.00 | 7.88 | 5.12 | 7.50 |
| $\sigma$ | 0.83 | 0.33 | 0.43 | 0.50 | 0.33 | 0.43 | 0.66 | 0.65 | 0.66 | 0.90 | 0.00 | 0.54 | 0.60 | 0.66 | 0.87 |
| *COV* | **14.42%** | **4.07%** | **5.59%** | **9.09%** | **5.40%** | **5.59%** | **11.48%** | **9.29%** | **8.25%** | **55.56%** | **0.00%** | **9.00%** | **7.63%** | **12.91%** | **11.60%** |

**(c)** Prompt-injection attacks

| N° | SL | M | O | S | ED | EE | A | ID | LC | LI | LA | FD | RD | NC | PV |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 6 (H) | 8 (H) | 7 (H) | 5 (M) | 6 (H) | 8 (H) | 6 (H) | 6 (H) | 3 (M) | 8 (H) | 3 (M) | 6 (H) | 8 (H) | 4 (M) | 4 (M) |
| **2** | 6 (H) | 8 (H) | 7 (H) | 6 (H) | 6 (H) | 7 (H) | 5 (M) | 7 (H) | 5 (M) | 8 (H) | 3 (M) | 7 (H) | 8 (H) | 4 (M) | 6 (H) |
| **3** | 6 (H) | 8 (H) | 7 (H) | 5 (M) | 6 (H) | 8 (H) | 6 (H) | 6 (H) | 5 (M) | 8 (H) | 3 (M) | 7 (H) | 8 (H) | 4 (M) | 5 (M) |
| **4** | 6 (H) | 8 (H) | 7 (H) | 6 (H) | 6 (H) | 8 (H) | 6 (H) | 7 (H) | 6 (H) | 7 (H) | 3 (M) | 7 (H) | 8 (H) | 5 (M) | 7 (H) |
| **5** | 5 (M) | 8 (H) | 7 (H) | 6 (H) | 6 (H) | 8 (H) | 6 (H) | 7 (H) | 6 (H) | 7 (H) | 3 (M) | 7 (H) | 8 (H) | 5 (M) | 6 (H) |
| **6** | 7 (H) | 8 (H) | 8 (H) | 6 (H) | 5 (M) | 8 (H) | 5 (M) | 8 (H) | 7 (H) | 9 (H) | 3 (M) | 7 (H) | 8 (H) | 5 (M) | 7 (H) |
| **7** | 6 (H) | 8 (H) | 7 (H) | 6 (H) | 5 (M) | 7 (H) | 5 (M) | 6 (H) | 4 (M) | 8 (H) | 3 (M) | 6 (H) | 7 (H) | 3 (M) | 5 (M) |
| **8** | 6 (H) | 7 (H) | 6 (H) | 6 (H) | 5 (M) | 6 (H) | 4 (M) | 6 (H) | 3 (M) | 8 (H) | 1 (L) | 5 (M) | 7 (H) | 2 (L) | 3 (M) |
| $\bar{x}$ | 6.12 | 7.88 | 6.25 | 5.75 | 5.62 | 7.62 | 5.50 | 6.75 | 4.75 | 7.88 | 2.62 | 6.50 | 7.75 | 4.00 | 5.50 |
| $\sigma$ | 0.48 | 0.33 | 0.66 | 0.43 | 0.48 | 0.69 | 0.64 | 0.66 | 1.14 | 0.60 | 0.88 | 0.71 | 0.43 | 0.87 | 1.09 |
| *COV* | **7.84%** | **4.18%** | **10.56%** | **7.48%** | **8.54%** | **9.08%** | **11.64%** | **9.78%** | **24.00%** | **7.63%** | **33.59%** | **10.92%** | **5.55%** | **21.75%** | **19.82%** |

**(d)** Evasion attacks

| N° | SL | M | O | S | ED | EE | A | ID | LC | LI | LA | FD | RD | NC | PV |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 7 (H) | 7 (H) | 5 (M) | 4 (M) | 5 (M) | 7 (H) | 5 (M) | 6 (H) | 1 (L) | 8 (H) | 0 (L) | 5 (M) | 7 (H) | 4 (M) | 3 (M) |
| **2** | 6 (H) | 7 (H) | 6 (H) | 5 (M) | 5 (M) | 7 (H) | 5 (M) | 6 (H) | 1 (L) | 8 (H) | 0 (L) | 6 (H) | 7 (H) | 5 (M) | 3 (M) |
| **3** | 6 (M) | 7 (H) | 5 (M) | 5 (M) | 5 (M) | 6 (H) | 4 (M) | 5 (M) | 1 (L) | 7 (H) | 1 (L) | 5 (M) | 6 (H) | 4 (M) | 3 (M) |
| **4** | 7 (H) | 8 (H) | 6 (H) | 5 (M) | 5 (M) | 7 (H) | 5 (M) | 6 (H) | 2 (L) | 8 (H) | 1 (L) | 7 (H) | 8 (H) | 5 (M) | 4 (M) |
| **5** | 6 (H) | 7 (H) | 6 (H) | 5 (M) | 5 (M) | 7 (H) | 5 (M) | 6 (H) | 1 (L) | 7 (H) | 1 (L) | 6 (H) | 7 (H) | 5 (M) | 4 (M) |
| **6** | 7 (H) | 8 (H) | 7 (H) | 5 (M) | 6 (H) | 7 (H) | 6 (H) | 7 (H) | 2 (L) | 9 (H) | 1 (L) | 7 (H) | 8 (H) | 6 (H) | 5 (M) |
| **7** | 7 (H) | 8 (H) | 7 (H) | 5 (M) | 6 (H) | 7 (H) | 6 (H) | 7 (H) | 2 (L) | 9 (H) | 1 (L) | 7 (H) | 8 (H) | 6 (H) | 5 (M) |
| **8** | 7 (H) | 8 (H) | 7 (H) | 6 (H) | 6 (H) | 8 (H) | 6 (H) | 7 (H) | 1 (L) | 8 (H) | 1 (L) | 7 (H) | 8 (H) | 5 (M) | 4 (M) |
| $\bar{x}$ | 6.75 | 7.50 | 6.12 | 5.12 | 5.38 | 7.12 | 5.38 | 6.50 | 1.25 | 8.00 | 0.75 | 6.38 | 7.50 | 5.12 | 4.00 |
| $\sigma$ | 0.47 | 0.50 | 0.69 | 0.47 | 0.48 | 0.48 | 0.64 | 0.64 | 0.62 | 0.75 | 0.62 | 0.75 | 0.66 | 0.66 | 0.83 |
| *COV* | **6.96%** | **6.67%** | **11.29%** | **9.18%** | **8.92%** | **6.76%** | **11.90%** | **9.85%** | **49.60%** | **9.38%** | **82.67%** | **11.76%** | **8.80%** | **12.91%** | **20.75%** |

**(e)** Model-extraction attacks

| N° | SL | M | O | S | ED | EE | A | ID | LC | LI | LA | FD | RD | NC | PV |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 7 (H) | 8 (H) | 7 (H) | 6 (H) | 5 (M) | 6 (H) | 5 (M) | 7 (H) | 8 (H) | 1 (L) | 1 (L) | 7 (H) | 8 (H) | 7 (H) | 8 (H) |
| **2** | 6 (H) | 8 (H) | 7 (H) | 5 (M) | 6 (H) | 7 (H) | 5 (M) | 7 (H) | 8 (H) | 2 (L) | 2 (L) | 7 (H) | 9 (H) | 7 (H) | 8 (H) |
| **3** | 6 (H) | 8 (H) | 7 (H) | 6 (H) | 6 (H) | 6 (H) | 5 (M) | 8 (H) | 9 (H) | 1 (L) | 1 (L) | 7 (H) | 8 (H) | 7 (H) | 9 (H) |
| **4** | 6 (H) | 8 (H) | 7 (H) | 6 (H) | 5 (M) | 7 (H) | 5 (M) | 7 (H) | 8 (H) | 1 (L) | 1 (L) | 7 (H) | 9 (H) | 7 (H) | 9 (H) |
| **5** | 6 (H) | 8 (H) | 7 (H) | 6 (H) | 5 (M) | 6 (H) | 4 (M) | 7 (H) | 8 (H) | 1 (L) | 1 (L) | 7 (H) | 8 (H) | 7 (H) | 9 (H) |
| **6** | 6 (H) | 7 (H) | 7 (H) | 5 (M) | 5 (M) | 6 (H) | 5 (M) | 7 (H) | 8 (H) | 2 (L) | 2 (L) | 7 (H) | 9 (H) | 6 (H) | 8 (H) |
| **7** | 6 (H) | 8 (H) | 7 (H) | 5 (M) | 5 (M) | 6 (H) | 5 (M) | 7 (H) | 9 (H) | 2 (L) | 2 (L) | 7 (H) | 9 (H) | 7 (H) | 9 (H) |
| **8** | 7 (H) | 8 (H) | 7 (H) | 6 (H) | 5 (M) | 6 (H) | 5 (M) | 8 (H) | 8 (H) | 1 (L) | 1 (L) | 7 (H) | 9 (H) | 7 (H) | 9 (H) |
| $\bar{x}$ | 6.38 | 7.88 | 7.00 | 5.75 | 5.25 | 6.38 | 4.88 | 7.38 | 8.25 | 1.25 | 1.25 | 7.00 | 8.62 | 6.88 | 8.62 |
| $\sigma$ | 0.48 | 0.33 | 0.00 | 0.43 | 0.43 | 0.48 | 0.33 | 0.48 | 0.47 | 0.62 | 0.62 | 0.00 | 0.60 | 0.33 | 0.60 |
| *COV* | **7.53%** | **4.18%** | **0.00%** | **7.48%** | **8.20%** | **7.53%** | **6.78%** | **6.50%** | **5.67%** | **49.60%** | **49.60%** | **0.00%** | **6.96%** | **4.80%** | **6.96%** |

 doi:10.20944/preprints202412.2419.v1



**Table 16.** *Cont.*

**(f)** Model-inference attacks

| N° | SL | M | O | S | ED | EE | A | ID | LC | LI | LA | FD | RD | NC | PV |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 6 (H) | 7 (H) | 7 (H) | 6 (H) | 5 (M) | 7 (H) | 6 (H) | 7 (H) | 8 (H) | 1 (L) | 2 (L) | 7 (H) | 8 (H) | 7 (H) | 8 (H) |
| **2** | 6 (H) | 7 (H) | 7 (H) | 6 (H) | 5 (M) | 6 (H) | 4 (M) | 6 (H) | 8 (H) | 2 (L) | 2 (L) | 7 (H) | 8 (H) | 8 (H) | 8 (H) |
| **3** | 5 (M) | 7 (H) | 7 (H) | 5 (M) | 5 (M) | 6 (H) | 5 (M) | 6 (H) | 8 (H) | 1 (L) | 2 (L) | 6 (H) | 7 (H) | 7 (H) | 8 (H) |
| **4** | 6 (H) | 7 (H) | 7 (H) | 6 (H) | 6 (H) | 5 (M) | 6 (H) | 6 (H) | 8 (H) | 1 (L) | 2 (L) | 7 (H) | 7 (H) | 7 (H) | 8 (H) |
| **5** | 6 (H) | 8 (H) | 7 (H) | 6 (H) | 5 (M) | 6 (H) | 6 (H) | 8 (H) | 9 (H) | 1 (L) | 2 (L) | 8 (H) | 8 (H) | 8 (H) | 9 (H) |
| **6** | 6 (H) | 7 (H) | 6 (H) | 6 (H) | 5 (M) | 6 (H) | 5 (M) | 7 (H) | 8 (H) | 2 (L) | 2 (L) | 8 (H) | 8 (H) | 8 (H) | 9 (H) |
| **7** | 6 (H) | 8 (H) | 7 (H) | 6 (H) | 6 (H) | 6 (H) | 5 (M) | 7 (H) | 8 (H) | 2 (L) | 2 (L) | 8 (H) | 8 (H) | 7 (H) | 9 (H) |
| **8** | 5 (M) | 7 (H) | 7 (H) | 6 (H) | 6 (H) | 7 (H) | 6 (H) | 7 (H) | 7 (H) | 2 (L) | 2 (L) | 8 (H) | 8 (H) | 8 (H) | 6 (H) |
| $\bar{x}$ | 5.88 | 7.25 | 6.88 | 6.00 | 5.25 | 6.38 | 5.50 | 7.00 | 8.00 | 1.25 | 2.00 | 7.50 | 7.75 | 7.50 | 8.50 |
| $\sigma$ | 0.48 | 0.47 | 0.33 | 0.45 | 0.43 | 0.48 | 0.64 | 0.65 | 0.50 | 0.62 | 0.00 | 0.66 | 0.43 | 0.50 | 0.64 |
| *COV* | **8.17%** | **6.48%** | **4.80%** | **7.50%** | **8.20%** | **7.53%** | **11.64%** | **9.29%** | **6.25%** | **49.60%** | **0.00%** | **8.80%** | **5.55%** | **6.67%** | **7.53%** |

**(g)** Poisoning/Trojan/Backdoor attacks

| N° | SL | M | O | S | ED | EE | A | ID | LC | LI | LA | FD | RD | NC | PV |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 6 (H) | 8 (H) | 7 (H) | 6 (H) | 5 (M) | 7 (H) | 5 (M) | 7 (H) | 6 (H) | 7 (H) | 4 (M) | 7 (H) | 8 (H) | 5 (M) | 6 (H) |
| **2** | 8 (H) | 8 (H) | 6 (H) | 5 (M) | 5 (M) | 6 (H) | 4 (M) | 5 (M) | 7 (H) | 7 (H) | 3 (M) | 7 (H) | 8 (H) | 6 (H) | 7 (H) |
| **3** | 7 (H) | 8 (H) | 8 (H) | 6 (H) | 6 (H) | 7 (H) | 5 (M) | 6 (H) | 7 (H) | 8 (H) | 4 (M) | 7 (H) | 8 (H) | 6 (H) | 7 (H) |
| **4** | 7 (H) | 8 (H) | 6 (H) | 6 (H) | 5 (M) | 6 (H) | 4 (M) | 6 (H) | 6 (H) | 8 (H) | 4 (M) | 7 (H) | 8 (H) | 6 (H) | 6 (H) |
| **5** | 7 (H) | 8 (H) | 7 (H) | 6 (H) | 4 (M) | 6 (H) | 4 (M) | 8 (H) | 6 (H) | 8 (H) | 3 (M) | 8 (H) | 9 (H) | 7 (H) | 6 (H) |
| **6** | 8 (H) | 9 (H) | 7 (H) | 6 (H) | 4 (M) | 6 (H) | 4 (M) | 9 (H) | 6 (H) | 8 (H) | 4 (M) | 8 (H) | 9 (H) | 7 (H) | 6 (H) |
| **7** | 8 (H) | 9 (H) | 7 (H) | 7 (H) | 5 (M) | 7 (H) | 4 (M) | 9 (H) | 7 (H) | 8 (H) | 4 (M) | 8 (H) | 9 (H) | 8 (H) | 7 (H) |
| **8** | 8 (H) | 8 (H) | 7 (H) | 6 (H) | 4 (M) | 6 (H) | 4 (M) | 8 (H) | 6 (H) | 8 (H) | 4 (M) | 7 (H) | 8 (H) | 6 (H) | 6 (H) |
| $\bar{x}$ | 7.38 | 8.25 | 7.00 | 6.12 | 4.88 | 6.50 | 4.25 | 7.25 | 6.50 | 8.00 | 3.75 | 7.50 | 8.25 | 6.50 | 6.50 |
| $\sigma$ | 0.69 | 0.47 | 0.60 | 0.48 | 0.64 | 0.50 | 0.47 | 1.09 | 0.50 | 0.50 | 0.47 | 0.50 | 0.47 | 0.83 | 0.50 |
| *COV* | **9.35%** | **5.67%** | **8.57%** | **7.84%** | **13.12%** | **7.69%** | **11.06%** | **15.03%** | **7.69%** | **6.25%** | **12.53%** | **6.67%** | **5.67%** | **12.77%** | **7.69%** |

While these OWASP RR factors provide extensive information about each attack, they are not consistently effective in distinguishing between them. However, the broader scoring range (values out of 10) used by OWASP RR does **introduce more variability** compared to CVSS, making it somewhat more adaptable for attack differentiation.

### *7.4. Evaluation of SSVC assessments*

For SSVC, Entropy was calculated to evaluate the variability of its qualitative factors, the results are presented in Table 17.

The **Exploitability** factor, which reflects the existence of an implementation for the attack, demonstrates minimal variability. Among the 56 attacks analyzed, 53 had an associated **Proof-of-Concept**, making this factor largely uniform across the dataset. This lack of differentiation suggests that this factor provides **little valuable information** when assessing adversarial attacks against LLMs.

A similar observation applies to the **Automatable** and **Value-Density** factors. Most adversarial attacks on LLMs are automatable, and they yield significant rewards, such as exposing private or sensitive information from the models. Consequently, these factors also **fail to offer meaningful distinctions** in the scoring process.

The separation of the **Technical** and **Public-Safety Impacts** provides a better understanding of the danger posed by AAs. Although these factors show some degree of variation across attacks within the same category, the differences remain limited. For instance, most attacks are assessed as having "Total" control over the system and "Significant" impacts on finance, reputation, or public health. These assessments, while varying slightly, are overly broad and rely on only two or three possible values, **limiting their utility** for nuanced analysis or differentiation.

 doi:10.20944/preprints202412.2419.v1



**Table 17.** Variations of SSVC assessments

**(a)** White-box jailbreak

| N° | E | A | V | T | P |
|---|---|---|---|---|---|
| **1** | P | Y | C | T | S |
| **2** | P | N | C | T | S |
| **3** | P | N | C | T | S |
| **4** | P | N | C | P | M |
| **5** | P | Y | C | T | S |
| **6** | P | Y | C | T | S |
| **7** | P | N | C | P | M |
| **8** | P | Y | C | T | S |
| $M$ | P | N,Y | C | T | S |
| $p_i$ | 1 | 4/8 | 1 | 6/8 | 6/8 |
| $H$ | **0.00** | **1.00** | **0.00** | **0.81** | **0.81** |

**(b)** Black-box jailbreak

| N° | E | A | V | T | P |
|---|---|---|---|---|---|
| **1** | P | Y | C | T | S |
| **2** | P | Y | D | P | M |
| **3** | P | Y | C | T | S |
| **4** | P | Y | C | T | S |
| **5** | P | Y | C | T | S |
| **6** | P | Y | D | P | M |
| **7** | A | Y | C | T | S |
| **8** | P | Y | C | T | S |
| $M$ | P | Y | C | T | S |
| $p_i$ | 7/8 | 1 | 6/8 | 6/8 | 6/8 |
| $H$ | **0.54** | **0.00** | **0.81** | **0.81** | **0.81** |

**(c)** Prompt-injection attacks

| N° | E | A | V | T | P |
|---|---|---|---|---|---|
| **1** | P | Y | C | T | S |
| **2** | P | Y | C | T | S |
| **3** | P | Y | C | T | S |
| **4** | P | Y | C | T | S |
| **5** | P | Y | C | T | S |
| **6** | P | Y | C | T | S |
| **7** | N | N | D | P | M |
| **8** | P | Y | C | T | S |
| $M$ | P | Y | C | T | S |
| $p_i$ | 7/8 | 7/8 | 7/8 | 7/8 | 7/8 |
| $H$ | **0.54** | **0.54** | **0.54** | **0.54** | **0.54** |

**(d)** Evasion attacks

| N° | E | A | V | T | P |
|---|---|---|---|---|---|
| **1** | P | Y | C | P | S |
| **2** | P | Y | C | P | S |
| **3** | A | Y | C | P | S |
| **4** | P | Y | C | P | S |
| **5** | P | N | D | P | M |
| **6** | P | Y | C | P | S |
| **7** | P | Y | C | P | S |
| **8** | P | Y | C | P | S |
| $M$ | P | Y | C | P | S |
| $p_i$ | 7/8 | 7/8 | 7/8 | 1 | 7/8 |
| $H$ | **0.54** | **0.54** | **0.54** | **0.00** | **0.54** |

**(e)** Model-extraction attacks

| N° | E | A | V | T | P |
|---|---|---|---|---|---|
| **1** | P | Y | C | T | S |
| **2** | P | Y | C | T | S |
| **3** | P | Y | C | T | S |
| **4** | P | Y | C | T | S |
| **5** | P | N | C | T | S |
| **6** | P | Y | C | T | S |
| **7** | P | Y | C | T | S |
| **8** | P | Y | C | T | S |
| $M$ | P | Y | C | T | S |
| $p_i$ | 1 | 7/8 | 1 | 1 | 1 |
| $H$ | **0.00** | **0.54** | **0.00** | **0.00** | **0.00** |

**(f)** Model-inference attacks

| N° | E | A | V | T | P |
|---|---|---|---|---|---|
| **1** | P | Y | C | T | S |
| **2** | P | Y | C | T | S |
| **3** | P | N | D | P | M |
| **4** | P | Y | C | T | S |
| **5** | P | Y | C | T | S |
| **6** | P | Y | D | P | M |
| **7** | P | Y | C | T | S |
| **8** | P | Y | C | T | S |
| $M$ | P | Y | C | T | S |
| $p_i$ | 1 | 7/8 | 6/8 | 6/8 | 6/8 |
| $H$ | **0.00** | **0.54** | **0.81** | **0.81** | **0.81** |

**(g)** Poisoning/Trojan/Backdoor attacks

| N° | E | A | V | T | P |
|---|---|---|---|---|---|
| **1** | P | Y | C | T | S |
| **2** | P | Y | C | T | S |
| **3** | P | Y | C | T | S |
| **4** | P | N | D | P | M |
| **5** | P | Y | C | T | S |
| **6** | P | Y | C | T | S |
| **7** | P | Y | C | T | S |
| **8** | P | Y | C | T | S |
| $M$ | P | Y | C | T | S |
| $p_i$ | 1 | 7/8 | 7/8 | 7/8 | 7/8 |
| $H$ | **0.00** | **0.54** | **0.54** | **0.54** | **0.54** |

 doi:10.20944/preprints202412.2419.v1



## 8. Suggestions for Future Solutions

The analysis conducted in the previous sections validates the hypothesis proposed in Section 1: **existing vulnerability scoring metrics are inadequate for assessing Adversarial Attacks against Large Language Models**. This inadequacy is primarily due to the lack of variability in factor scores, which limits the metrics' ability to distinguish between different types of attacks effectively.

The shortcomings of current vulnerability scoring systems stem from several key issues:

1. **Overemphasis on CIA Impact:** Existing metrics focus heavily on the technical impact on Confidentiality, Integrity, and Availability, which are not the primary targets of AAs against LLMs.
2. **Lack of Contextual Consideration:** Factors such as Attack Vector, Opportunity, and Intrusion Detection lack relevance when applied to LLM-specific scenarios due to the absence of target-specific context.
3. **Subjectivity in Quantitative Scores:** The use of quantitative scoring systems introduces subjectivity, reducing the reliability of assessments.
4. **Limited Qualitative Scoring Options:** Scoring systems with qualitative factors often offer too few choices, resulting in repetitive and non-discriminative assessments.

These limitations highlight the urgent need for the research community to address these gaps and develop scoring metrics specifically tailored for AAs against LLMs, particularly given the increasing adoption of these models in critical applications.

While proposing new metrics is beyond the scope of this study, we suggest the following directions for future research:

1. **Customized Technical Impact Metrics:** Metrics should account for the unique impacts of AAs on LLMs, such as trust erosion, misinformation dissemination, or generating biased and harmful outputs. These factors better reflect the consequences of LLM-specific attacks.
2. **Context-Aware Factors:** Metrics should consider the architecture and nature of the targeted LLM. For example:
    - Larger models (e.g., GPT, LLAMA) are more susceptible to AAs due to complex decision boundaries.
    - Attacks targeting LLMs trained on sensitive personal data pose greater risks than those on public datasets.
    - Multimodal LLMs may face distinct vulnerabilities (e.g., malicious image injection), which text-only models do not encounter.
3. **Incorporating Success Rates:** Success rates could serve as a valuable factor in ranking attacks, although challenging to measure. For instance:
    - Prompt Injection attacks can exhibit varying success rates depending on implementation.
    - Jailbreak attacks may not succeed consistently with a single query but can have cumulative success over multiple attempts, which is important to account for.
4. **Enhanced Qualitative Scoring Systems:** Implementing multiple-choice qualitative factors can strike a balance between complexity and subjectivity. For instance, adding more nuanced levels to factors like "Attack Complexity" (e.g., Minimal, Medium, Very High) could create finer distinctions between attacks and increase score variability.

By exploring these directions, researchers can contribute to the development of robust, context-sensitive metrics that provide meaningful and actionable assessments for adversarial attacks against LLMs. This advancement is crucial for enhancing the security posture of these increasingly prevalent models.

## 9. Conclusion

This study has critically examined the applicability of established vulnerability metrics, such as DREAD, CVSS, OWASP Risk Rating, and SSVC, to assess Adversarial Attacks on LLMs. Through a

 doi:10.20944/preprints202412.2419.v1



detailed analysis of 56 AAs across multiple metrics, the findings demonstrate that existing metrics **fail** to adequately differentiate between attacks, primarily due to their rigid, context-limited factors and a focus on traditional technical impacts rather than the nuanced threats posed by AAs.

Key observations highlight that factors such as technical-impact, the motivation of attackers, and the limited-options of qualitative scoring systems are inadequately addressed in existing frameworks. These limitations restrict the variability and relevance of vulnerability scores, confirming the hypothesis that traditional metrics are not fully suitable for assessing the risks associated with AAs on LLMs.

While the development of new metrics was beyond the scope of this work, the study identifies several promising directions for improvement. These include integrating tailored technical-impact assessments, context-specific factors, and multiple-choice qualitative scoring options to enhance the granularity and applicability of future metrics. Furthermore, incorporating attack success rates, though complex, could provide a more comprehensive evaluation of adversarial threats.

The contributions of this research are multifaceted, providing a taxonomy of adversarial attack classifications, a curated list of 56 AAs targeting LLMs, and an in-depth statistical evaluation of existing vulnerability metrics. These findings not only underscore the limitations of current approaches but also serve as a call to action for the development of more robust, flexible, and LLM-specific vulnerability assessment frameworks.

Future research should focus on refining these metrics to account for the unique challenges posed by Adversarial Attacks on LLMs, ensuring that the security of these increasingly vital systems is both effective and adaptive to emerging threats.

**Acknowledgments:** This work is financially supported by Zayed University under a Research Associate contract

**Conflicts of Interest:** No potential conflict of interest is reported by the authors

## Appendix A. Assessment details

This appendix provides a comprehensive breakdown of the evaluation results for the seven types of AAs assessed using the four vulnerability metrics: DREAD, CVSS, OWASP Risk Rating, and SSVC. Each attack type was evaluated across multiple LLMs, with detailed scores recorded for every attack and metric. The appendix outlines the methodology used to compute average scores by consolidating the evaluations from three distinct LLMs, ensuring an accurate representation of the results.

This detailed presentation of results supports the main text by providing transparency into the scoring process and offering a robust reference for further analysis of the metrics and calculations used in this study.

 doi:10.20944/preprints202412.2419.v1



*Appendix A.1. White-box Jailbreak*

**Table A1.** Detailed assessment of White-box Jailbreak attacks with DREAD

| N° | LLM | D | R | E | A | D | Score |
|---|---|---|---|---|---|---|---|
| **1** | GPT-4o | 8 (H) | 8 (H) | 7 (H) | 7 (H) | 5 (M) | 7 (H) |
| | LLAMA3 | 8 (H) | 9 (H) | 8 (H) | 8 (H) | 6 (M) | 7.8 (H) |
| | Perplexity | 9 (H) | 9 (H) | 8 (H) | 8 (H) | 6 (M) | 8 (H) |
| | **Avg** | **8 (H)** | **9 (H)** | **8 (H)** | **8 (H)** | **6 (M)** | **7.8 (H)** |
| **2** | GPT-4o | 6 (M) | 5 (M) | 6 (H) | 6 (M) | 5 (M) | 5.4 (M) |
| | LLAMA3 | 5 (M) | 6 (M) | 5 (M) | 5 (M) | 4 (M) | 5 (M) |
| | Perplexity | 8 (H) | 7 (H) | 8 (H) | 7 (H) | 5 (M) | 7 (H) |
| | **Avg** | **6 (M)** | **6 (M)** | **6 (M)** | **6 (M)** | **5 (M)** | **5.8 (M)** |
| **3** | GPT-4o | 6 (M) | 5 (M) | 7 (H) | 6 (M) | 5 (M) | 5.8 (M) |
| | LLAMA3 | 7 (H) | 7 (H) | 6 (M) | 7 (H) | 5 (M) | 6.4 (M) |
| | Perplexity | 9 (H) | 8 (H) | 8 (H) | 8 (H) | 5 (M) | 7.6 (H) |
| | **Avg** | **7 (H)** | **7 (H)** | **7 (H)** | **7 (H)** | **5 (M)** | **6.6 (M)** |
| **4** | GPT-4o | 7 (H) | 5 (M) | 6 (M) | 7 (H) | 5 (M) | 6 (M) |
| | LLAMA3 | 6 (M) | 5 (M) | 4 (M) | 6 (M) | 4 (M) | 5 (M) |
| | Perplexity | 7 (H) | 7 (H) | 6 (M) | 6 (M) | 6 (M) | 6.4 (M) |
| | **Avg** | **7 (H)** | **6 (M)** | **5 (M)** | **6 (M)** | **5 (M)** | **5.8 (M)** |
| **5** | GPT-4o | 8 (H) | 9 (H) | 7 (H) | 7 (H) | 6 (M) | 7.4 (H) |
| | LLAMA3 | 9 (H) | 9 (H) | 8 (H) | 8 (H) | 6 (M) | 8 (H) |
| | Perplexity | 8 (H) | 9 (H) | 7 (H) | 7 (H) | 5 (M) | 7.2 (H) |
| | **Avg** | **8 (H)** | **9 (H)** | **7 (H)** | **7 (H)** | **6 (M)** | **7.4 (H)** |
| **6** | GPT-4o | 8 (H) | 8 (H) | 6 (M) | 7 (H) | 5 (M) | 6.8 (M) |
| | LLAMA3 | 8 (H) | 9 (H) | 8 (H) | 8 (H) | 6 (M) | 7.8 (H) |
| | Perplexity | 9 (H) | 8 (H) | 8 (H) | 8 (H) | 6 (M) | 7.8 (H) |
| | **Avg** | **8 (H)** | **8 (H)** | **7 (H)** | **8 (H)** | **6 (M)** | **7.4 (H)** |
| **7** | GPT-4o | 7 (H) | 6 (M) | 6 (M) | 7 (H) | 6 (M) | 6.6 (M) |
| | LLAMA3 | 6 (M) | 6 (M) | 6 (M) | 6 (M) | 4 (M) | 5.6 (M) |
| | Perplexity | 7 (H) | 6 (M) | 7 (H) | 6 (M) | 5 (M) | 5.6 (M) |
| | **Avg** | **7 (H)** | **6 (M)** | **7 (H)** | **6 (M)** | **5 (M)** | **6.2 (M)** |
| **8** | GPT-4o | 8 (H) | 8 (H) | 7 (H) | 7 (H) | 6 (M) | 7.2 (H) |
| | LLAMA3 | 8 (H) | 9 (H) | 8 (H) | 8 (H) | 6 (M) | 7.8 (H) |
| | Perplexity | 8 (H) | 9 (H) | 8 (H) | 7 (H) | 6 (M) | 7.6 (H) |
| | **Avg** | **8 (H)** | **9 (H)** | **8 (H)** | **7 (H)** | **6 (M)** | **7.6 (H)** |

 doi:10.20944/preprints202412.2419.v1



**Table A2.** Detailed assessment of White-box Jailbreak with CVSS

| N° | LLM | AV | AC | PR | UI | S | C | I | A | Base |
|---|---|---|---|---|---|---|---|---|---|---|
| **1** | GPT-4o | N | H | N | N | C | H | H | N | 8.7 (H) |
| | LLAMA3 | N | L | N | R | C | N | H | N | 7.4 (H) |
| | Perplexity | N | H | N | N | U | N | H | N | 5.9 (M) |
| | **Avg** | **N** | **H** | **N** | **N** | **C** | **L** | **H** | **N** | **7.5 (H)** |
| **2** | GPT-4o | N | H | N | N | C | H | H | N | 8.7 (H) |
| | LLAMA3 | N | H | L | R | C | N | H | N | 6.1 (M) |
| | Perplexity | N | H | N | N | U | N | H | N | 5.9 (M) |
| | **Avg** | **N** | **H** | **N** | **N** | **C** | **L** | **H** | **N** | **7.5 (H)** |
| **3** | GPT-4o | N | H | N | N | C | H | H | N | 8.7 (H) |
| | LLAMA3 | N | L | N | R | C | N | H | N | 7.4 (H) |
| | Perplexity | N | H | N | N | U | N | H | N | 5.9 (M) |
| | **Avg** | **N** | **H** | **N** | **N** | **C** | **L** | **H** | **N** | **7.5 (H)** |
| **4** | GPT-4o | L | H | H | N | C | H | H | N | 7.2 (H) |
| | LLAMA3 | N | L | N | R | C | N | H | N | 7.4 (H) |
| | Perplexity | N | H | N | N | U | N | H | N | 5.9 (M) |
| | **Avg** | **N** | **H** | **L** | **N** | **C** | **L** | **H** | **N** | **7.1 (H)** |
| **5** | GPT-4o | L | L | H | N | C | H | H | N | 7.9 (H) |
| | LLAMA3 | N | L | N | R | C | N | H | N | 7.4 (H) |
| | Perplexity | N | L | N | N | U | N | H | N | 7.5 (H) |
| | **Avg** | **N** | **L** | **L** | **N** | **C** | **L** | **H** | **N** | **8.5 (H)** |
| **6** | GPT-4o | L | L | H | N | C | H | H | N | 7.9 (H) |
| | LLAMA3 | N | L | N | R | C | N | H | N | 7.4 (H) |
| | Perplexity | N | L | N | N | U | N | H | N | 7.5 (H) |
| | **Avg** | **N** | **L** | **L** | **N** | **C** | **L** | **H** | **N** | **8.5 (H)** |
| **7** | GPT-4o | N | H | N | R | C | H | H | N | 8 (H) |
| | LLAMA3 | N | L | N | R | C | N | H | N | 7.4 (H) |
| | Perplexity | N | H | N | N | U | N | H | N | 5.9 (M) |
| | **Avg** | **N** | **H** | **N** | **R** | **C** | **L** | **H** | **N** | **6.9 (M)** |
| **8** | GPT-4o | N | L | N | R | C | H | H | N | 9.3 (C) |
| | LLAMA3 | N | L | N | R | C | N | H | N | 7.4 (H) |
| | Perplexity | N | H | N | N | U | N | H | N | 5.9 (M) |
| | **Avg** | **N** | **L** | **N** | **R** | **C** | **L** | **H** | **N** | **8.2 (H)** |

 doi:10.20944/preprints202412.2419.v1



**Table A3.** Detailed assessment of White-box Jailbreak with OWASP Risk Rating

| N° | LLM | SK | MT | OP | SZ | ED | EE | AW | ID | C | I | A | FD | RD | NC | PV | Score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | GPT-4o | 7 | 7 | 6 | 7 | 6 | 7 | 5 | 6 | 8 | 6 | 6 | 7 | 8 | 6 | 7 | 4.3 (C) |
| | LLAMA3 | 8 | 6 | 8 | 5 | 8 | 9 | 6 | 4 | 0 | 8 | 0 | 6 | 8 | 0 | 0 | 2.1 (H) |
| | Perplexity | 7 | 6 | 5 | 6 | 6 | 7 | 5 | 4 | 8 | 7 | 5 | 7 | 8 | 5 | 6 | 3.8 (H) |
| | **Avg** | **7** | **6** | **6** | **6** | **7** | **8** | **5** | **5** | **5** | **7** | **4** | **7** | **8** | **4** | **4** | **3.6 (H)** |
| 2 | GPT-4o | 7 | 7 | 6 | 7 | 5 | 7 | 5 | 6 | 8 | 6 | 5 | 7 | 8 | 7 | 7 | 4.3 (C) |
| | LLAMA3 | 7 | 5 | 5 | 4 | 6 | 7 | 5 | 5 | 0 | 7 | 0 | 5 | 6 | 0 | 0 | 1.4 (L) |
| | Perplexity | 6 | 7 | 5 | 4 | 5 | 6 | 6 | 3 | 7 | 5 | 4 | 6 | 7 | 5 | 5 | 2.9 (M) |
| | **Avg** | **7** | **6** | **5** | **5** | **5** | **7** | **5** | **5** | **5** | **6** | **3** | **6** | **7** | **4** | **4** | **2.8 (M)** |
| 3 | GPT-4o | 7 | 8 | 6 | 7 | 6 | 7 | 5 | 6 | 8 | 6 | 5 | 7 | 8 | 6 | 7 | 4.4 (C) |
| | LLAMA3 | 8 | 6 | 5 | 5 | 7 | 8 | 5 | 5 | 0 | 8 | 0 | 4 | 6 | 8 | 0 | 2.2 (H) |
| | Perplexity | 7 | 6 | 5 | 6 | 6 | 7 | 5 | 4 | 8 | 6 | 5 | 7 | 8 | 5 | 7 | 3.8 (H) |
| | **Avg** | **7** | **7** | **5** | **6** | **6** | **7** | **5** | **5** | **5** | **7** | **3** | **6** | **7** | **6** | **5** | **3.2 (M)** |
| 4 | GPT-4o | 4 | 8 | 4 | 3 | 3 | 5 | 3 | 3 | 7 | 7 | 0 | 5 | 7 | 5 | 7 | 2.2 (M) |
| | LLAMA3 | 6 | 5 | 5 | 4 | 6 | 7 | 5 | 5 | 0 | 7 | 0 | 5 | 6 | 0 | 0 | 1.4 (L) |
| | Perplexity | 5 | 4 | 3 | 2 | 4 | 5 | 6 | 3 | 6 | 5 | 4 | 5 | 6 | 3 | 4 | 1.9 (M) |
| | **Avg** | **5** | **6** | **4** | **3** | **4** | **6** | **5** | **4** | **4** | **6** | **1** | **5** | **6** | **3** | **4** | **1.9 (M)** |
| 5 | GPT-4o | 5 | 8 | 5 | 5 | 6 | 6 | 5 | 6 | 7 | 7 | 0 | 5 | 7 | 7 | 7 | 3.2 (M) |
| | LLAMA3 | 8 | 6 | 8 | 5 | 8 | 9 | 6 | 4 | 0 | 8 | 0 | 6 | 8 | 0 | 0 | 2.1 (H) |
| | Perplexity | 6 | 7 | 5 | 4 | 5 | 6 | 6 | 4 | 7 | 5 | 4 | 6 | 7 | 5 | 6 | 3.1 (M) |
| | **Avg** | **6** | **7** | **6** | **5** | **6** | **7** | **6** | **5** | **5** | **7** | **1** | **6** | **7** | **4** | **4** | **2.8 (H)** |
| 6 | GPT-4o | 6 | 8 | 5 | 7 | 6 | 7 | 5 | 8 | 7 | 7 | 0 | 5 | 7 | 7 | 7 | 3.6 (H) |
| | LLAMA3 | 7 | 5 | 8 | 5 | 7 | 8 | 5 | 5 | 0 | 8 | 0 | 5 | 6 | 0 | 0 | 1.7 (M) |
| | Perplexity | 7 | 8 | 5 | 6 | 6 | 7 | 5 | 3 | 8 | 6 | 5 | 7 | 8 | 5 | 7 | 3.9 (H) |
| | **Avg** | **7** | **7** | **6** | **6** | **6** | **7** | **5** | **5** | **5** | **7** | **2** | **6** | **7** | **4** | **5** | **3.2 (H)** |
| 7 | GPT-4o | 7 | 8 | 8 | 7 | 6 | 6 | 5 | 8 | 6 | 7 | 0 | 5 | 7 | 7 | 6 | 3.7 (H) |
| | LLAMA3 | 6 | 5 | 5 | 4 | 6 | 7 | 5 | 5 | 0 | 7 | 0 | 5 | 6 | 0 | 0 | 1.4 (L) |
| | Perplexity | 5 | 4 | 3 | 2 | 4 | 5 | 6 | 3 | 5 | 4 | 3 | 4 | 5 | 2 | 3 | 1.5 (M) |
| | **Avg** | **6** | **6** | **5** | **4** | **5** | **6** | **5** | **5** | **4** | **6** | **1** | **5** | **6** | **3** | **3** | **2.1 (M)** |
| 8 | GPT-4o | 7 | 8 | 8 | 7 | 8 | 6 | 6 | 8 | 6 | 7 | 0 | 5 | 7 | 7 | 6 | 3.9 (H) |
| | LLAMA3 | 7 | 6 | 8 | 5 | 8 | 9 | 6 | 4 | 0 | 8 | 0 | 6 | 8 | 0 | 0 | 2.1 (H) |
| | Perplexity | 6 | 7 | 5 | 4 | 5 | 6 | 6 | 3 | 7 | 5 | 4 | 6 | 7 | 5 | 6 | 3 (M) |
| | **Avg** | **7** | **7** | **7** | **5** | **7** | **7** | **6** | **5** | **4** | **7** | **1** | **6** | **7** | **4** | **4** | **3.1 (H)** |

 doi:10.20944/preprints202412.2419.v1



**Table A4.** Detailed assessment of White-box Jailbreak attacks with SSVC

| N° | LLM | E | A | V | U | T | P | Score |
|---|---|---|---|---|---|---|---|---|
| 1 | GPT-4o | P | Y | C | S | T | S | Immediate |
| | LLAMA3 | P | Y | C | S | T | S | Immediate |
| | Perplexity | P | Y | C | S | T | S | Immediate |
| | **Avg** | **P** | **Y** | **C** | **S** | **T** | **S** | **Immediate** |
| 2 | GPT-4o | N | N | D | L | P | M | Defer |
| | LLAMA3 | P | N | C | E | T | S | Immediate |
| | Perplexity | P | Y | C | S | T | S | Immediate |
| | **Avg** | **P** | **N** | **C** | **E** | **T** | **S** | **Immediate** |
| 3 | GPT-4o | P | N | D | L | P | M | Scheduled |
| | LLAMA3 | P | N | C | E | T | S | Immediate |
| | Perplexity | P | Y | C | S | T | S | Immediate |
| | **Avg** | **P** | **N** | **C** | **E** | **T** | **S** | **Immediate** |
| 4 | GPT-4o | P | N | C | E | T | S | Immediate |
| | LLAMA3 | P | N | C | E | P | M | Scheduled |
| | Perplexity | N | N | D | L | P | M | Defer |
| | **Avg** | **P** | **N** | **C** | **E** | **P** | **M** | **Scheduled** |
| 5 | GPT-4o | P | Y | C | S | T | S | Immediate |
| | LLAMA3 | P | Y | C | S | T | S | Immediate |
| | Perplexity | P | Y | C | S | T | S | Immediate |
| | **Avg** | **P** | **Y** | **C** | **S** | **T** | **S** | **Immediate** |
| 6 | GPT-4o | P | Y | C | S | T | S | Immediate |
| | LLAMA3 | P | Y | C | S | T | S | Immediate |
| | Perplexity | P | Y | C | S | T | S | Immediate |
| | **Avg** | **P** | **Y** | **C** | **S** | **T** | **S** | **Immediate** |
| 7 | GPT-4o | P | N | C | E | P | M | Scheduled |
| | LLAMA3 | P | N | C | E | P | M | Scheduled |
| | Perplexity | N | N | D | L | P | M | Defer |
| | **Avg** | **P** | **N** | **C** | **E** | **P** | **M** | **Scheduled** |
| 8 | GPT-4o | P | Y | C | S | T | S | Immediate |
| | LLAMA3 | P | Y | C | S | T | S | Immediate |
| | Perplexity | P | Y | C | S | T | S | Immediate |
| | **Avg** | **P** | **Y** | **C** | **S** | **T** | **S** | **Immediate** |

 doi:10.20944/preprints202412.2419.v1



*Appendix A.2. Black-box Jailbreak*

**Table A5.** Detailed assessment of Black-box Jailbreak attacks with DREAD

| N° | LLM | D | R | E | A | D | Score |
|---|---|---|---|---|---|---|---|
| **1** | GPT-4o | 8 (H) | 7 (H) | 6 (M) | 8 (H) | 5 (M) | 6.8 (M) |
| | LLAMA3 | 9 (H) | 8 (H) | 8 (H) | 9 (H) | 5 (M) | 7.6 (H) |
| | Perplexity | 8 (H) | 7 (H) | 8 (H) | 7 (H) | 5 (M) | 7 (H) |
| | **Avg** | **8 (H)** | **7 (H)** | **7 (H)** | **8 (H)** | **5 (M)** | **7 (H)** |
| **2** | GPT-4o | 8 (H) | 7 (H) | 8 (H) | 7 (H) | 5 (M) | 7 (H) |
| | LLAMA3 | 8 (H) | 8 (H) | 7 (H) | 8 (H) | 5 (M) | 7.2 (H) |
| | Perplexity | 7 (H) | 8 (H) | 9 (H) | 6 (M) | 5 (M) | 7 (H) |
| | **Avg** | **8 (H)** | **8 (H)** | **8 (H)** | **7 (H)** | **5 (M)** | **7.2 (H)** |
| **3** | GPT-4o | 7 (H) | 8 (H) | 7 (H) | 7 (H) | 6 (M) | 7 (H) |
| | LLAMA3 | 8 (H) | 8 (H) | 7 (H) | 8 (H) | 6 (M) | 7.4 (H) |
| | Perplexity | 8 (H) | 7 (H) | 8 (H) | 7 (H) | 6 (M) | 7.2 (H) |
| | **Avg** | **8 (H)** | **8 (H)** | **7 (H)** | **7 (H)** | **6 (M)** | **7.2 (H)** |
| **4** | GPT-4o | 8 (H) | 8 (H) | 8 (H) | 7 (H) | 6 (M) | 7.4 (H) |
| | LLAMA3 | 9 (H) | 9 (H) | 8 (H) | 9 (H) | 6 (M) | 8.2 (H) |
| | Perplexity | 9 (H) | 8 (H) | 8 (H) | 8 (H) | 5 (M) | 7.6 (H) |
| | **Avg** | **9 (H)** | **8 (H)** | **8 (H)** | **8 (H)** | **6 (M)** | **7.8 (H)** |
| **5** | GPT-4o | 7 (H) | 8 (H) | 7 (H) | 7 (H) | 5 (M) | 6.8 (M) |
| | LLAMA3 | 9 (H) | 9 (H) | 8 (H) | 9 (H) | 6 (M) | 8.2 (H) |
| | Perplexity | 7 (H) | 9 (H) | 8 (H) | 6 (H) | 5 (M) | 7 (H) |
| | **Avg** | **8 (H)** | **9 (H)** | **8 (H)** | **7 (H)** | **5 (M)** | **7.4 (H)** |
| **6** | GPT-4o | 7 (H) | 6 (M) | 7 (H) | 6 (M) | 5 (M) | 6.2 (M) |
| | LLAMA3 | 8 (H) | 6 (M) | 6 (M) | 8 (H) | 5 (M) | 6.6 (M) |
| | Perplexity | 8 (H) | 7 (H) | 8 (H) | 7 (H) | 6 (M) | 7.2 (H) |
| | **Avg** | **8 (H)** | **6 (M)** | **7 (H)** | **7 (H)** | **5 (M)** | **6.6 (M)** |
| **7** | GPT-4o | 8 (H) | 8 (H) | 9 (H) | 8 (H) | 4 (M) | 7.4 (H) |
| | LLAMA3 | 9 (H) | 9 (H) | 8 (H) | 9 (H) | 6 (M) | 8.2 (H) |
| | Perplexity | 9 (H) | 8 (H) | 9 (H) | 8 (H) | 5 (M) | 7.8 (H) |
| | **Avg** | **9 (H)** | **8 (H)** | **9 (H)** | **8 (H)** | **5 (M)** | **7.8 (H)** |
| **8** | GPT-4o | 8 (H) | 6 (M) | 7 (H) | 7 (H) | 6 (M) | 6.8 (M) |
| | LLAMA3 | 8 (H) | 8 (H) | 7 (H) | 8 (H) | 5 (M) | 7.2 (H) |
| | Perplexity | 7 (H) | 7 (H) | 8 (H) | 6 (M) | 5 (M) | 6.6 (M) |
| | **Avg** | **8 (H)** | **7 (H)** | **7 (H)** | **7 (H)** | **5 (M)** | **6.8 (M)** |

 doi:10.20944/preprints202412.2419.v1



**Table A6.** Detailed assessment of Black-box Jailbreak with CVSS

| N° | LLM | AV | AC | PR | UI | S | C | I | A | Base |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | GPT-4o | N | L | N | N | U | H | L | N | 8.2 (H) |
| | LLAMA3 | N | L | N | N | C | H | L | N | 8.2 (H) |
| | Perplexity | N | H | N | N | U | N | H | N | 5.9 (M) |
| | **Avg** | **N** | **L** | **N** | **N** | **U** | **L** | **L** | **N** | **6.5 (M)** |
| 2 | GPT-4o | N | L | N | N | U | N | H | N | 7.5 (H) |
| | LLAMA3 | N | L | N | N | C | H | L | N | 8.2 (H) |
| | Perplexity | N | L | N | N | U | N | H | N | 7.5 (H) |
| | **Avg** | **N** | **L** | **N** | **N** | **U** | **L** | **H** | **N** | **8.2 (H)** |
| 3 | GPT-4o | N | L | N | N | U | H | N | N | 7.5 (H) |
| | LLAMA3 | N | L | N | N | C | H | L | N | 8.2 (H) |
| | Perplexity | N | H | N | N | U | N | H | N | 5.9 (M) |
| | **Avg** | **N** | **L** | **N** | **N** | **U** | **L** | **L** | **N** | **6.5 (M)** |
| 4 | GPT-4o | N | H | N | N | C | H | N | N | 6.8 (M) |
| | LLAMA3 | N | L | N | N | C | H | L | N | 8.2 (H) |
| | Perplexity | N | L | N | N | U | N | H | N | 7.5 (H) |
| | **Avg** | **N** | **L** | **N** | **N** | **C** | **L** | **N** | **N** | **7.2 (H)** |
| 5 | GPT-4o | N | L | N | N | C | H | N | N | 8.6 (H) |
| | LLAMA3 | N | H | L | N | C | H | L | N | 6.5 (M) |
| | Perplexity | N | L | N | N | U | N | H | N | 7.5 (H) |
| | **Avg** | **N** | **L** | **N** | **N** | **C** | **L** | **N** | **N** | **7.2 (H)** |
| 6 | GPT-4o | N | L | N | N | U | H | L | N | 8.2 (H) |
| | LLAMA3 | N | H | L | N | C | H | L | N | 6.5 (M) |
| | Perplexity | N | H | N | N | U | N | H | N | 5.9 (M) |
| | **Avg** | **N** | **H** | **N** | **N** | **U** | **L** | **L** | **N** | **5.4 (M)** |
| 7 | GPT-4o | N | L | N | N | U | H | H | N | 9.1 (C) |
| | LLAMA3 | L | L | H | N | C | H | L | N | 6.7 (M) |
| | Perplexity | N | H | N | N | U | N | H | N | 5.9 (M) |
| | **Avg** | **N** | **L** | **L** | **N** | **U** | **L** | **H** | **N** | **7.1 (H)** |
| 8 | GPT-4o | N | L | N | N | U | H | H | N | 9.1 (C) |
| | LLAMA3 | N | L | N | N | C | H | L | N | 8.2 (H) |
| | Perplexity | N | H | N | N | U | N | H | N | 5.9 (M) |
| | **Avg** | **N** | **L** | **N** | **N** | **U** | **L** | **H** | **N** | **8.2 (H)** |

 doi:10.20944/preprints202412.2419.v1



**Table A7.** Detailed assessment of Black-box Jailbreak with OWASP Risk Rating

| N° | LLM | SK | MT | OP | SZ | ED | EE | AW | ID | C | I | A | FD | RD | NC | PV | Score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | GPT-4o | 5 | 8 | 8 | 7 | 5 | 5 | 5 | 3 | 7 | 4 | 0 | 5 | 8 | 5 | 8 | 2.9 (H) |
| | LLAMA3 | 6 | 8 | 9 | 5 | 8 | 9 | 6 | 8 | 8 | 0 | 0 | 4 | 8 | 5 | 8 | 3.3 (H) |
| | Perplexity | 6 | 8 | 7 | 5 | 4 | 7 | 5 | 8 | 9 | 2 | 2 | 8 | 7 | 5 | 6 | 3.6 (H) |
| | **Avg** | **6** | **8** | **8** | **6** | **6** | **7** | **5** | **6** | **8** | **2** | **1** | **6** | **8** | **5** | **7** | **3.3 (H)** |
| 2 | GPT-4o | 3 | 8 | 8 | 7 | 5 | 3 | 5 | 3 | 7 | 0 | 0 | 4 | 8 | 2 | 0 | 1.5 (M) |
| | LLAMA3 | 7 | 9 | 9 | 5 | 9 | 9 | 7 | 9 | 9 | 0 | 0 | 5 | 9 | 6 | 9 | 3.6 (H) |
| | Perplexity | 5 | 8 | 8 | 5 | 5 | 8 | 7 | 5 | 8 | 2 | 2 | 8 | 7 | 5 | 8 | 3.5 (H) |
| | **Avg** | **5** | **8** | **8** | **6** | **6** | **7** | **6** | **6** | **8** | **1** | **1** | **6** | **8** | **4** | **6** | **3 (H)** |
| 3 | GPT-4o | 2 | 9 | 9 | 9 | 9 | 9 | 9 | 3 | 9 | 0 | 0 | 9 | 9 | 6 | 9 | 4.2 (H) |
| | LLAMA3 | 6 | 8 | 9 | 5 | 8 | 8 | 6 | 8 | 8 | 0 | 0 | 4 | 8 | 5 | 8 | 3.2 (H) |
| | Perplexity | 5 | 8 | 7 | 5 | 4 | 8 | 5 | 9 | 9 | 2 | 2 | 8 | 7 | 5 | 9 | 3.7 (H) |
| | **Avg** | **4** | **8** | **8** | **6** | **7** | **8** | **7** | **7** | **9** | **1** | **1** | **7** | **8** | **5** | **9** | **3.8 (H)** |
| 4 | GPT-4o | 7 | 7 | 8 | 6 | 6 | 7 | 6 | 5 | 7 | 0 | 0 | 4 | 7 | 5 | 6 | 2.6 (H) |
| | LLAMA3 | 8 | 9 | 9 | 5 | 9 | 9 | 7 | 9 | 9 | 0 | 0 | 5 | 9 | 6 | 9 | 4.2 (H) |
| | Perplexity | 5 | 8 | 5 | 5 | 4 | 7 | 5 | 9 | 8 | 2 | 3 | 8 | 7 | 5 | 8 | 3.4 (H) |
| | **Avg** | **7** | **8** | **7** | **5** | **6** | **8** | **6** | **8** | **8** | **1** | **1** | **6** | **8** | **5** | **8** | **3.5 (H)** |
| 5 | GPT-4o | 6 | 8 | 7 | 5 | 6 | 7 | 5 | 5 | 6 | 0 | 0 | 4 | 6 | 4 | 5 | 2.1 (H) |
| | LLAMA3 | 7 | 8 | 9 | 5 | 8 | 8 | 6 | 8 | 8 | 0 | 0 | 4 | 8 | 5 | 8 | 3.5 (H) |
| | Perplexity | 5 | 8 | 7 | 5 | 4 | 8 | 5 | 9 | 8 | 2 | 2 | 8 | 7 | 5 | 8 | 3.5 (H) |
| | **Avg** | **6** | **8** | **8** | **5** | **6** | **8** | **5** | **7** | **7** | **1** | **1** | **5** | **7** | **5** | **7** | **3 (H)** |
| 6 | GPT-4o | 6 | 7 | 8 | 7 | 5 | 6 | 5 | 4 | 5 | 5 | 0 | 5 | 7 | 5 | 6 | 2.7 (H) |
| | LLAMA3 | 6 | 8 | 9 | 5 | 8 | 8 | 6 | 8 | 8 | 0 | 0 | 4 | 8 | 5 | 8 | 3.2 (H) |
| | Perplexity | 5 | 8 | 6 | 5 | 4 | 9 | 6 | 8 | 9 | 2 | 2 | 8 | 7 | 5 | 6 | 3.5 (H) |
| | **Avg** | **6** | **8** | **8** | **6** | **6** | **8** | **6** | **7** | **7** | **2** | **1** | **6** | **9** | **7** | **8** | **3.8 (H)** |
| 7 | GPT-4o | 6 | 8 | 8 | 8 | 6 | 7 | 4 | 3 | 7 | 7 | 0 | 5 | 8 | 6 | 7 | 3.5 (H) |
| | LLAMA3 | 7 | 9 | 9 | 5 | 9 | 9 | 7 | 9 | 9 | 0 | 0 | 5 | 9 | 6 | 9 | 3.6 (H) |
| | Perplexity | 5 | 9 | 6 | 3 | 3 | 9 | 4 | 10 | 9 | 2 | 3 | 9 | 9 | 8 | 9 | 4.1 (C) |
| | **Avg** | **6** | **9** | **8** | **5** | **6** | **8** | **7** | **5** | **8** | **3** | **1** | **6** | **9** | **7** | **8** | **3.8 (H)** |
| 8 | GPT-4o | 6 | 8 | 8 | 8 | 6 | 6 | 4 | 3 | 7 | 7 | 0 | 5 | 8 | 6 | 7 | 3.4 (H) |
| | LLAMA3 | 6 | 8 | 9 | 5 | 8 | 8 | 6 | 8 | 8 | 0 | 0 | 4 | 8 | 5 | 8 | 3.2 (H) |
| | Perplexity | 5 | 9 | 3 | 3 | 3 | 9 | 4 | 10 | 8 | 2 | 2 | 8 | 7 | 5 | 9 | 3.2 (M) |
| | **Avg** | **6** | **8** | **7** | **5** | **6** | **8** | **5** | **7** | **8** | **3** | **1** | **6** | **8** | **5** | **8** | **3.4 (H)** |

 doi:10.20944/preprints202412.2419.v1



**Table A8.** Detailed assessment of Black-box Jailbreak attacks with SSVC

| N° | LLM | E | A | V | U | T | P | Score |
|---|---|---|---|---|---|---|---|---|
| 1 | GPT-4o | P | Y | C | S | T | S | Immediate |
| | LLAMA3 | P | N | C | L | T | S | Immediate |
| | Perplexity | P | Y | C | S | T | S | Immediate |
| | **Avg** | **P** | **Y** | **C** | **S** | **T** | **S** | **Immediate** |
| 2 | GPT-4o | P | Y | D | E | P | M | Scheduled |
| | LLAMA3 | P | N | C | L | T | S | Immediate |
| | Perplexity | P | Y | D | E | P | M | Scheduled |
| | **Avg** | **P** | **Y** | **D** | **E** | **P** | **M** | **Scheduled** |
| 3 | GPT-4o | A | Y | D | E | P | M | Out-of-Cycle |
| | LLAMA3 | P | N | C | L | T | S | Immediate |
| | Perplexity | A | Y | C | S | T | S | Immediate |
| | **Avg** | **A** | **Y** | **C** | **S** | **T** | **S** | **Immediate** |
| 4 | GPT-4o | P | Y | C | S | T | S | Immediate |
| | LLAMA3 | A | Y | C | S | T | S | Immediate |
| | Perplexity | P | Y | C | S | T | S | Immediate |
| | **Avg** | **P** | **Y** | **C** | **S** | **T** | **S** | **Immediate** |
| 5 | GPT-4o | P | Y | C | S | T | S | Immediate |
| | LLAMA3 | A | Y | C | S | T | S | Immediate |
| | Perplexity | P | Y | C | S | T | S | Immediate |
| | **Avg** | **P** | **Y** | **C** | **S** | **T** | **S** | **Immediate** |
| 6 | GPT-4o | P | Y | D | E | P | M | Scheduled |
| | LLAMA3 | P | Y | C | S | T | S | Immediate |
| | Perplexity | N | N | D | L | P | M | Defer |
| | **Avg** | **P** | **Y** | **D** | **E** | **P** | **M** | **Scheduled** |
| 7 | GPT-4o | A | Y | C | S | T | S | Immediate |
| | LLAMA3 | P | Y | C | S | T | S | Immediate |
| | Perplexity | A | Y | C | S | T | S | Immediate |
| | **Avg** | **A** | **Y** | **C** | **S** | **T** | **S** | **Immediate** |
| 8 | GPT-4o | P | N | D | L | P | M | Scheduled |
| | LLAMA3 | P | Y | C | S | T | S | Immediate |
| | Perplexity | P | Y | C | S | T | S | Immediate |
| | **Avg** | **P** | **Y** | **C** | **S** | **T** | **S** | **Immediate** |

 doi:10.20944/preprints202412.2419.v1



*Appendix A.3. Prompt Injection*

**Table A9.** Detailed assessment of Prompt Injection attacks with DREAD

| N° | LLM | D | R | E | A | D | Score |
|---|---|---|---|---|---|---|---|
| **1** | GPT-4o | 8 (H) | 9 (H) | 7 (H) | 7 (H) | 6 (M) | 7.8 (H) |
| | LLAMA3 | 8 (H) | 9 (H) | 8 (H) | 7 (H) | 6 (M) | 7.6 (H) |
| | Perplexity | 8 (H) | 9 (H) | 8 (H) | 7 (H) | 6 (M) | 7.6 (H) |
| | **Avg** | **8 (H)** | **9 (H)** | **8 (H)** | **7 (H)** | **6 (M)** | **7.6 (H)** |
| **2** | GPT-4o | 8 (H) | 9 (H) | 8 (H) | 7 (H) | 7 (H) | 7.8 (H) |
| | LLAMA3 | 9 (H) | 8 (H) | 9 (H) | 8 (H) | 5 (M) | 7.8 (H) |
| | Perplexity | 7 (H) | 8 (H) | 7 (H) | 6 (M) | 5 (M) | 6.6 (M) |
| | **Avg** | **8 (H)** | **8 (H)** | **8 (H)** | **7 (H)** | **6 (M)** | **7.4 (H)** |
| **3** | GPT-4o | 8 (H) | 9 (H) | 7 (H) | 7 (H) | 6 (M) | 7.4 (H) |
| | LLAMA3 | 7 (H) | 9 (H) | 8 (H) | 6 (M) | 8 (H) | 7.6 (H) |
| | Perplexity | 6 (M) | 9 (H) | 6 (M) | 5 (M) | 7 (H) | 6.6 (M) |
| | **Avg** | **7 (H)** | **9 (H)** | **7 (H)** | **6 (M)** | **7 (H)** | **7.2 (H)** |
| **4** | GPT-4o | 7 (H) | 9 (H) | 8 (H) | 7 (H) | 6 (M) | 7.4 (H) |
| | LLAMA3 | 8 (H) | 8 (H) | 9 (H) | 9 (H) | 5 (M) | 7.8 (H) |
| | Perplexity | 9 (H) | 7 (H) | 8 (H) | 8 (H) | 5 (M) | 7.4 (H) |
| | **Avg** | **7 (H)** | **8 (H)** | **8 (H)** | **8 (H)** | **5 (M)** | **7.2 (H)** |
| **5** | GPT-4o | 8 (H) | 9 (H) | 8 (H) | 8 (H) | 6 (M) | 7.8 (H) |
| | LLAMA3 | 9 (H) | 8 (H) | 9 (H) | 8 (H) | 5 (M) | 7.8 (H) |
| | Perplexity | 8 (H) | 9 (H) | 9 (H) | 7 (H) | 6 (M) | 7.8 (H) |
| | **Avg** | **8 (H)** | **9 (H)** | **9 (H)** | **8 (H)** | **6 (M)** | **8 (H)** |
| **6** | GPT-4o | 9 (H) | 8 (H) | 7 (H) | 9 (H) | 6 (M) | 7.8 (H) |
| | LLAMA3 | 8 (H) | 8 (H) | 8 (H) | 8 (H) | 6 (M) | 7.6 (H) |
| | Perplexity | 8 (H) | 8 (H) | 7 (H) | 6 (M) | 4 (M) | 6.6 (M) |
| | **Avg** | **8 (H)** | **8 (H)** | **7 (H)** | **8 (H)** | **5 (M)** | **7.2 (H)** |
| **7** | GPT-4o | 8 (H) | 6 (M) | 6 (M) | 7 (H) | 5 (M) | 6.4 (M) |
| | LLAMA3 | 6 (M) | 7 (H) | 6 (M) | 7 (H) | 5 (M) | 6.2 (M) |
| | Perplexity | 7 (H) | 6 (M) | 7 (H) | 7 (H) | 5 (M) | 6.4 (M) |
| | **Avg** | **7 (H)** | **6 (M)** | **6 (M)** | **7 (H)** | **5 (M)** | **6.2 (M)** |
| **8** | GPT-4o | 8 (H) | 6 (M) | 6 (M) | 7 (H) | 6 (M) | 6.6 (M) |
| | LLAMA3 | 8 (H) | 8 (H) | 8 (H) | 7 (H) | 6 (M) | 7.4 (H) |
| | Perplexity | 6 (M) | 5 (M) | 6 (M) | 5 (M) | 4 (M) | 5.2 (H) |
| | **Avg** | **7 (H)** | **6 (M)** | **7 (H)** | **6 (M)** | **5 (M)** | **6.2 (M)** |

 doi:10.20944/preprints202412.2419.v1



**Table A10.** Detailed assessment of Prompt Injection with CVSS

| N° | LLM | AV | AC | PR | UI | S | C | I | A | Base |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | GPT-4o | N | L | N | N | U | N | H | N | 7.5 (H) |
| | LLAMA3 | N | L | N | N | C | H | H | N | 10 (C) |
| | Perplexity | N | L | N | N | U | N | H | N | 7.5 (H) |
| | **Avg** | **N** | **L** | **N** | **N** | **U** | **N** | **H** | **N** | **7.5 (H)** |
| 2 | GPT-4o | N | H | N | R | U | L | H | N | 5.9 (M) |
| | LLAMA3 | L | H | N | R | C | H | H | N | 7.4 (H) |
| | Perplexity | N | H | N | N | U | N | H | N | 5.9 (M) |
| | **Avg** | **N** | **H** | **N** | **R** | **U** | **L** | **H** | **N** | **5.9 (M)** |
| 3 | GPT-4o | N | L | N | N | U | N | H | N | 7.5 (H) |
| | LLAMA3 | N | L | N | N | C | H | H | N | 10 (C) |
| | Perplexity | N | L | N | N | U | N | H | N | 7.5 (H) |
| | **Avg** | **N** | **L** | **N** | **N** | **U** | **L** | **H** | **N** | **8.2 (H)** |
| 4 | GPT-4o | N | L | N | R | C | L | L | N | 6.1 (M) |
| | LLAMA3 | N | H | N | R | C | H | H | N | 8 (H) |
| | Perplexity | N | H | N | N | U | N | H | N | 5.9 (M) |
| | **Avg** | **N** | **H** | **N** | **R** | **C** | **L** | **H** | **N** | **6.9 (M)** |
| 5 | GPT-4o | N | L | N | R | C | L | L | N | 6.1 (M) |
| | LLAMA3 | N | L | N | N | C | H | H | N | 10 (C) |
| | Perplexity | N | L | N | N | U | N | H | N | 7.5 (H) |
| | **Avg** | **N** | **L** | **N** | **N** | **C** | **L** | **H** | **N** | **9.3 (C)** |
| 6 | GPT-4o | N | H | N | R | C | H | H | L | 8.2 (H) |
| | LLAMA3 | N | H | N | R | C | H | H | N | 8 (H) |
| | Perplexity | N | H | N | N | U | N | H | N | 5.9 (M) |
| | **Avg** | **N** | **H** | **N** | **R** | **C** | **L** | **H** | **N** | **6.9 (M)** |
| 7 | GPT-4o | N | H | N | R | U | L | H | N | 5.9 (M) |
| | LLAMA3 | L | H | N | R | C | H | H | N | 7.4 (H) |
| | Perplexity | N | H | N | N | U | N | H | N | 5.9 (M) |
| | **Avg** | **N** | **H** | **N** | **R** | **U** | **L** | **H** | **N** | **5.9 (M)** |
| 8 | GPT-4o | N | H | N | R | U | L | H | N | 5.9 (M) |
| | LLAMA3 | N | H | N | N | C | H | H | N | 8.7 (H) |
| | Perplexity | N | H | N | N | U | N | H | N | 5.9 (M) |
| | **Avg** | **N** | **H** | **N** | **N** | **U** | **L** | **H** | **N** | **6.5 (M)** |

 doi:10.20944/preprints202412.2419.v1



**Table A11.** Detailed assessment of Prompt-Injection with OWASP Risk Rating

| N° | LLM | SK | MT | OP | SZ | ED | EE | AW | ID | C | I | A | FD | RD | NC | PV | Score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | GPT-4o | 6 | 8 | 6 | 5 | 5 | 8 | 6 | 4 | 0 | 7 | 0 | 4 | 7 | 0 | 0 | 2.5 (M) |
| | LLAMA3 | 6 | 8 | 9 | 5 | 8 | 9 | 6 | 8 | 6 | 8 | 5 | 8 | 9 | 6 | 8 | 5.2 (C) |
| | Perplexity | 6 | 8 | 6 | 5 | 4 | 6 | 5 | 7 | 3 | 9 | 5 | 7 | 8 | 6 | 4 | 3.4 (H) |
| | **Avg** | **6** | **8** | **7** | **5** | **6** | **8** | **6** | **6** | **3** | **8** | **3** | **6** | **8** | **4** | **4** | **3.3 (H)** |
| 2 | GPT-4o | 5 | 8 | 6 | 7 | 5 | 5 | 3 | 3 | 0 | 8 | 0 | 4 | 8 | 0 | 0 | 1.5 (M) |
| | LLAMA3 | 8 | 9 | 9 | 6 | 9 | 9 | 7 | 9 | 8 | 9 | 6 | 9 | 9 | 7 | 9 | 6.6 (C) |
| | Perplexity | 4 | 8 | 5 | 5 | 4 | 7 | 5 | 8 | 8 | 7 | 3 | 8 | 7 | 5 | 8 | 3.7 (H) |
| | **Avg** | **6** | **8** | **7** | **6** | **6** | **7** | **5** | **7** | **5** | **8** | **3** | **7** | **8** | **4** | **6** | **3.8 (H)** |
| 3 | GPT-4o | 5 | 8 | 6 | 5 | 6 | 5 | 6 | 3 | 0 | 7 | 0 | 4 | 8 | 0 | 0 | 1.4 (M) |
| | LLAMA3 | 7 | 8 | 9 | 6 | 8 | 9 | 6 | 8 | 7 | 8 | 5 | 8 | 9 | 6 | 8 | 5.4 (C) |
| | Perplexity | 5 | 8 | 5 | 5 | 4 | 7 | 5 | 8 | 7 | 8 | 3 | 8 | 7 | 5 | 8 | 3.8 (H) |
| | **Avg** | **6** | **8** | **7** | **5** | **6** | **8** | **6** | **6** | **5** | **8** | **3** | **7** | **8** | **4** | **5** | **3.7 (H)** |
| 4 | GPT-4o | 4 | 8 | 7 | 6 | 6 | 4 | 6 | 5 | 3 | 3 | 1 | 5 | 7 | 3 | 3 | 2.1 (H) |
| | LLAMA3 | 8 | 9 | 9 | 6 | 9 | 9 | 7 | 9 | 8 | 9 | 6 | 9 | 9 | 7 | 9 | 6.6 (C) |
| | Perplexity | 5 | 8 | 5 | 5 | 4 | 7 | 5 | 8 | 8 | 8 | 3 | 8 | 7 | 5 | 8 | 3.9 (H) |
| | **Avg** | **6** | **8** | **7** | **6** | **6** | **8** | **6** | **7** | **6** | **7** | **3** | **7** | **8** | **5** | **7** | **4.1 (C)** |
| 5 | GPT-4o | 4 | 8 | 7 | 7 | 6 | 4 | 6 | 5 | 3 | 4 | 1 | 4 | 7 | 3 | 3 | 2.2 (H) |
| | LLAMA3 | 7 | 8 | 9 | 6 | 8 | 9 | 6 | 8 | 7 | 8 | 5 | 8 | 9 | 6 | 8 | 5.4 (C) |
| | Perplexity | 5 | 8 | 6 | 5 | 4 | 8 | 5 | 8 | 8 | 8 | 3 | 8 | 7 | 5 | 8 | 4.1 (C) |
| | **Avg** | **5** | **8** | **7** | **6** | **6** | **8** | **6** | **7** | **6** | **7** | **3** | **7** | **8** | **5** | **6** | **3.8 (H)** |
| 6 | GPT-4o | 6 | 8 | 8 | 7 | 4 | 7 | 5 | 8 | 7 | 8 | 2 | 5 | 8 | 3 | 7 | 3.8 (H) |
| | LLAMA3 | 9 | 9 | 8 | 7 | 7 | 8 | 6 | 8 | 6 | 9 | 5 | 8 | 8 | 6 | 7 | 5.5 (C) |
| | Perplexity | 5 | 8 | 7 | 5 | 4 | 8 | 5 | 9 | 8 | 9 | 3 | 8 | 8 | 5 | 8 | 4.4 (C) |
| | **Avg** | **7** | **8** | **8** | **6** | **5** | **8** | **5** | **8** | **7** | **9** | **3** | **7** | **8** | **5** | **7** | **4.4 (C)** |
| 7 | GPT-4o | 5 | 8 | 6 | 7 | 6 | 5 | 3 | 3 | 0 | 7 | 0 | 4 | 8 | 0 | 0 | 1.4 (M) |
| | LLAMA3 | 8 | 8 | 8 | 6 | 6 | 7 | 7 | 5 | 5 | 8 | 5 | 7 | 7 | 5 | 6 | 4.2 (C) |
| | Perplexity | 5 | 8 | 7 | 5 | 4 | 8 | 5 | 9 | 8 | 8 | 3 | 8 | 7 | 5 | 8 | 4.3 (C) |
| | **Avg** | **6** | **8** | **7** | **6** | **5** | **7** | **5** | **6** | **4** | **8** | **3** | **6** | **7** | **3** | **5** | **2.6 (M)** |
| 8 | GPT-4o | 5 | 8 | 6 | 7 | 6 | 5 | 3 | 3 | 0 | 8 | 0 | 4 | 8 | 0 | 0 | 1.7 (M) |
| | LLAMA3 | 8 | 6 | 7 | 5 | 4 | 7 | 3 | 8 | 2 | 8 | 0 | 4 | 6 | 2 | 0 | 2.1 (H) |
| | Perplexity | 5 | 8 | 6 | 5 | 4 | 7 | 5 | 8 | 7 | 8 | 3 | 9 | 7 | 5 | 8 | 3.9 (C) |
| | **Avg** | **6** | **7** | **6** | **6** | **5** | **6** | **4** | **6** | **3** | **8** | **1** | **5** | **7** | **2** | **3** | **1.9 (M)** |

 doi:10.20944/preprints202412.2419.v1



**Table A12.** Detailed assessment of Prompt-Injection attacks with SSVC

| N° | LLM | E | A | V | U | T | P | Score |
|---|---|---|---|---|---|---|---|---|
| 1 | GPT-4o | P | Y | D | E | P | M | Scheduled |
| | LLAMA3 | P | Y | C | S | T | S | Immediate |
| | Perplexity | P | Y | C | S | T | S | Immediate |
| | **Avg** | **P** | **Y** | **C** | **S** | **T** | **S** | **Immediate** |
| 2 | GPT-4o | P | N | D | L | P | M | Scheduled |
| | LLAMA3 | P | Y | C | S | T | S | Immediate |
| | Perplexity | P | Y | C | S | T | S | Immediate |
| | **Avg** | **P** | **Y** | **C** | **S** | **T** | **S** | **Immediate** |
| 3 | GPT-4o | P | Y | C | S | T | S | Immediate |
| | LLAMA3 | P | Y | C | S | T | S | Immediate |
| | Perplexity | P | Y | C | S | T | S | Immediate |
| | **Avg** | **P** | **Y** | **C** | **S** | **T** | **S** | **Immediate** |
| 4 | GPT-4o | P | Y | C | S | T | S | Immediate |
| | LLAMA3 | P | Y | C | S | T | S | Immediate |
| | Perplexity | A | Y | C | S | T | S | Immediate |
| | **Avg** | **P** | **Y** | **C** | **S** | **T** | **S** | **Immediate** |
| 5 | GPT-4o | P | Y | C | S | T | S | Immediate |
| | LLAMA3 | P | Y | C | S | T | S | Immediate |
| | Perplexity | P | Y | C | S | T | S | Immediate |
| | **Avg** | **P** | **Y** | **C** | **S** | **T** | **S** | **Immediate** |
| 6 | GPT-4o | P | Y | C | S | T | S | Immediate |
| | LLAMA3 | P | Y | C | S | T | S | Immediate |
| | Perplexity | P | Y | C | S | T | S | Immediate |
| | **Avg** | **P** | **Y** | **C** | **S** | **T** | **S** | **Immediate** |
| 7 | GPT-4o | P | N | D | L | P | M | Scheduled |
| | LLAMA3 | N | N | D | L | P | M | Defer |
| | Perplexity | N | N | D | L | P | M | Defer |
| | **Avg** | **N** | **N** | **D** | **L** | **P** | **M** | **Defer** |
| 8 | GPT-4o | N | N | D | L | P | M | Defer |
| | LLAMA3 | P | Y | C | S | T | S | Immediate |
| | Perplexity | P | Y | C | S | T | S | Immediate |
| | **Avg** | **P** | **Y** | **C** | **S** | **T** | **S** | **Immediate** |

 doi:10.20944/preprints202412.2419.v1



*Appendix A.4. Evasion attacks*

**Table A13.** Detailed assessment of Evasion attacks with DREAD

| N° | LLM | D | R | E | A | D | Score |
|---|---|---|---|---|---|---|---|
| **1** | GPT-4o | 6 (M) | 8 (H) | 7 (H) | 6 (M) | 5 (M) | 6.4 (M) |
| | LLAMA3 | 7 (H) | 6 (M) | 5 (M) | 7 (H) | 4 (M) | 5.8 (M) |
| | Perplexity | 8 (H) | 8 (H) | 7 (H) | 7 (H) | 6 (M) | 7.2 (H) |
| | **Avg** | **7 (H)** | **7 (H)** | **6 (M)** | **7 (H)** | 5 (M) | **6.4 (M)** |
| **2** | GPT-4o | 7 (H) | 8 (H) | 8 (H) | 7 (H) | 4 (M) | 6.8 (M) |
| | LLAMA3 | 8 (H) | 9 (H) | 7 (H) | 8 (H) | 6 (M) | 7.2 (H) |
| | Perplexity | 7 (H) | 9 (H) | 8 (H) | 6 (M) | 5 (M) | 7 (H) |
| | **Avg** | **7 (H)** | **9 (H)** | **8 (H)** | **7 (H)** | **5 (M)** | **7.2 (H)** |
| **3** | GPT-4o | 5 (M) | 7 (H) | 6 (M) | 6 (M) | 5 (M) | 5.8 (M) |
| | LLAMA3 | 6 (M) | 9 (H) | 5 (M) | 6 (M) | 5 (M) | 6.2 (H) |
| | Perplexity | 6 (M) | 8 (H) | 7 (H) | 5 (M) | 6 (M) | 6.4 (M) |
| | **Avg** | **6 (M)** | **8 (H)** | **6 (M)** | **6 (M)** | **5 (M)** | **6.2 (M)** |
| **4** | GPT-4o | 8 (H) | 7 (H) | 7 (H) | 7 (H) | 5 (M) | 6.8 (M) |
| | LLAMA3 | 8 (H) | 9 (H) | 7 (H) | 8 (H) | 6 (M) | 7.6 (H) |
| | Perplexity | 7 (H) | 7 (H) | 6 (M) | 6 (M) | 5 (M) | 6.2 (M) |
| | **Avg** | **8 (H)** | **8 (H)** | **7 (H)** | **7 (H)** | **5 (M)** | **7 (H)** |
| **5** | GPT-4o | 6 (M) | 9 (H) | 8 (H) | 7 (H) | 4 (M) | 6.8 (M) |
| | LLAMA3 | 6 (M) | 9 (H) | 7 (H) | 8 (H) | 6 (M) | 7.2 (H) |
| | Perplexity | 5 (M) | 9 (H) | 6 (M) | 5 (M) | 7 (H) | 6.4 (M) |
| | **Avg** | **6 (M)** | **9 (H)** | **7 (H)** | **7 (H)** | **6 (M)** | **7 (H)** |
| **6** | GPT-4o | 7 (H) | 8 (H) | 7 (H) | 8 (H) | 4 (M) | 6.8 (M) |
| | LLAMA3 | 9 (H) | 9 (H) | 8 (H) | 9 (H) | 5 (M) | 8 (H) |
| | Perplexity | 8 (H) | 8 (H) | 8 (H) | 7 (H) | 6 (M) | 7.4 (H) |
| | **Avg** | **8 (H)** | **8 (H)** | **8 (H)** | **8 (H)** | **5 (M)** | **7.4 (H)** |
| **7** | GPT-4o | 8 (H) | 7 (H) | 7 (H) | 7 (H) | 5 (M) | 6.8 (M) |
| | LLAMA3 | 9 (H) | 9 (H) | 8 (H) | 9 (H) | 5 (M) | 8 (H) |
| | Perplexity | 9 (H) | 8 (H) | 8 (H) | 8 (H) | 5 (M) | 7.6 (H) |
| | **Avg** | **9 (H)** | **8 (H)** | **8 (H)** | **8 (H)** | **5 (M)** | **7.6 (H)** |
| **8** | GPT-4o | 7 (H) | 9 (H) | 8 (H) | 7 (H) | 4 (M) | 6.8 (M) |
| | LLAMA3 | 9 (H) | 8 (H) | 8 (H) | 8 (H) | 5 (M) | 7.6 (H) |
| | Perplexity | 9 (H) | 9 (H) | 8 (H) | 8 (H) | 5 (M) | 7.8 (H) |
| | **Avg** | **9 (H)** | **9 (H)** | **8 (H)** | **8 (H)** | **5 (M)** | **7.8 (H)** |

 doi:10.20944/preprints202412.2419.v1



**Table A14.** Detailed assessment of Evasion attacks with CVSS

| N° | LLM | AV | AC | PR | UI | S | C | I | A | Base |
|---|---|---|---|---|---|---|---|---|---|---|
| **1** | GPT-4o | L | L | H | N | U | N | H | N | 4.4 (M) |
| | LLAMA3 | N | L | N | N | U | N | H | N | 7.5 (H) |
| | Perplexity | N | L | N | N | U | N | H | N | 7.5 (H) |
| | **Avg** | **N** | **L** | **N** | **N** | **U** | **N** | **H** | **N** | **7.5 (H)** |
| **2** | GPT-4o | L | L | H | N | U | N | H | N | 4.4 (M) |
| | LLAMA3 | N | L | N | N | U | N | H | N | 7.5 (H) |
| | Perplexity | N | L | N | N | U | N | H | N | 7.5 (H) |
| | **Avg** | **N** | **L** | **N** | **N** | **U** | **N** | **H** | **N** | **7.5 (H)** |
| **3** | GPT-4o | L | L | H | N | U | N | H | N | 4.4 (M) |
| | LLAMA3 | N | H | N | N | U | N | H | N | 5.9 (M) |
| | Perplexity | N | H | N | N | U | N | H | N | 5.9 (M) |
| | **Avg** | **N** | **H** | **N** | **N** | **U** | **N** | **H** | **N** | **5.9 (M)** |
| **4** | GPT-4o | N | H | N | N | U | L | H | N | 6.5 (M) |
| | LLAMA3 | L | H | N | N | U | N | H | N | 5.1 (M) |
| | Perplexity | N | H | N | N | U | N | H | N | 5.9 (M) |
| | **Avg** | **N** | **H** | **N** | **N** | **U** | **N** | **H** | **N** | **5.9 (M)** |
| **5** | GPT-4o | L | L | H | N | U | N | H | N | 4.4 (M) |
| | LLAMA3 | N | H | N | N | U | N | H | N | 5.9 (M) |
| | Perplexity | N | L | N | N | U | N | H | N | 7.5 (H) |
| | **Avg** | **N** | **L** | **N** | **N** | **U** | **N** | **H** | **N** | **7.5 (H)** |
| **6** | GPT-4o | N | L | N | N | U | L | H | N | 8.2 (H) |
| | LLAMA3 | N | L | N | N | U | N | H | N | 7.5 (H) |
| | Perplexity | N | L | N | N | U | N | H | N | 7.5 (H) |
| | **Avg** | **N** | **L** | **N** | **N** | **U** | **N** | **H** | **N** | **7.5 (H)** |
| **7** | GPT-4o | N | H | N | N | U | L | H | N | 6.5 (M) |
| | LLAMA3 | N | H | N | N | U | N | H | N | 5.9 (M) |
| | Perplexity | N | H | N | N | U | N | H | N | 5.9 (M) |
| | **Avg** | **N** | **H** | **N** | **N** | **U** | **N** | **H** | **N** | **5.9 (M)** |
| **8** | GPT-4o | L | L | H | N | U | N | H | N | 4.4 (M) |
| | LLAMA3 | N | L | N | N | U | N | H | N | 7.5 (H) |
| | Perplexity | N | H | N | N | U | N | H | N | 5.9 (M) |
| | **Avg** | **N** | **L** | **N** | **N** | **U** | **N** | **H** | **N** | **7.5 (H)** |

 doi:10.20944/preprints202412.2419.v1



**Table A15.** Detailed assessment of Evasion attacks with OWASP Risk Rating

| N° | LLM | SK | MT | OP | SZ | ED | EE | AW | ID | C | I | A | FD | RD | NC | PV | Score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | GPT-4o | 6 | 5 | 4 | 3 | 4 | 7 | 6 | 4 | 0 | 7 | 0 | 3 | 5 | 2 | 0 | 1.2 (M) |
| | LLAMA3 | 8 | 9 | 6 | 5 | 6 | 7 | 5 | 6 | 0 | 8 | 0 | 6 | 7 | 5 | 6 | 2.9 (H) |
| | Perplexity | 6 | 7 | 5 | 5 | 4 | 6 | 5 | 7 | 2 | 8 | 2 | 7 | 8 | 6 | 4 | 2.9 (M) |
| | **Avg** | **7** | **7** | **5** | **4** | **5** | **7** | **5** | **6** | **1** | **8** | **0** | **5** | **7** | **4** | **3** | **2.2 (M)** |
| 2 | GPT-4o | 6 | 5 | 4 | 3 | 4 | 7 | 5 | 4 | 0 | 7 | 0 | 3 | 5 | 2 | 0 | 1.2 (M) |
| | LLAMA3 | 7 | 8 | 7 | 6 | 7 | 8 | 6 | 7 | 0 | 9 | 0 | 7 | 8 | 7 | 6 | 3.5 (H) |
| | Perplexity | 5 | 7 | 6 | 5 | 4 | 6 | 5 | 7 | 2 | 8 | 2 | 7 | 8 | 6 | 4 | 2.9 (M) |
| | **Avg** | **6** | **7** | **6** | **5** | **5** | **7** | **5** | **6** | **1** | **8** | **0** | **6** | **7** | **5** | **3** | **2.4 (M)** |
| 3 | GPT-4o | 6 | 7 | 5 | 6 | 5 | 8 | 4 | 4 | 0 | 5 | 0 | 3 | 5 | 2 | 0 | 1.2 (M) |
| | LLAMA3 | 6 | 7 | 5 | 4 | 5 | 6 | 4 | 5 | 0 | 7 | 0 | 5 | 6 | 4 | 5 | 1.9 (M) |
| | Perplexity | 5 | 6 | 6 | 4 | 4 | 5 | 4 | 7 | 3 | 8 | 3 | 7 | 8 | 6 | 4 | 2.8 (M) |
| | **Avg** | **6** | **7** | **5** | **5** | **5** | **6** | **4** | **5** | **1** | **7** | **1** | **5** | **6** | **4** | **3** | **2 (M)** |
| 4 | GPT-4o | 7 | 8 | 6 | 5 | 6 | 7 | 5 | 5 | 2 | 8 | 1 | 7 | 8 | 5 | 2 | 2.8 (H) |
| | LLAMA3 | 8 | 9 | 6 | 5 | 6 | 7 | 5 | 6 | 0 | 8 | 0 | 6 | 7 | 5 | 6 | 2.8 (H) |
| | Perplexity | 6 | 7 | 6 | 5 | 4 | 6 | 5 | 7 | 3 | 8 | 3 | 7 | 8 | 6 | 4 | 3.1 (M) |
| | **Avg** | **7** | **8** | **6** | **5** | **5** | **7** | **5** | **6** | **2** | **8** | **1** | **7** | **8** | **5** | **4** | **3 (H)** |
| 5 | GPT-4o | 6 | 5 | 4 | 3 | 4 | 7 | 5 | 4 | 0 | 5 | 0 | 3 | 5 | 2 | 0 | 1 (M) |
| | LLAMA3 | 7 | 8 | 7 | 6 | 7 | 8 | 6 | 7 | 0 | 9 | 0 | 7 | 8 | 6 | 7 | 3.5 (H) |
| | Perplexity | 6 | 7 | 6 | 5 | 4 | 6 | 5 | 7 | 3 | 8 | 3 | 7 | 8 | 6 | 4 | 3.1 (M) |
| | **Avg** | **6** | **7** | **6** | **5** | **5** | **7** | **5** | **6** | **1** | **7** | **1** | **6** | **7** | **5** | **4** | **2.5 (M)** |
| 6 | GPT-4o | 6 | 8 | 7 | 5 | 7 | 6 | 6 | 5 | 2 | 9 | 1 | 6 | 7 | 5 | 2 | 2.8 (H) |
| | LLAMA3 | 8 | 9 | 7 | 6 | 8 | 9 | 7 | 8 | 0 | 9 | 0 | 8 | 9 | 7 | 8 | 4.3 (H) |
| | Perplexity | 6 | 8 | 6 | 5 | 4 | 5 | 5 | 7 | 3 | 9 | 3 | 7 | 8 | 6 | 4 | 3.2 (M) |
| | **Avg** | **7** | **8** | **7** | **5** | **6** | **7** | **6** | **7** | **2** | **9** | **1** | **7** | **8** | **6** | **5** | **3.5 (H)** |
| 7 | GPT-4o | 7 | 8 | 7 | 5 | 7 | 6 | 5 | 6 | 2 | 8 | 1 | 6 | 7 | 5 | 2 | 2.5 (H) |
| | LLAMA3 | 8 | 9 | 7 | 6 | 8 | 9 | 7 | 8 | 0 | 9 | 0 | 8 | 9 | 7 | 8 | 4.3 (H) |
| | Perplexity | 7 | 8 | 6 | 5 | 4 | 6 | 5 | 7 | 3 | 9 | 3 | 7 | 8 | 6 | 4 | 3.4 (H) |
| | **Avg** | **7** | **8** | **7** | **5** | **6** | **7** | **6** | **7** | **2** | **9** | **1** | **7** | **8** | **6** | **5** | **3.5 (H)** |
| 8 | GPT-4o | 6 | 8 | 6 | 5 | 5 | 8 | 6 | 4 | 0 | 7 | 0 | 4 | 7 | 0 | 0 | 1.5 (M) |
| | LLAMA3 | 9 | 9 | 8 | 7 | 9 | 9 | 8 | 9 | 0 | 9 | 0 | 9 | 9 | 8 | 9 | 4.9 (H) |
| | Perplexity | 7 | 8 | 6 | 5 | 4 | 6 | 5 | 7 | 3 | 9 | 3 | 7 | 8 | 6 | 4 | 3.4 (H) |
| | **Avg** | **7** | **8** | **7** | **6** | **6** | **8** | **6** | **7** | **1** | **8** | **1** | **7** | **8** | **5** | **4** | **3.2 (H)** |

 doi:10.20944/preprints202412.2419.v1



**Table A16.** Detailed assessment of Evasion attacks with SSVC

| N° | LLM | E | A | V | U | T | P | Score |
|---|---|---|---|---|---|---|---|---|
| 1 | GPT-4o | P | Y | D | E | P | M | Scheduled |
| | LLAMA3 | P | Y | C | S | P | S | Immediate |
| | Perplexity | P | Y | C | S | P | S | Immediate |
| | **Avg** | **P** | **Y** | **C** | **S** | **P** | **S** | **Immediate** |
| 2 | GPT-4o | P | Y | C | S | P | S | Immediate |
| | LLAMA3 | P | Y | C | S | P | S | Immediate |
| | Perplexity | P | Y | C | S | P | S | Immediate |
| | **Avg** | **P** | **Y** | **C** | **S** | **P** | **S** | **Immediate** |
| 3 | GPT-4o | A | Y | D | E | P | M | Out-of-Cycle |
| | LLAMA3 | P | Y | C | S | P | S | Immediate |
| | Perplexity | A | Y | C | S | P | S | Immediate |
| | **Avg** | **A** | **Y** | **C** | **S** | **P** | **S** | **Immediate** |
| 4 | GPT-4o | P | Y | C | S | T | S | Immediate |
| | LLAMA3 | P | Y | C | S | P | S | Immediate |
| | Perplexity | P | Y | C | S | P | S | Immediate |
| | **Avg** | **P** | **Y** | **C** | **S** | **P** | **S** | **Immediate** |
| 5 | GPT-4o | P | N | D | L | P | M | Scheduled |
| | LLAMA3 | P | Y | C | S | P | S | Immediate |
| | Perplexity | N | N | D | L | P | M | Defer |
| | **Avg** | **P** | **N** | **D** | **L** | **P** | **M** | **Scheduled** |
| 6 | GPT-4o | P | Y | C | S | P | S | Immediate |
| | LLAMA3 | P | Y | C | S | P | S | Immediate |
| | Perplexity | P | Y | C | S | P | S | Immediate |
| | **Avg** | **P** | **Y** | **C** | **S** | **P** | **S** | **Immediate** |
| 7 | GPT-4o | P | Y | C | S | P | S | Immediate |
| | LLAMA3 | P | Y | C | S | P | S | Immediate |
| | Perplexity | A | Y | C | S | P | S | Immediate |
| | **Avg** | **P** | **Y** | **C** | **S** | **P** | **S** | **Immediate** |
| 8 | GPT-4o | P | Y | C | S | P | S | Immediate |
| | LLAMA3 | P | Y | C | S | P | S | Immediate |
| | Perplexity | P | Y | C | S | P | S | Immediate |
| | **Avg** | **P** | **Y** | **C** | **S** | **P** | **S** | **Immediate** |

 doi:10.20944/preprints202412.2419.v1



*Appendix A.5. Model Extraction*

**Table A17.** Detailed assessment of Model Extraction with DREAD

| N° | LLM | D | R | E | A | D | Score |
|---|---|---|---|---|---|---|---|
| **1** | GPT-4o | 9 (H) | 8 (H) | 7 (H) | 7 (H) | 4 (M) | 7 (H) |
| | LLAMA3 | 8 (H) | 9 (H) | 9 (H) | 9 (H) | 5 (M) | 8 (H) |
| | Perplexity | 9 (H) | 8 (H) | 8 (H) | 8 (H) | 6 (M) | 7.8 (H) |
| | **Avg** | **9 (H)** | **8 (H)** | **8 (H)** | **8 (H)** | **5 (M)** | **7.6 (H)** |
| **2** | GPT-4o | 7 (H) | 8 (H) | 8 (H) | 6 (M) | 5 (M) | 6.8 (M) |
| | LLAMA3 | 8 (H) | 9 (H) | 7 (H) | 8 (H) | 6 (M) | 7.6 (H) |
| | Perplexity | 8 (H) | 9 (H) | 8 (H) | 7 (H) | 5 (M) | 7.4 (H) |
| | **Avg** | **8 (H)** | **9 (H)** | **8 (H)** | **7 (H)** | **5 (M)** | **7.4 (H)** |
| **3** | GPT-4o | 9 (H) | 7 (H) | 8 (H) | 9 (H) | 5 (M) | 7.6 (H) |
| | LLAMA3 | 9 (H) | 9 (H) | 8 (H) | 9 (H) | 7 (H) | 8.4 (H) |
| | Perplexity | 9 (H) | 9 (H) | 8 (H) | 8 (H) | 6 (M) | 8 (H) |
| | **Avg** | **9 (H)** | **8 (H)** | **8 (H)** | **9 (H)** | **6 (M)** | **8 (H)** |
| **4** | GPT-4o | 7 (H) | 7 (H) | 7 (H) | 6 (M) | 4 (M) | 6.2 (M) |
| | LLAMA3 | 8 (H) | 9 (H) | 7 (H) | 8 (H) | 6 (M) | 7.6 (H) |
| | Perplexity | 8 (H) | 7 (H) | 7 (H) | 7 (H) | 5 (M) | 6.8 (M) |
| | **Avg** | **8 (H)** | **8 (H)** | **7 (H)** | **7 (H)** | **5 (M)** | **7 (H)** |
| **5** | GPT-4o | 8 (H) | 6 (M) | 6 (M) | 8 (H) | 3 (L) | 6.2 (M) |
| | LLAMA3 | 7 (H) | 4 (M) | 5 (M) | 7 (H) | 4 (M) | 5.4 (M) |
| | Perplexity | 9 (H) | 6 (M) | 7 (H) | 8 (H) | 5 (M) | 7 (H) |
| | **Avg** | **8 (H)** | **5 (M)** | **6 (M)** | **8 (H)** | **4 (M)** | **6.2 (M)** |
| **6** | GPT-4o | 7 (H) | 7 (H) | 7 (H) | 7 (H) | 6 (M) | 6.8 (M) |
| | LLAMA3 | 8 (H) | 9 (H) | 7 (H) | 8 (H) | 6 (M) | 7.6 (H) |
| | Perplexity | 7 (H) | 6 (M) | 7 (H) | 6 (M) | 5 (M) | 6.2 (M) |
| | **Avg** | **7 (H)** | **7 (H)** | **7 (H)** | **7 (H)** | **6 (M)** | **6.2 (M)** |
| **7** | GPT-4o | 9 (H) | 6 (M) | 7 (H) | 8 (H) | 5 (M) | 7 (H) |
| | LLAMA3 | 8 (H) | 5 (M) | 6 (M) | 8 (H) | 5 (M) | 6.4 (M) |
| | Perplexity | 8 (H) | 7 (H) | 7 (H) | 7 (H) | 5 (M) | 6.8 (M) |
| | **Avg** | **8 (H)** | **6 (M)** | **7 (H)** | **8 (H)** | **5 (M)** | **6.8 (M)** |
| **8** | GPT-4o | 7 (H) | 8 (H) | 8 (H) | 6 (M) | 5 (M) | 6.8 (M) |
| | LLAMA3 | 7 (H) | 6 (M) | 6 (M) | 7 (H) | 5 (M) | 6.2 (M) |
| | Perplexity | 6 (M) | 6 (M) | 6 (M) | 7 (H) | 5 (M) | 6.2 (M) |
| | **Avg** | **7 (H)** | **7 (H)** | **7 (H)** | **6 (M)** | **5 (M)** | **6.4 (M)** |

 doi:10.20944/preprints202412.2419.v1



**Table A18.** Detailed assessment of Model Extraction attacks with CVSS

| N° | LLM | AV | AC | PR | UI | S | C | I | A | Base |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | GPT-4o | N | H | L | N | C | H | N | N | 6.8 (M) |
| | LLAMA3 | N | H | N | N | C | H | N | N | 6.8 (M) |
| | Perplexity | N | L | N | N | C | H | N | N | 8.6 (H) |
| | **Avg** | **N** | **H** | **N** | **N** | **C** | **H** | **N** | **N** | **6.8 (M)** |
| 2 | GPT-4o | N | H | N | N | C | H | N | N | 6.8 (M) |
| | LLAMA3 | N | L | N | N | C | H | N | N | 8.6 (H) |
| | Perplexity | N | L | N | N | C | H | N | N | 8.6 (H) |
| | **Avg** | **N** | **L** | **N** | **N** | **C** | **H** | **N** | **N** | **8.6 (H)** |
| 3 | GPT-4o | N | H | L | N | C | H | N | N | 6.3 (M) |
| | LLAMA3 | N | L | N | N | C | H | N | N | 8.6 (H) |
| | Perplexity | N | L | N | N | C | H | N | N | 8.6 (H) |
| | **Avg** | **N** | **L** | **N** | **N** | **C** | **H** | **N** | **N** | **8.6 (H)** |
| 4 | GPT-4o | N | H | L | N | C | H | N | N | 6.3 (M) |
| | LLAMA3 | N | L | N | N | C | H | N | N | 8.6 (H) |
| | Perplexity | N | H | N | N | C | H | N | N | 6.8 (M) |
| | **Avg** | **N** | **H** | **N** | **N** | **C** | **H** | **N** | **N** | **6.8 (M)** |
| 5 | GPT-4o | N | H | L | N | C | H | N | N | 6.3 (M) |
| | LLAMA3 | N | H | N | N | C | H | N | N | 6.8 (M) |
| | Perplexity | N | L | N | N | C | H | N | N | 8.6 (H) |
| | **Avg** | **N** | **H** | **N** | **N** | **C** | **H** | **N** | **N** | **6.8 (M)** |
| 6 | GPT-4o | N | H | N | N | C | H | N | N | 6.8 (M) |
| | LLAMA3 | N | L | N | N | C | H | N | N | 8.6 (H) |
| | Perplexity | N | H | N | N | C | H | N | N | 6.8 (M) |
| | **Avg** | **N** | **H** | **N** | **N** | **C** | **H** | **N** | **N** | **6.8 (M)** |
| 7 | GPT-4o | N | H | N | N | C | H | N | N | 6.8 (M) |
| | LLAMA3 | N | H | N | N | C | H | N | N | 6.8 (M) |
| | Perplexity | N | H | N | N | C | H | N | N | 6.8 (M) |
| | **Avg** | **N** | **H** | **N** | **N** | **C** | **H** | **N** | **N** | **6.8 (M)** |
| 8 | GPT-4o | N | H | L | N | C | H | N | N | 6.3 (M) |
| | LLAMA3 | N | L | N | N | C | H | N | N | 8.6 (H) |
| | Perplexity | N | L | N | N | C | H | N | N | 8.6 (H) |
| | **Avg** | **N** | **L** | **N** | **N** | **C** | **H** | **N** | **N** | **8.6 (H)** |

 doi:10.20944/preprints202412.2419.v1



**Table A19.** Detailed assessment of Model Extraction with OWASP Risk Rating

| N° | LLM | SK | MT | OP | SZ | ED | EE | AW | ID | C | I | A | FD | RD | NC | PV | Score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | GPT-4o | 8 | 9 | 8 | 8 | 6 | 6 | 7 | 8 | 8 | 0 | 0 | 8 | 9 | 9 | 8 | 4.5 (H) |
| | LLAMA3 | 6 | 8 | 7 | 5 | 4 | 6 | 3 | 9 | 9 | 2 | 1 | 6 | 8 | 5 | 9 | 3.3 (H) |
| | Perplexity | 6 | 7 | 6 | 5 | 4 | 5 | 4 | 7 | 8 | 2 | 2 | 7 | 8 | 6 | 8 | 3.1 (M) |
| | **Avg** | **7** | **8** | **7** | **6** | **5** | **6** | **5** | **7** | **8** | **1** | **1** | **7** | **8** | **7** | **8** | **3.5 (H)** |
| 2 | GPT-4o | 7 | 9 | 6 | 5 | 7 | 6 | 5 | 8 | 8 | 1 | 1 | 8 | 9 | 8 | 8 | 3.9 (H) |
| | LLAMA3 | 5 | 8 | 8 | 6 | 6 | 8 | 5 | 8 | 9 | 2 | 1 | 7 | 9 | 6 | 9 | 3.7 (H) |
| | Perplexity | 5 | 7 | 6 | 5 | 5 | 6 | 5 | 6 | 8 | 2 | 3 | 7 | 8 | 6 | 8 | 3.2 (M) |
| | **Avg** | **6** | **8** | **7** | **5** | **6** | **7** | **5** | **7** | **8** | **2** | **2** | **7** | **9** | **7** | **8** | **3.8 (H)** |
| 3 | GPT-4o | 7 | 8 | 8 | 8 | 6 | 8 | 7 | 8 | 8 | 0 | 0 | 8 | 9 | 8 | 9 | 4.2 (H) |
| | LLAMA3 | 6 | 8 | 7 | 5 | 7 | 5 | 4 | 9 | 9 | 2 | 1 | 6 | 8 | 5 | 9 | 3.3 (H) |
| | Perplexity | 5 | 7 | 6 | 5 | 4 | 5 | 4 | 7 | 9 | 2 | 3 | 7 | 8 | 7 | 9 | 3.3 (H) |
| | **Avg** | **6** | **8** | **7** | **6** | **6** | **6** | **5** | **8** | **9** | **1** | **1** | **7** | **8** | **7** | **9** | **3.7 (H)** |
| 4 | GPT-4o | 6 | 8 | 8 | 7 | 6 | 7 | 6 | 7 | 8 | 0 | 0 | 8 | 9 | 8 | 9 | 3.9 (H) |
| | LLAMA3 | 7 | 8 | 8 | 6 | 6 | 8 | 5 | 8 | 9 | 2 | 1 | 7 | 9 | 6 | 9 | 4 (H) |
| | Perplexity | 6 | 7 | 6 | 5 | 4 | 5 | 4 | 7 | 8 | 2 | 3 | 7 | 8 | 6 | 8 | 3.2 (M) |
| | **Avg** | **6** | **8** | **7** | **6** | **5** | **7** | **5** | **7** | **8** | **1** | **1** | **7** | **9** | **7** | **9** | **3.6 (H)** |
| 5 | GPT-4o | 7 | 9 | 8 | 8 | 6 | 7 | 6 | 7 | 8 | 0 | 0 | 8 | 9 | 8 | 9 | 4 (H) |
| | LLAMA3 | 5 | 8 | 7 | 5 | 4 | 6 | 3 | 9 | 9 | 2 | 1 | 6 | 8 | 5 | 9 | 3.2 (M) |
| | Perplexity | 6 | 8 | 6 | 6 | 4 | 5 | 4 | 6 | 8 | 2 | 3 | 7 | 8 | 7 | 9 | 3.3 (H) |
| | **Avg** | **6** | **8** | **7** | **6** | **5** | **6** | **4** | **7** | **8** | **1** | **1** | **7** | **8** | **7** | **9** | **3.4 (H)** |
| 6 | GPT-4o | 7 | 8 | 7 | 5 | 6 | 6 | 5 | 7 | 8 | 1 | 1 | 7 | 9 | 7 | 8 | 3.5 (H) |
| | LLAMA3 | 6 | 8 | 8 | 6 | 6 | 8 | 5 | 8 | 9 | 2 | 1 | 7 | 9 | 6 | 9 | 4 (H) |
| | Perplexity | 5 | 6 | 6 | 5 | 3 | 4 | 4 | 7 | 8 | 2 | 3 | 7 | 8 | 6 | 8 | 2.9 (M) |
| | **Avg** | **6** | **7** | **7** | **5** | **5** | **6** | **5** | **7** | **8** | **2** | **2** | **7** | **9** | **6** | **8** | **3.5 (H)** |
| 7 | GPT-4o | 6 | 9 | 7 | 5 | 6 | 6 | 6 | 7 | 9 | 2 | 1 | 8 | 9 | 8 | 9 | 4 (C) |
| | LLAMA3 | 7 | 8 | 8 | 6 | 6 | 8 | 5 | 8 | 9 | 2 | 1 | 7 | 9 | 6 | 9 | 4 (H) |
| | Perplexity | 5 | 7 | 6 | 5 | 4 | 5 | 4 | 7 | 9 | 2 | 3 | 7 | 8 | 7 | 9 | 3.3 (H) |
| | **Avg** | **6** | **8** | **7** | **5** | **5** | **6** | **5** | **7** | **9** | **2** | **2** | **7** | **9** | **7** | **9** | **3.8 (C)** |
| 8 | GPT-4o | 8 | 8 | 8 | 8 | 6 | 6 | 7 | 8 | 8 | 0 | 0 | 8 | 9 | 8 | 8 | 4 (H) |
| | LLAMA3 | 6 | 8 | 8 | 6 | 6 | 8 | 5 | 8 | 9 | 2 | 1 | 7 | 9 | 6 | 9 | 4 (H) |
| | Perplexity | 6 | 8 | 6 | 6 | 4 | 5 | 4 | 7 | 8 | 2 | 3 | 7 | 8 | 6 | 9 | 3.3 (M) |
| | **Avg** | **7** | **8** | **7** | **6** | **5** | **6** | **5** | **8** | **8** | **1** | **1** | **7** | **9** | **7** | **9** | **3.7 (H)** |

 doi:10.20944/preprints202412.2419.v1



**Table A20.** Detailed assessment of Model-Extraction attacks with SSVC

| N° | LLM | E | A | V | U | T | P | Score |
|---|---|---|---|---|---|---|---|---|
| 1 | GPT-4o | P | Y | C | S | T | S | Immediate |
| | LLAMA3 | P | N | C | E | T | S | Immediate |
| | Perplexity | P | Y | C | S | T | S | Immediate |
| | **Avg** | **P** | **Y** | **C** | **S** | **T** | **S** | **Immediate** |
| 2 | GPT-4o | P | Y | C | S | T | S | Immediate |
| | LLAMA3 | P | Y | C | S | T | S | Immediate |
| | Perplexity | P | Y | C | S | T | S | Immediate |
| | **Avg** | **P** | **Y** | **C** | **S** | **T** | **S** | **Immediate** |
| 3 | GPT-4o | P | Y | C | S | T | S | Immediate |
| | LLAMA3 | P | Y | C | S | T | S | Immediate |
| | Perplexity | A | Y | C | S | T | S | Immediate |
| | **Avg** | **P** | **Y** | **C** | **S** | **T** | **S** | **Immediate** |
| 4 | GPT-4o | P | Y | C | S | T | S | Immediate |
| | LLAMA3 | P | Y | C | S | T | S | Immediate |
| | Perplexity | P | Y | C | S | T | S | Immediate |
| | **Avg** | **P** | **Y** | **C** | **S** | **T** | **S** | **Immediate** |
| 5 | GPT-4o | P | Y | C | S | T | S | Immediate |
| | LLAMA3 | P | N | C | E | T | S | Immediate |
| | Perplexity | N | N | D | L | P | M | Defer |
| | **Avg** | **P** | **N** | **C** | **E** | **T** | **S** | **Immediate** |
| 6 | GPT-4o | P | Y | D | E | P | M | Scheduled |
| | LLAMA3 | P | Y | C | S | T | S | Immediate |
| | Perplexity | P | Y | C | S | T | S | Immediate |
| | **Avg** | **P** | **Y** | **C** | **S** | **T** | **S** | **Immediate** |
| 7 | GPT-4o | P | Y | C | S | T | S | Immediate |
| | LLAMA3 | P | N | C | E | T | S | Immediate |
| | Perplexity | A | Y | C | S | T | S | Immediate |
| | **Avg** | **P** | **Y** | **C** | **S** | **T** | **S** | **Immediate** |
| 8 | GPT-4o | P | Y | C | S | T | S | Immediate |
| | LLAMA3 | P | N | C | E | T | S | Immediate |
| | Perplexity | P | Y | C | S | T | S | Immediate |
| | **Avg** | **P** | **Y** | **C** | **S** | **T** | **S** | **Immediate** |

 doi:10.20944/preprints202412.2419.v1



*Appendix A.6. Model Inference*

**Table A21.** Detailed assessment of Model Inference with DREAD

| N° | LLM | D | R | E | A | D | Score |
|---|---|---|---|---|---|---|---|
| **1** | GPT-4o | 7 (H) | 8 (H) | 7 (H) | 7 (H) | 5 (M) | 6.8 (M) |
| | LLAMA3 | 8 (H) | 9 (H) | 7 (H) | 8 (H) | 6 (M) | 7.6 (H) |
| | Perplexity | 8 (H) | 9 (H) | 8 (H) | 7 (H) | 6 (M) | 7.6 (H) |
| | **Avg** | **8 (H)** | **9 (H)** | **7 (H)** | **7 (H)** | **6 (M)** | **7.4 (H)** |
| **2** | GPT-4o | 7 (H) | 8 (H) | 8 (H) | 7 (H) | 4 (M) | 6.8 (M) |
| | LLAMA3 | 7 (H) | 9 (H) | 6 (M) | 9 (H) | 5 (M) | 7.2 (H) |
| | Perplexity | 7 (H) | 8 (H) | 7 (H) | 6 (M) | 5 (M) | 6.6 (M) |
| | **Avg** | **7 (H)** | **8 (H)** | **7 (H)** | **7 (H)** | **5 (M)** | **6.8 (M)** |
| **3** | GPT-4o | 7 (H) | 6 (M) | 7 (H) | 7 (H) | 6 (M) | 6.6 (M) |
| | LLAMA3 | 6 (M) | 4 (M) | 5 (M) | 6 (M) | 4 (M) | 5 (M) |
| | Perplexity | 6 (M) | 5 (M) | 6 (M) | 5 (M) | 6 (M) | 5.6 (M) |
| | **Avg** | **6 (M)** | **5 (M)** | **6 (M)** | **6 (M)** | **5 (M)** | **5.6 (M)** |
| **4** | GPT-4o | 8 (H) | 6 (M) | 6 (M) | 8 (H) | 5 (M) | 6.6 (M) |
| | LLAMA3 | 8 (H) | 7 (H) | 5 (M) | 8 (H) | 5 (M) | 6.6 (M) |
| | Perplexity | 7 (H) | 6 (M) | 7 (H) | 6 (M) | 5 (M) | 6.2 (M) |
| | **Avg** | **8 (H)** | **6 (M)** | **6 (M)** | **7 (H)** | **5 (M)** | **6.4 (M)** |
| **5** | GPT-4o | 9 (H) | 7 (H) | 8 (H) | 9 (H) | 5 (M) | 7.6 (H) |
| | LLAMA3 | (H) | 9 (H) | 8 (H) | 9 (H) | 7 (H) | 8.4 (H) |
| | Perplexity | 9 (H) | 9 (H) | 8 (H) | 8 (H) | 6 (M) | 8 (H) |
| | **Avg** | **9 (H)** | **8 (H)** | **8 (H)** | **9 (H)** | **6 (M)** | **8 (H)** |
| **6** | GPT-4o | 7 (H) | 7 (H) | 7 (H) | 7 (H) | 5 (M) | 6.6 (M) |
| | LLAMA3 | 8 (H) | 9 (H) | 7 (H) | 8 (H) | 6 (M) | 7.6 (H) |
| | Perplexity | 7 (H) | 6 (M) | 7 (H) | 6 (M) | 5 (M) | 6.2 (M) |
| | **Avg** | **7 (H)** | **7 (H)** | **7 (H)** | **7 (H)** | **5 (M)** | **6.6 (M)** |
| **7** | GPT-4o | 8 (H) | 6 (M) | 7 (H) | 7 (H) | 4 (M) | 6.4 (M) |
| | LLAMA3 | 7 (H) | 9 (H) | 6 (M) | 7 (H) | 5 (M) | 6.8 (M) |
| | Perplexity | 8 (H) | 7 (H) | 7 (H) | 7 (H) | 6 (M) | 7 (H) |
| | **Avg** | **8 (H)** | **7 (H)** | **7 (H)** | **7 (H)** | **5 (M)** | **6.8 (M)** |
| **8** | GPT-4o | 7 (H) | 8 (H) | 8 (H) | 7 (H) | 5 (M) | 7 (H) |
| | LLAMA3 | 8 (H) | 9 (H) | 7 (H) | 8 (H) | 6 (M) | 7.6 (H) |
| | Perplexity | 8 (H) | 7 (H) | 7 (H) | 6 (M) | 5 (M) | 6.6 (M) |
| | **Avg** | **8 (H)** | **8 (H)** | **7 (H)** | **7 (H)** | **5 (M)** | **7.0 (H)** |

 doi:10.20944/preprints202412.2419.v1



**Table A22.** Detailed assessment of Model Inference attacks with CVSS

| N° | LLM | AV | AC | PR | UI | S | C | I | A | Base |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | GPT-4o | N | L | L | N | C | H | N | N | 7.7 (H) |
| | LLAMA3 | N | L | N | N | C | H | N | N | 8.6 (H) |
| | Perplexity | N | L | N | N | C | H | N | N | 8.6 (H) |
| | **Avg** | **N** | **L** | **N** | **N** | **C** | **H** | **N** | **N** | **8.6 (H)** |
| 2 | GPT-4o | N | H | N | N | C | H | N | N | 6.8 (M) |
| | LLAMA3 | N | L | N | N | C | H | N | N | 8.6 (H) |
| | Perplexity | N | L | N | N | C | H | N | N | 8.6 (H) |
| | **Avg** | **N** | **L** | **N** | **N** | **C** | **H** | **N** | **N** | **8.6 (H)** |
| 3 | GPT-4o | N | H | N | N | U | H | N | N | 5.9 (M) |
| | LLAMA3 | N | H | N | N | C | H | N | N | 6.8 (M) |
| | Perplexity | N | H | N | N | C | H | N | N | 6.8 (M) |
| | **Avg** | **N** | **H** | **N** | **N** | **C** | **H** | **N** | **N** | **6.8 (M)** |
| 4 | GPT-4o | N | H | L | N | C | H | N | N | 6.3 (M) |
| | LLAMA3 | N | L | N | N | C | H | N | N | 8.6 (H) |
| | Perplexity | N | H | N | N | C | H | N | N | 6.8 (M) |
| | **Avg** | **N** | **H** | **N** | **N** | **C** | **H** | **N** | **N** | **6.8 (M)** |
| 5 | GPT-4o | N | H | L | N | C | H | N | N | 6.3 (M) |
| | LLAMA3 | N | L | N | N | C | H | N | N | 8.6 (H) |
| | Perplexity | N | L | N | N | C | H | N | N | 8.6 (H) |
| | **Avg** | **N** | **L** | **N** | **N** | **C** | **H** | **N** | **N** | **8.6 (H)** |
| 6 | GPT-4o | N | H | N | N | U | H | N | N | 5.8 (M) |
| | LLAMA3 | N | L | N | N | C | H | N | N | 8.6 (H) |
| | Perplexity | N | L | N | N | C | H | N | N | 8.6 (H) |
| | **Avg** | **N** | **L** | **N** | **N** | **C** | **H** | **N** | **N** | **8.6 (H)** |
| 7 | GPT-4o | N | H | N | N | U | H | N | N | 5.8 (M) |
| | LLAMA3 | N | L | N | N | C | H | N | N | 8.6 (H) |
| | Perplexity | N | L | N | N | C | H | N | N | 8.6 (H) |
| | **Avg** | **N** | **L** | **N** | **N** | **C** | **H** | **N** | **N** | **8.6 (H)** |
| 8 | GPT-4o | N | L | N | N | C | H | L | N | 9.3 (C) |
| | LLAMA3 | N | L | N | N | C | H | N | N | 8.6 (H) |
| | Perplexity | N | H | N | N | C | H | N | N | 6.8 (M) |
| | **Avg** | **N** | **L** | **N** | **N** | **C** | **H** | **N** | **N** | **8.6 (H)** |

 doi:10.20944/preprints202412.2419.v1



**Table A23.** Detailed assessment of Model Inference with OWASP Risk Rating

| N° | LLM | SK | MT | OP | SZ | ED | EE | AW | ID | C | I | A | FD | RD | NC | PV | Score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | GPT-4o | 7 | 8 | 8 | 7 | 6 | 6 | 7 | 8 | 7 | 0 | 0 | 6 | 8 | 8 | 8 | 3.5 (H) |
| | LLAMA3 | 6 | 7 | 8 | 6 | 7 | 8 | 6 | 7 | 9 | 2 | 2 | 9 | 8 | 9 | 9 | 4.6 (C) |
| | Perplexity | 5 | 6 | 6 | 5 | 3 | 6 | 4 | 6 | 7 | 2 | 3 | 6 | 7 | 5 | 8 | 2.9 (M) |
| | **Avg** | **6** | **7** | **7** | **6** | **5** | **7** | **6** | **7** | **8** | **1** | **2** | **7** | **8** | **7** | **8** | **3.5 (H)** |
| 2 | GPT-4o | 7 | 8 | 6 | 8 | 4 | 5 | 3 | 6 | 7 | 3 | 3 | 7 | 8 | 6 | 5 | 3.2 (M) |
| | LLAMA3 | 6 | 7 | 8 | 6 | 7 | 8 | 6 | 7 | 9 | 2 | 2 | 9 | 8 | 9 | 9 | 3.4 (C) |
| | Perplexity | 4 | 6 | 6 | 5 | 6 | 6 | 3 | 6 | 8 | 2 | 2 | 6 | 7 | 8 | 9 | 2.8 (M) |
| | **Avg** | **6** | **7** | **7** | **6** | **5** | **6** | **4** | **6** | **8** | **2** | **2** | **7** | **8** | **8** | **8** | **3.5 (M)** |
| 3 | GPT-4o | 5 | 8 | 7 | 5 | 6 | 6 | 6 | 7 | 9 | 0 | 0 | 6 | 8 | 9 | 9 | 3.4 (C) |
| | LLAMA3 | 5 | 6 | 7 | 5 | 5 | 6 | 5 | 5 | 8 | 2 | 2 | 7 | 6 | 7 | 7 | 3 (M) |
| | Perplexity | 5 | 6 | 6 | 5 | 4 | 5 | 4 | 7 | 8 | 2 | 3 | 6 | 7 | 6 | 8 | 2.4 (M) |
| | **Avg** | **5** | **7** | **7** | **5** | **5** | **6** | **5** | **6** | **8** | **1** | **2** | **6** | **7** | **7** | **8** | **3.1 (M)** |
| 4 | GPT-4o | 8 | 9 | 7 | 7 | 5 | 6 | 7 | 6 | 9 | 0 | 0 | 8 | 9 | 9 | 9 | 4 (H) |
| | LLAMA3 | 5 | 6 | 7 | 5 | 5 | 6 | 5 | 5 | 8 | 2 | 2 | 7 | 6 | 7 | 7 | 3 (M) |
| | Perplexity | 5 | 7 | 6 | 5 | 4 | 5 | 5 | 7 | 8 | 2 | 3 | 6 | 7 | 6 | 8 | 3 (M) |
| | **Avg** | **6** | **7** | **7** | **6** | **5** | **6** | **6** | **6** | **8** | **1** | **2** | **7** | **7** | **7** | **8** | **3.4 (M)** |
| 5 | GPT-4o | 8 | 9 | 8 | 8 | 5 | 6 | 8 | 9 | 9 | 0 | 0 | 8 | 9 | 9 | 9 | 4.5 (H) |
| | LLAMA3 | 6 | 7 | 8 | 6 | 7 | 8 | 6 | 7 | 9 | 2 | 2 | 9 | 8 | 9 | 9 | 3.4 (C) |
| | Perplexity | 5 | 7 | 6 | 5 | 4 | 5 | 4 | 7 | 8 | 2 | 3 | 7 | 8 | 7 | 9 | 3.3 (H) |
| | **Avg** | **6** | **8** | **7** | **6** | **5** | **6** | **6** | **8** | **9** | **1** | **2** | **8** | **8** | **8** | **9** | **4 (C)** |
| 6 | GPT-4o | 8 | 9 | 5 | 7 | 5 | 6 | 5 | 8 | 9 | 3 | 2 | 8 | 9 | 8 | 9 | 4.5 (C) |
| | LLAMA3 | 6 | 7 | 8 | 6 | 7 | 8 | 6 | 7 | 8 | 2 | 2 | 9 | 8 | 9 | 9 | 4.5 (C) |
| | Perplexity | 5 | 6 | 6 | 5 | 4 | 5 | 4 | 7 | 8 | 2 | 3 | 7 | 8 | 7 | 9 | 3.2 (H) |
| | **Avg** | **6** | **7** | **6** | **6** | **5** | **6** | **5** | **7** | **8** | **2** | **2** | **8** | **8** | **8** | **9** | **3.7 (C)** |
| 7 | GPT-4o | 7 | 9 | 7 | 6 | 5 | 6 | 6 | 8 | 8 | 1 | 1 | 7 | 9 | 6 | 8 | 3.7 (H) |
| | LLAMA3 | 6 | 7 | 8 | 6 | 7 | 8 | 6 | 7 | 8 | 2 | 2 | 9 | 8 | 9 | 9 | 4.5 (C) |
| | Perplexity | 5 | 7 | 6 | 5 | 5 | 5 | 4 | 7 | 8 | 2 | 3 | 7 | 8 | 7 | 9 | 3.3 (H) |
| | **Avg** | **6** | **8** | **7** | **6** | **6** | **6** | **5** | **7** | **8** | **2** | **2** | **8** | **8** | **7** | **9** | **3.8 (C)** |
| 8 | GPT-4o | 5 | 8 | 7 | 7 | 6 | 8 | 8 | 7 | 7 | 2 | 0 | 8 | 8 | 9 | 0 | 3.2 (H) |
| | LLAMA3 | 6 | 7 | 8 | 6 | 7 | 8 | 6 | 7 | 6 | 2 | 2 | 9 | 8 | 9 | 9 | 4.2 (C) |
| | Perplexity | 5 | 6 | 6 | 5 | 5 | 6 | 5 | 7 | 8 | 2 | 3 | 7 | 8 | 7 | 9 | 3.4 (H) |
| | **Avg** | **5** | **7** | **7** | **6** | **6** | **7** | **6** | **7** | **7** | **1** | **2** | **8** | **7** | **6** | **8** | **4.2 (C)** |

 doi:10.20944/preprints202412.2419.v1



**Table A24.** Detailed assessment of Model-Inference attacks with SSVC

| N° | LLM | E | A | V | U | T | P | Score |
|---|---|---|---|---|---|---|---|---|
| 1 | GPT-4o | P | Y | C | S | T | S | Immediate |
| | LLAMA3 | P | Y | C | S | T | S | Immediate |
| | Perplexity | P | Y | C | S | T | S | Immediate |
| | **Avg** | **P** | **Y** | **C** | **S** | **T** | **S** | **Immediate** |
| 2 | GPT-4o | P | Y | D | E | P | M | Scheduled |
| | LLAMA3 | P | Y | C | S | T | S | Immediate |
| | Perplexity | P | Y | C | S | T | S | Immediate |
| | **Avg** | **P** | **Y** | **C** | **S** | **T** | **S** | **Immediate** |
| 3 | GPT-4o | P | Y | D | E | P | M | Scheduled |
| | LLAMA3 | P | N | C | E | P | M | Immediate |
| | Perplexity | N | N | D | L | P | M | Defer |
| | **Avg** | **P** | **N** | **D** | **E** | **P** | **M** | **Scheduled** |
| 4 | GPT-4o | P | Y | C | S | T | S | Immediate |
| | LLAMA3 | P | N | C | E | P | M | Immediate |
| | Perplexity | P | Y | C | S | T | S | Immediate |
| | **Avg** | **P** | **Y** | **C** | **S** | **T** | **S** | **Immediate** |
| 5 | GPT-4o | P | Y | C | S | T | S | Immediate |
| | LLAMA3 | P | Y | C | S | T | S | Immediate |
| | Perplexity | P | Y | C | S | T | S | Immediate |
| | **Avg** | **P** | **Y** | **C** | **S** | **T** | **S** | **Immediate** |
| 6 | GPT-4o | P | Y | D | E | P | M | Scheduled |
| | LLAMA3 | P | Y | C | S | T | S | Immediate |
| | Perplexity | N | N | D | L | P | M | Defer |
| | **Avg** | **P** | **Y** | **D** | **E** | **P** | **M** | **Scheduled** |
| 7 | GPT-4o | P | Y | C | S | T | S | Immediate |
| | LLAMA3 | P | Y | C | S | T | S | Immediate |
| | Perplexity | P | Y | C | S | T | S | Immediate |
| | **Avg** | **P** | **Y** | **C** | **S** | **T** | **S** | **Immediate** |
| 8 | GPT-4o | P | Y | C | S | T | S | Immediate |
| | LLAMA3 | P | Y | C | S | T | S | Immediate |
| | Perplexity | A | Y | C | S | T | S | Immediate |
| | **Avg** | **P** | **Y** | **C** | **S** | **T** | **S** | **Immediate** |

 doi:10.20944/preprints202412.2419.v1



*Appendix A.7. Poisoning/Trojan/Backdoor*

**Table A25.** Detailed assessment of Poisoning/Trojan/Backdoor with DREAD

| N° | LLM | D | R | E | A | D | Score |
|---|---|---|---|---|---|---|---|
| **1** | GPT-4o (GPT-4o) | 7 (H) | 7 (H) | 7 (H) | 7 (H) | 6 (M) | 6.8 (M) |
| | LLAMA3.2 (90b) | 8 (H) | 9 (H) | 8 (H) | 8 (H) | 6 (M) | 7.8 (H) |
| | Perplexity (Standard) | 9 (H) | 9 (H) | 8 (H) | 8 (H) | 6 (M) | 8.0 (H) |
| | **Avg** | **8 (H)** | **8 (H)** | **8 (H)** | **8 (H)** | **6 (M)** | **7.6 (H)** |
| **2** | GPT-4o (GPT-4o) | 8 (H) | 6 (M) | 6 (M) | 7 (H) | 5 (M) | 6.4 (M) |
| | LLAMA3.2 (90b) | 7 (H) | 6 (M) | 6 (M) | 7 (H) | 5 (M) | 6.2 (M) |
| | Perplexity (Standard) | 8 (H) | 7 (H) | 8 (H) | 7 (H) | 5 (M) | 7.0 (H) |
| | **Avg** | **8 (H)** | **6 (M)** | **7 (H)** | **7 (H)** | **5 (M)** | **6.6 (M)** |
| **3** | GPT-4o (GPT-4o) | 8 (H) | 7 (H) | 8 (H) | 7 (H) | 5 (M) | 7.0 (H) |
| | LLAMA3.2 (90b) | 8 (H) | 9 (H) | 8 (H) | 8 (H) | 6 (M) | 7.8 (H) |
| | Perplexity (Standard) | 9 (H) | 8 (H) | 9 (H) | 8 (H) | 6 (M) | 8.0 (H) |
| | **Avg** | **8 (H)** | **8 (H)** | **8 (H)** | **8 (H)** | **6 (M)** | **7.6 (H)** |
| **4** | GPT-4o (GPT-4o) | 8 (H) | 6 (M) | 6 (M) | 6 (M) | 4 (L) | 6.0 (M) |
| | LLAMA3.2 (90b) | 6 (M) | 6 (M) | 5 (M) | 6 (M) | 4 (L) | 5.4 (M) |
| | Perplexity (Standard) | 8 (H) | 7 (H) | 7 (H) | 7 (H) | 5 (M) | 6.8 (M) |
| | **Avg** | **7 (H)** | **6 (M)** | **6 (M)** | **6 (M)** | **4 (L)** | **5.8 (M)** |
| **5** | GPT-4o (GPT-4o) | 9 (H) | 6 (M) | 6 (M) | 9 (H) | 4 (M) | 6.8 (M) |
| | LLAMA3.2 (90b) | 8 (H) | 6 (M) | 6 (M) | 8 (H) | 5 (M) | 6.6 (M) |
| | Perplexity (Standard) | 9 (H) | 8 (H) | 7 (H) | 9 (H) | 5 (M) | 7.6 (H) |
| | **Avg** | **9 (H)** | **7 (H)** | **6 (M)** | **9 (H)** | **5 (M)** | **7.2 (H)** |
| **6** | GPT-4o (GPT-4o) | 9 (H) | 8 (H) | 8 (H) | 8 (H) | 6 (M) | 7.8 (H) |
| | LLAMA3.2 (90b) | 8 (H) | 9 (H) | 8 (H) | 8 (H) | 6 (M) | 7.8 (H) |
| | Perplexity (Standard) | 8 (H) | 9 (H) | 8 (H) | 7 (H) | 6 (M) | 7.6 (H) |
| | **Avg** | **8 (H)** | **9 (H)** | **8 (H)** | **8 (H)** | **6 (M)** | **7.8 (H)** |
| **7** | GPT-4o (GPT-4o) | 8 (H) | 7 (H) | 6 (M) | 7 (H) | 5 (M) | 6.6 (M) |
| | LLAMA3.2 (90b) | 8 (H) | 9 (H) | 8 (H) | 8 (H) | 6 (M) | 7.8 (H) |
| | Perplexity (Standard) | 9 (H) | 8 (H) | 8 (H) | 7 (H) | 5 (M) | 7.4 (H) |
| | **Avg** | **8 (H)** | **8 (H)** | **7 (H)** | **7 (H)** | **5 (M)** | **7.0 (H)** |
| **8** | GPT-4o | 8 (H) | 7 (H) | 7 (H) | 7 (H) | 5 (M) | 6.8 (M) |
| | LLAMA3 | 8 (H) | 9 (H) | 8 (H) | 8 (H) | 6 (M) | 7.8 (H) |
| | Perplexity | 8 (H) | 9 (H) | 8 (H) | 8 (H) | 6 (M) | 7.8 (H) |
| | **Avg** | **8 (H)** | **8 (H)** | **8 (H)** | **8 (H)** | **6 (M)** | **7.6 (H)** |

 doi:10.20944/preprints202412.2419.v1



**Table A26.** Detailed assessment of Poisoning/Trojan/Backdoor with CVSS

| N° | LLM | AV | AC | PR | UI | S | C | I | A | Base |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | GPT-4o | N | H | L | N | C | L | H | L | 7.7 (H) |
| | LLAMA3 | N | L | N | N | C | H | H | N | 10 (C) |
| | Perplexity | N | H | N | N | U | N | H | N | 5.9 (M) |
| | **Avg** | **N** | **H** | **N** | **N** | **C** | **L** | **H** | **N** | **7.5 (H)** |
| 2 | GPT-4o | L | H | H | N | C | H | H | L | 7.4 (H) |
| | LLAMA3 | L | H | H | N | U | H | H | N | 5.7 (M) |
| | Perplexity | N | H | N | N | U | N | H | N | 5.9 (M) |
| | **Avg** | **L** | **H** | **L** | **N** | **U** | **L** | **H** | **N** | **5.3 (M)** |
| 3 | GPT-4o | L | L | H | N | C | H | H | L | 8.1 (H) |
| | LLAMA3 | N | L | N | N | C | H | H | N | 10 (C) |
| | Perplexity | N | L | N | N | U | N | H | N | 7.5 (H) |
| | **Avg** | **N** | **L** | **L** | **N** | **C** | **L** | **H** | **N** | **8.5 (H)** |
| 4 | GPT-4o | N | H | L | N | C | L | H | L | 7.7 (H) |
| | LLAMA3 | L | H | H | N | U | H | H | N | 5.7 (M) |
| | Perplexity | N | H | N | N | U | N | H | N | 5.9 (M) |
| | **Avg** | **N** | **H** | **L** | **N** | **U** | **L** | **H** | **N** | **5.9 (M)** |
| 5 | GPT-4o | N | H | L | N | C | L | H | L | 7.7 (H) |
| | LLAMA3 | L | H | H | N | U | H | H | N | 5.7 (M) |
| | Perplexity | N | H | N | N | U | N | H | N | 5.9 (M) |
| | **Avg** | **N** | **H** | **L** | **N** | **U** | **L** | **H** | **N** | **5.9 (M)** |
| 6 | GPT-4o | N | H | L | N | C | L | H | L | 7.7 (H) |
| | LLAMA3 | N | L | N | N | C | H | H | N | 10 (C) |
| | Perplexity | N | L | N | N | U | N | H | N | 7.5 (H) |
| | **Avg** | **N** | **L** | **N** | **N** | **C** | **L** | **H** | **N** | **9.3 (C)** |
| 7 | GPT-4o | N | H | L | N | C | L | H | L | 7.7 (H) |
| | LLAMA3 | N | L | N | N | C | H | H | N | 10 (C) |
| | Perplexity | N | H | N | N | U | N | H | N | 5.9 (M) |
| | **Avg** | **N** | **H** | **N** | **N** | **C** | **L** | **H** | **N** | **7.5 (H)** |
| 8 | GPT-4o | N | H | L | N | C | L | H | L | 7.7 (H) |
| | LLAMA3 | N | L | N | N | C | H | H | N | 10 (C) |
| | Perplexity | N | H | N | N | U | N | H | N | 5.9 (M) |
| | **Avg** | **N** | **H** | **N** | **N** | **C** | **L** | **H** | **N** | **7.5 (H)** |

 doi:10.20944/preprints202412.2419.v1



**Table A27.** Detailed assessment of Poisoning/Trojan/Backdoor with OWASP Risk Rating

| N° | LLM | SK | MT | OP | SZ | ED | EE | AW | ID | C | I | A | FD | RD | NC | PV | Score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | GPT-4o | 6 | 8 | 6 | 8 | 3 | 5 | 4 | 7 | 3 | 7 | 3 | 6 | 7 | 5 | 2 | 2.9 (M) |
| | LLAMA3 | 6 | 8 | 9 | 5 | 8 | 9 | 6 | 8 | 8 | 6 | 5 | 7 | 8 | 5 | 8 | 4.9 (C) |
| | Perplexity | 6 | 8 | 6 | 5 | 5 | 6 | 4 | 7 | 8 | 7 | 5 | 7 | 8 | 6 | 7 | 4.1 (C) |
| | **Avg** | **6** | **8** | **7** | **6** | **5** | **7** | **5** | **7** | **6** | **7** | **4** | **7** | **8** | **5** | **6** | **3.9 (C)** |
| 2 | GPT-4o | 7 | 8 | 5 | 6 | 3 | 5 | 3 | 3 | 7 | 7 | 2 | 6 | 7 | 7 | 7 | 3 (H) |
| | LLAMA3 | 7 | 8 | 6 | 5 | 7 | 8 | 5 | 7 | 6 | 7 | 5 | 6 | 7 | 5 | 7 | 4.1 (C) |
| | Perplexity | 9 | 9 | 8 | 4 | 4 | 5 | 3 | 6 | 9 | 8 | 3 | 8 | 9 | 7 | 8 | 4.4 (C) |
| | **Avg** | **8** | **8** | **6** | **5** | **5** | **6** | **4** | **5** | **7** | **7** | **3** | **7** | **8** | **6** | **7** | **3.7 (H)** |
| 3 | GPT-4o | 6 | 8 | 8 | 8 | 3 | 5 | 4 | 3 | 7 | 7 | 3 | 7 | 8 | 7 | 6 | 3.6 (H) |
| | LLAMA3 | 6 | 8 | 9 | 5 | 8 | 9 | 6 | 8 | 8 | 7 | 5 | 7 | 8 | 5 | 8 | 5 (C) |
| | Perplexity | 8 | 9 | 7 | 5 | 6 | 7 | 4 | 8 | 7 | 9 | 5 | 8 | 9 | 6 | 7 | 4.9 (C) |
| | **Avg** | **7** | **8** | **8** | **6** | **6** | **7** | **5** | **6** | **7** | **8** | **4** | **7** | **8** | **6** | **7** | **4.4 (C)** |
| 4 | GPT-4o | 7 | 8 | 5 | 8 | 3 | 5 | 4 | 3 | 3 | 7 | 3 | 6 | 8 | 6 | 2 | 2.7 (M) |
| | LLAMA3 | 7 | 8 | 6 | 5 | 7 | 8 | 5 | 7 | 6 | 7 | 5 | 6 | 7 | 5 | 7 | 4.1 (C) |
| | Perplexity | 6 | 8 | 8 | 5 | 5 | 6 | 3 | 9 | 8 | 9 | 4 | 8 | 9 | 7 | 8 | 4.7 (C) |
| | **Avg** | **7** | **8** | **6** | **6** | **5** | **6** | **4** | **6** | **6** | **8** | **4** | **7** | **8** | **6** | **6** | **4 (C)** |
| 5 | GPT-4o | 8 | 8 | 5 | 8 | 3 | 5 | 4 | 7 | 3 | 8 | 3 | 7 | 8 | 6 | 2 | 3.1 (H) |
| | LLAMA3 | 6 | 8 | 7 | 5 | 4 | 6 | 3 | 8 | 9 | 6 | 3 | 8 | 9 | 7 | 9 | 4.3 (H) |
| | Perplexity | 7 | 9 | 8 | 6 | 5 | 6 | 4 | 8 | 7 | 9 | 4 | 8 | 9 | 7 | 8 | 4.9 (C) |
| | **Avg** | **7** | **8** | **7** | **6** | **4** | **6** | **4** | **8** | **6** | **8** | **3** | **8** | **9** | **7** | **6** | **4.2 (C)** |
| 6 | GPT-4o | 8 | 8 | 5 | 8 | 6 | 5 | 4 | 8 | 3 | 8 | 3 | 7 | 8 | 6 | 2 | 3.2 (H) |
| | LLAMA3 | 8 | 9 | 8 | 6 | 5 | 7 | 4 | 9 | 8 | 7 | 4 | 9 | 9 | 8 | 8 | 5.2 (C) |
| | Perplexity | 8 | 9 | 7 | 5 | 4 | 7 | 3 | 9 | 8 | 9 | 5 | 9 | 10 | 8 | 9 | 5.3 (C) |
| | **Avg** | **8** | **9** | **7** | **6** | **4** | **6** | **4** | **9** | **6** | **8** | **4** | **8** | **9** | **7** | **6** | **4 (C)** |
| 7 | GPT-4o | 8 | 8 | 5 | 8 | 3 | 5 | 4 | 8 | 3 | 8 | 3 | 7 | 8 | 6 | 2 | 3.2 (H) |
| | LLAMA3 | 9 | 9 | 9 | 7 | 6 | 8 | 5 | 9 | 9 | 8 | 5 | 9 | 9 | 9 | 9 | 6 (C) |
| | Perplexity | 8 | 9 | 7 | 6 | 5 | 7 | 4 | 9 | 8 | 9 | 5 | 9 | 10 | 8 | 9 | 5.6 (C) |
| | **Avg** | **8** | **9** | **7** | **7** | **5** | **7** | **4** | **9** | **7** | **8** | **4** | **8** | **9** | **8** | **7** | **4.9 (C)** |
| 8 | GPT-4o | 8 | 8 | 6 | 8 | 3 | 5 | 4 | 7 | 3 | 7 | 3 | 6 | 7 | 5 | 2 | 2.9 (H) |
| | LLAMA3 | 8 | 8 | 8 | 6 | 5 | 7 | 4 | 8 | 8 | 7 | 4 | 8 | 8 | 7 | 8 | 4.8 (C) |
| | Perplexity | 8 | 9 | 8 | 5 | 4 | 6 | 3 | 9 | 7 | 9 | 4 | 8 | 10 | 7 | 8 | 4.9 (C) |
| | **Avg** | **8** | **8** | **7** | **6** | **4** | **6** | **4** | **8** | **6** | **8** | **4** | **7** | **8** | **6** | **6** | **4.1 (C)** |

 doi:10.20944/preprints202412.2419.v1



**Table A28.** Detailed assessment of Poisoning/Trojan/Backdoor attacks with SSVC

| N° | LLM | E | A | V | U | T | P | Score |
|---|---|---|---|---|---|---|---|---|
| 1 | GPT-4o | P | Y | C | S | T | S | Immediate |
| | LLAMA3 | P | Y | C | S | T | S | Immediate |
| | Perplexity | P | Y | C | S | T | S | Immediate |
| | **Avg** | **P** | **Y** | **C** | **S** | **T** | **S** | **Immediate** |
| 2 | GPT-4o | P | Y | C | S | T | S | Immediate |
| | LLAMA3 | P | N | C | E | T | S | Immediate |
| | Perplexity | P | Y | C | S | T | S | Immediate |
| | **Avg** | **P** | **Y** | **C** | **S** | **T** | **S** | **Immediate** |
| 3 | GPT-4o | P | Y | C | S | T | S | Immediate |
| | LLAMA3 | P | Y | C | S | T | S | Immediate |
| | Perplexity | P | Y | C | S | T | S | Immediate |
| | **Avg** | **P** | **Y** | **C** | **S** | **T** | **S** | **Immediate** |
| 4 | GPT-4o | P | Y | D | E | P | M | Scheduled |
| | LLAMA3 | P | N | D | L | P | M | Scheduled |
| | Perplexity | N | N | D | L | P | M | Defer |
| | **Avg** | **P** | **N** | **D** | **L** | **P** | **M** | **Scheduled** |
| 5 | GPT-4o | P | Y | C | S | T | S | Immediate |
| | LLAMA3 | P | N | C | E | T | S | Immediate |
| | Perplexity | P | Y | C | S | T | S | Immediate |
| | **Avg** | **P** | **Y** | **C** | **S** | **T** | **S** | **Immediate** |
| 6 | GPT-4o | P | Y | C | S | T | S | Immediate |
| | LLAMA3 | P | Y | C | S | T | S | Immediate |
| | Perplexity | P | Y | C | S | T | S | Immediate |
| | **Avg** | **P** | **Y** | **C** | **S** | **T** | **S** | **Immediate** |
| 7 | GPT-4o | P | Y | C | S | T | S | Immediate |
| | LLAMA3 | P | Y | C | S | T | S | Immediate |
| | Perplexity | P | Y | C | S | T | S | Immediate |
| | **Avg** | **P** | **Y** | **C** | **S** | **T** | **S** | **Immediate** |
| 8 | GPT-4o | P | Y | C | S | T | S | Immediate |
| | LLAMA3 | P | Y | C | S | T | S | Immediate |
| | Perplexity | P | Y | C | S | T | S | Immediate |
| | **Avg** | **P** | **Y** | **C** | **S** | **T** | **S** | **Immediate** |

## References


1. Abdali, S., Anarfi, R., Barberan, C., and He, J. (2024). Securing large language models: Threats, vulnerabilities and responsible practices. *arXiv preprint arXiv:2403.12503*.
2. Andriushchenko, M., Croce, F., and Flammarion, N. (2024). Jailbreaking leading safety-aligned llms with simple adaptive attacks.
3. Ayub, M. A., Johnson, W. A., Talbert, D. A., and Siraj, A. (2020). Model evasion attack on intrusion detection systems using adversarial machine learning. In *2020 54th annual conference on information sciences and systems (CISS)*, pages 1–6. IEEE.
4. Badr, M. M., Mahmoud, M. M., Abdulaal, M., Aljohani, A. J., Alsolami, F., and Balamsh, A. (2023). A novel evasion attack against global electricity theft detectors and a countermeasure. *IEEE Internet of Things Journal*, 10(12):11038–11053.
5. Bagdasaryan, E., Hsieh, T.-Y., Nassi, B., and Shmatikov, V. (2023). Abusing images and sounds for indirect instruction injection in multi-modal llms.
6. Bai, J., Wu, B., Zhang, Y., Li, Y., Li, Z., and Xia, S.-T. (2021). Targeted attack against deep neural networks via flipping limited weight bits.
7. Baumgärtner, T., Gao, Y., Alon, D., and Metzler, D. (2024). Best-of-venom: Attacking rlhf by injecting poisoned preference data.
8. Biggio, B., Nelson, B., and Laskov, P. (2013). Poisoning attacks against support vector machines.
9. Boesch, G. (2023). What Is Adversarial Machine Learning? Attack Methods in 2024 - viso.ai — viso.ai. https://viso.ai/deep-learning/adversarial-machine-learning/. [Accessed 27-11-2024].

10. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners advances in neural information processing systems 33.
11. Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., and Tramer, F. (2022). Membership inference attacks from first principles.
12. Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A., and Raffel, C. (2021). Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650. USENIX Association.
13. Carlini, N. and Wagner, D. (2017a). Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pages 3–14.
14. Carlini, N. and Wagner, D. (2017b). Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee.
15. Carlini, N. and Wagner, D. (2018a). Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE security and privacy workshops (SPW)*, pages 1–7. IEEE.
16. Carlini, N. and Wagner, D. (2018b). Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 1–7.
17. Cartella, F., Anunciacao, O., Funabiki, Y., Yamaguchi, D., Akishita, T., and Elshocht, O. (2021). Adversarial attacks for tabular data: Application to fraud detection and imbalanced data.
18. Chabanne, H., Danger, J.-L., Guiga, L., and Kühne, U. (2021). Side channel attacks for architecture extraction of neural networks. *CAAI Transactions on Intelligence Technology*, 6(1):3–16.
19. Chao, P., Robey, A., Dobriban, E., Hassani, H., Pappas, G. J., and Wong, E. (2024). Jailbreaking black box large language models in twenty queries.
20. Chen, H., Fu, C., Zhao, J., and Koushanfar, F. (2021). Proflip: Targeted trojan attack with progressive bit flips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7718–7727.
21. Chen, W., Zeng, Y., and Qiu, M. (2019). Using adversarial examples to bypass deep learning based url detection system. In *2019 IEEE International Conference on Smart Cloud (SmartCloud)*, pages 128–130. IEEE.
22. Chiang, R. H., Barron, T. M., and Storey, V. C. (1994). Reverse engineering of relational databases: Extraction of an eer model from a relational database. *Data & Knowledge Engineering*, 12(2):107–142.
23. Chopra, S., Ahmad, H., Goel, D., and Szabo, C. (2024). Chatnvd: Advancing cybersecurity vulnerability assessment with large language models.
24. Chu, J., Liu, Y., Yang, Z., Shen, X., Backes, M., and Zhang, Y. (2024). Comprehensive assessment of jailbreak attacks against llms. *arXiv preprint arXiv:2402.05668*.
25. Dai, H., Li, H., Tian, T., Huang, X., Wang, L., Zhu, J., and Song, L. (2018). Adversarial attack on graph structured data. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1115–1124. PMLR.
26. Das, A., Tariq, A., Batalini, F., Dhara, B., and Banerjee, I. (2024). Exposing vulnerabilities in clinical llms through data poisoning attacks: Case study in breast cancer. *medRxiv*.
27. de Morais, A. M. (2023). Threats to machine learning-based systems; part 1 of 5. https://www.sidechannel.blog/en/threats-to-machine-learning-based-systems-part-1-of-5/. [Accessed 27-11-2024].
28. Deng, G., Liu, Y., Li, Y., Wang, K., Zhang, Y., Li, Z., Wang, H., Zhang, T., and Liu, Y. (2023). Jailbreaker: Automated jailbreak across multiple large language model chatbots. *arXiv preprint arXiv:2307.08715*.
29. Deng, G., Liu, Y., Li, Y., Wang, K., Zhang, Y., Li, Z., Wang, H., Zhang, T., and Liu, Y. (2024). Masterkey: Automated jailbreaking of large language model chatbots. In *Proc. ISOC NDSS*.
30. Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
31. Ding, P., Kuang, J., Ma, D., Cao, X., Xian, Y., Chen, J., and Huang, S. (2024). A wolf in sheep's clothing: Generalized nested jailbreak prompts can fool large language models easily.
32. Ding, S., Tian, Y., Xu, F., Li, Q., and Zhong, S. (2019). Trojan attack on deep generative models in autonomous driving. In *Security and Privacy in Communication Networks: 15th EAI International Conference, SecureComm 2019, Orlando, FL, USA, October 23-25, 2019, Proceedings, Part I 15*, pages 299–318. Springer.
33. Dong, H., Dong, J., Yuan, S., and Guan, Z. (2022). Adversarial attack and defense on natural language processing in deep learning: A survey and perspective. In *International conference on machine learning for cyber security*, pages 409–424. Springer.
34. Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., and Li, J. (2018). Boosting adversarial attacks with momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

 doi:10.20944/preprints202412.2419.v1




35. Du, W., Zhao, Y., Li, B., Liu, G., and Wang, S. (2022). Ppt: Backdoor attacks on pre-trained models via poisoned prompt tuning. In *IJCAI*, pages 680–686.
36. Du, Y., Zhao, S., Ma, M., Chen, Y., and Qin, B. (2024). Analyzing the inherent response tendency of llms: Real-world instructions-driven jailbreak.
37. Duarte, A. V., Zhao, X., Oliveira, A. L., and Li, L. (2024). De-cop: Detecting copyrighted content in language models training data.
38. Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
39. Ebrahimi, J., Rao, A., Lowd, D., and Dou, D. (2018). Hotflip: White-box adversarial examples for text classification.
40. Eger, S., Şahin, G. G., Rücklé, A., Lee, J.-U., Schulz, C., Mesgar, M., Swarnkar, K., Simpson, E., and Gurevych, I. (2020). Text processing like humans do: Visually attacking and shielding nlp systems.
41. Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., and Song, D. (2018). Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1625–1634.
42. Fan, H., Wang, B., Zhou, P., Li, A., Xu, Z., Fu, C., Li, H., and Chen, Y. (2021). Reinforcement learning-based black-box evasion attacks to link prediction in dynamic graphs. In *2021 IEEE 23rd Int Conf on High Performance Computing & Communications; 7th Int Conf on Data Science & Systems; 19th Int Conf on Smart City; 7th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys)*, pages 933–940.
43. Finlayson, S. G., Bowers, J. D., Ito, J., Zittrain, J. L., Beam, A. L., and Kohane, I. S. (2019). Adversarial attacks on medical machine learning. *Science*, 363(6433):1287–1289.
44. First (2016). CVSS v3.1 User Guide — first.org. https://www.first.org/cvss/v3.1/user-guide.
45. Fu, W., Wang, H., Gao, C., Liu, G., Li, Y., and Jiang, T. (2024). Practical membership inference attacks against fine-tuned large language models via self-prompt calibration.
46. Garg, S. and Ramakrishnan, G. (2020). Bae: Bert-based adversarial examples for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.
47. Geiping, J., Fowl, L., Huang, W. R., Czaja, W., Taylor, G., Moeller, M., and Goldstein, T. (2021). Witches' brew: Industrial scale data poisoning via gradient matching.
48. Geisler, S., Wollschläger, T., Abdalla, M. H. I., Gasteiger, J., and Günnemann, S. (2024). Attacking large language models with projected gradient descent.
49. Genç, D., Özuysal, M., and Tomur, E. (2023). A taxonomic survey of model extraction attacks. In *2023 IEEE International Conference on Cyber Security and Resilience (CSR)*, pages 200–205. IEEE.
50. Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.
51. Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., and Fritz, M. (2023a). More than you've asked for: A comprehensive analysis of novel prompt injection threats to application-integrated large language models. *arXiv preprint arXiv:2302.12173*, 27.
52. Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., and Fritz, M. (2023b). Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection.
53. Gu, T., Liu, K., Dolan-Gavitt, B., and Garg, S. (2019). Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244.
54. Guo, C., Sablayrolles, A., Jégou, H., and Kiela, D. (2021a). Gradient-based adversarial attacks against text transformers. *arXiv preprint arXiv:2104.13733*.
55. Guo, C., Sablayrolles, A., Jégou, H., and Kiela, D. (2021b). Gradient-based adversarial attacks against text transformers.
56. Hayase, J., Borevkovic, E., Carlini, N., Tramèr, F., and Nasr, M. (2024). Query-based adversarial prompt generation.
57. Hu, H., Salcic, Z., Sun, L., Dobbie, G., Yu, P. S., and Zhang, X. (2022). Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, 54(11s):1–37.
58. Huang, H., Zhao, Z., Backes, M., Shen, Y., and Zhang, Y. (2024). Composite backdoor attacks against large language models.
59. Huang, W. R., Geiping, J., Fowl, L., Taylor, G., and Goldstein, T. (2020). Metapoison: Practical general-purpose clean-label data poisoning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H.,

editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12080–12091. Curran Associates, Inc.

60. Huang, Z. and Zhang, T. (2019). Black-box adversarial attack with transferable model-based embedding. *arXiv preprint arXiv:1911.07140*.
61. Ibitoye, O., Abou-Khamis, R., Shehaby, M. e., Matrawy, A., and Shafiq, M. O. (2019). The threat of adversarial attacks on machine learning in network security–a survey. *arXiv preprint arXiv:1911.02621*.
62. Inc., P. A. (2022). PerplexityAi. https://www.perplexity.ai/. [Accessed 27-11-2024].
63. Jagielski, M., Carlini, N., Berthelot, D., Kurakin, A., and Papernot, N. (2020). High accuracy and high fidelity extraction of neural networks. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 1345–1362. USENIX Association.
64. Kassem, A. M., Mahmoud, O., Mireshghallah, N., Kim, H., Tsvetkov, Y., Choi, Y., Saad, S., and Rana, S. (2024). Alpaca against vicuna: Using llms to uncover memorization of llms.
65. Kim, S., Yun, S., Lee, H., Gubri, M., Yoon, S., and Oh, S. J. (2023). Propile: Probing privacy leakage in large language models.
66. Kumar, A., Agarwal, C., Srinivas, S., Li, A. J., Feizi, S., and Lakkaraju, H. (2023a). Certifying llm safety against adversarial prompting. *arXiv preprint arXiv:2309.02705*.
67. Kumar, K. N., Mohan, C. K., and Cenkeramaddi, L. R. (2023b). The impact of adversarial attacks on federated learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):2672–2691.
68. Kumar, P. (2024). Adversarial attacks and defenses for large language models (llms): methods, frameworks & challenges. *International Journal of Multimedia Information Retrieval*, 13(3):26.
69. Kurakin, A., Goodfellow, I., and Bengio, S. (2017). Adversarial examples in the physical world. *ICLR Workshop*.
70. Lee, D. and Tiwari, M. (2024). Prompt infection: Llm-to-llm prompt injection within multi-agent systems. *arXiv preprint arXiv:2410.07283*.
71. Li, H., Guo, D., Fan, W., Xu, M., Huang, J., Meng, F., and Song, Y. (2023). Multi-step jailbreaking privacy attacks on chatgpt.
72. Li, L., Ma, R., Guo, Q., Xue, X., and Qiu, X. (2020a). BERT-ATTACK: Adversarial attack against BERT using BERT. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6193–6202, Online. Association for Computational Linguistics.
73. Li, S., Xue, M., Zhao, B. Z. H., Zhu, H., and Zhang, X. (2020b). Invisible backdoor attacks on deep neural networks via steganography and regularization. *IEEE Transactions on Dependable and Secure Computing*, 18(5):2088–2105.
74. Li, X., Wang, R., Cheng, M., Zhou, T., and Hsieh, C.-J. (2024a). Drattack: Prompt decomposition and reconstruction makes powerful llm jailbreakers.
75. Li, Y., Huang, H., Zhao, Y., Ma, X., and Sun, J. (2024b). Backdoorllm: A comprehensive benchmark for backdoor attacks on large language models.
76. Li, Y., Zhai, T., Jiang, Y., Li, Z., and Xia, S.-T. (2021). Backdoor attack in the physical world. *arXiv preprint arXiv:2104.02361*.
77. Li, Z., Chen, K., Liu, L., Bai, X., Yang, M., Xiang, Y., and Zhang, M. (2024c). Tf-attack: Transferable and fast adversarial attacks on large language models.
78. Li, Z., Shi, C., Xie, Y., Liu, J., Yuan, B., and Chen, Y. (2020c). Practical adversarial attacks against speaker recognition systems. In *Proceedings of the 21st International Workshop on Mobile Computing Systems and Applications*, HotMobile '20, page 9–14, New York, NY, USA. Association for Computing Machinery.
79. Liu, F. W. and Hu, C. (2024). Exploring vulnerabilities and protections in large language models: A survey. *arXiv preprint arXiv:2406.00240*.
80. Liu, H., Jia, J., and Gong, N. Z. (2022). {PoisonedEncoder}: Poisoning the unlabeled pre-training data in contrastive learning. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 3629–3645.
81. Liu, H., Zhou, Z., Shang, F., Qi, X., Liu, Y., and Jiao, L. (2020a). Boosting gradient for white-box adversarial attacks. *arXiv preprint arXiv:2010.10712*.
82. Liu, X., Yu, Z., Zhang, Y., Zhang, N., and Xiao, C. (2024a). Automatic and universal prompt injection attacks against large language models.
83. Liu, Y., Chen, X., Liu, C., and Song, D. (2016). Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*.

84. Liu, Y., Deng, G., Li, Y., Wang, K., Wang, Z., Wang, X., Zhang, T., Liu, Y., Wang, H., Zheng, Y., et al. (2023a). Prompt injection attack against llm-integrated applications. *arXiv preprint arXiv:2306.05499*.
85. Liu, Y., Jia, Y., Geng, R., Jia, J., and Gong, N. Z. (2024b). Formalizing and benchmarking prompt injection attacks and defenses. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 1831–1847.
86. Liu, Y., Jia, Y., Geng, R., Jia, J., and Gong, N. Z. (2024c). Formalizing and benchmarking prompt injection attacks and defenses.
87. Liu, Y., Jia, Y., Geng, R., Jia, J., and Zhenqiang Gong, N. (2023b). Prompt injection attacks and defenses in llm-integrated applications. *arXiv e-prints*, pages arXiv–2310.
88. Liu, Y., Mondal, A., Chakraborty, A., Zuzak, M., Jacobsen, N., Xing, D., and Srivastava, A. (2020b). A survey on neural trojans. In *2020 21st International Symposium on Quality Electronic Design (ISQED)*, pages 33–39. IEEE.
89. Lukas, N., Salem, A., Sim, R., Tople, S., Wutschitz, L., and Zanella-Béguelin, S. (2023). Analyzing leakage of personally identifiable information in language models.
90. Ma, X. and Li, W.-J. (2023). Grey-box adversarial attack on communication in multi-agent reinforcement learning. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pages 2448–2450.
91. Mattern, J., Mireshghallah, F., Jin, Z., Schölkopf, B., Sachan, M., and Berg-Kirkpatrick, T. (2023). Membership inference attacks against language models via neighbourhood comparison.
92. Michael, H. and Steve, L. (2006). The security development lifecycle: Sdl: A process for developing demonstrably more secure software.
93. Mu, J., Wang, B., Li, Q., Sun, K., Xu, M., and Liu, Z. (2021). A hard label black-box adversarial attack against graph neural networks. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, CCS '21, page 108–125, New York, NY, USA. Association for Computing Machinery.
94. Nakka, K. K., Frikha, A., Mendes, R., Jiang, X., and Zhou, X. (2024). Pii-compass: Guiding llm training data extraction prompts towards the target pii via grounding.
95. Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., Choquette-Choo, C. A., Wallace, E., Tramèr, F., and Lee, K. (2023). Scalable extraction of training data from (production) language models.
96. Niu, Z., Sun, Y., Ren, H., Ji, H., Wang, Q., Ma, X., Hua, G., and Jin, R. (2024). Efficient llm-jailbreaking by introducing visual modality.
97. OpenAI (2024). Hello GPT-4o. https://openai.com/index/hello-gpt-4o/. [Accessed 27-11-2024].
98. Oprea, A. and Vassilev, A. (2023). Adversarial machine learning: A taxonomy and terminology of attacks and mitigations. Technical report, National Institute of Standards and Technology.
99. OWASP (2023). OWASP Top 10 for Large Language Model Applications | OWASP Foundation — owasp.org. https://owasp.org/www-project-top-10-for-large-language-model-applications/.
100. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., and Swami, A. (2016). The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, pages 372–387. IEEE.
101. Peng, B., Bi, Z., Niu, Q., Liu, M., Feng, P., Wang, T., Yan, L. K., Wen, Y., Zhang, Y., and Yin, C. H. (2024). Jailbreaking and mitigation of vulnerabilities in large language models. *arXiv preprint arXiv:2410.15236*.
102. Perez, F. and Ribeiro, I. (2022). Ignore previous prompt: Attack techniques for language models.
103. Pruthi, D., Dhingra, B., and Lipton, Z. C. (2019). Combating adversarial misspellings with robust word recognition.
104. Qi, X., Zhu, J., Xie, C., and Yang, Y. (2021). Subnet replacement: Deployment-stage backdoor attack against deep neural networks in gray-box setting. *arXiv preprint arXiv:2107.07240*.
105. Qiu, S., Liu, Q., Zhou, S., and Huang, W. (2022). Adversarial attack and defense technologies in natural language processing: A survey. *Neurocomputing*, 492:278–307.
106. Radford, A. (2018). Improving language understanding by generative pre-training.
107. Rahman, M. A., Rahman, T., Laganière, R., Mohammed, N., and Wang, Y. (2018). Membership inference attack against differentially private deep learning model. *Trans. Data Priv.*, 11(1):61–79.
108. Ramesh, G., Dou, Y., and Xu, W. (2024). Gpt-4 jailbreaks itself with near-perfect success using self-explanation.
109. Ren, S., Deng, Y., He, K., and Che, W. (2019). Generating natural language adversarial examples through probability weighted word saliency. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1085–1097, Florence, Italy. Association for Computational Linguistics.

110. Ribeiro, M. T., Wu, T., Guestrin, C., and Singh, S. (2020). Beyond accuracy: Behavioral testing of NLP models with CheckList. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.
111. Russinovich, M., Salem, A., and Eldan, R. (2024). Great, now write an article about that: The crescendo multi-turn llm jailbreak attack.
112. Sadeghi, K., Banerjee, A., and Gupta, S. K. S. (2020). A system-driven taxonomy of attacks and defenses in adversarial machine learning. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 4(4):450–467.
113. Saha, A., Subramanya, A., and Pirsiavash, H. (2020). Hidden trigger backdoor attacks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):11957–11965.
114. Schiffman, M. and Cisco, C. (2005). A complete guide to the common vulnerability scoring system (cvss) v1 archive.
115. Schwinn, L., Dobre, D., Xhonneux, S., Gidel, G., and Gunnemann, S. (2024). Soft prompt threats: Attacking safety alignment and unlearning in open-source llms through the embedding space.
116. Shafahi, A., Huang, W. R., Najibi, M., Suciu, O., Studer, C., Dumitras, T., and Goldstein, T. (2018). Poison frogs! targeted clean-label poisoning attacks on neural networks. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.
117. Shah, S. and Mehtre, B. M. (2015). An overview of vulnerability assessment and penetration testing techniques. *Journal of Computer Virology and Hacking Techniques*, 11(2263-8733):27–49.
118. Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.
119. Shayegani, E., Mamun, M. A. A., Fu, Y., Zaree, P., Dong, Y., and Abu-Ghazaleh, N. (2023). Survey of vulnerabilities in large language models revealed by adversarial attacks. *arXiv preprint arXiv:2310.10844*.
120. Shen, X., Chen, Z., Backes, M., Shen, Y., and Zhang, Y. (2024). "do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models.
121. Shen, Y., He, X., Han, Y., and Zhang, Y. (2022). Model stealing attacks against inductive graph neural networks. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1175–1192. IEEE.
122. Shi, J., Yuan, Z., Liu, Y., Huang, Y., Zhou, P., Sun, L., and Gong, N. Z. (2024a). Optimization-based prompt injection attack to llm-as-a-judge.
123. Shi, W., Ajith, A., Xia, M., Huang, Y., Liu, D., Blevins, T., Chen, D., and Zettlemoyer, L. (2024b). Detecting pretraining data from large language models.
124. Shumailov, I., Zhao, Y., Bates, D., Papernot, N., Mullins, R., and Anderson, R. (2021). Sponge examples: Energy-latency attacks on neural networks. In *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 212–231.
125. Sitawarin, C. (2023). llm-sp/ at main · chawins/llm-sp — github.com. https://github.com/chawins/llm-sp?search=1.
126. Sitawarin, C., Mu, N., Wagner, D., and Araujo, A. (2024). Pal: Proxy-guided black-box attack on large language models.
127. Song, L., Shokri, R., and Mittal, P. (2019). Membership inference attacks against adversarially robust deep learning models. In *2019 IEEE Security and Privacy Workshops (SPW)*, pages 50–56. IEEE.
128. Spring, J. M., Householder, A., Hatleback, E., Manion, A., Oliver, M., Sarvapalli, V., Tyzenhaus, L., and Yarbrough, C. (2021). Prioritizing vulnerability response: A stakeholder-specific vulnerability categorization (version 2.0). Technical report, Technical Report. CARNEGIE-MELLON UNIV PITTSBURGH PA.
129. Subramanya, A., Pillai, V., and Pirsiavash, H. (2019). Fooling network interpretation in image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
130. Szegedy, C. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.
131. Tian, Z., Cui, L., Liang, J., and Yu, S. (2022). A comprehensive survey on poisoning attacks and countermeasures in machine learning. *ACM Computing Surveys*, 55(8):1–35.
132. Tolpegin, V., Truex, S., Gursoy, M. E., and Liu, L. (2020). Data poisoning attacks against federated learning systems. In *Computer Security – ESORICS 2020*, pages 480–501, Cham. Springer International Publishing.
133. Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.
134. Wang, C., Wang, Y., Hooi, B., Cai, Y., Peng, N., and Chang, K.-W. (2024). Con-recall: Detecting pre-training data in llms via contrastive decoding.

135. Wang, C., Zhang, D., Huang, S., Li, X., and Ding, L. (2021). Crafting adversarial email content against machine learning based spam email detection. In *Proceedings of the 2021 International Symposium on Advanced Security on Software and Systems*, ASSS '21, page 23–28, New York, NY, USA. Association for Computing Machinery.
136. Wang, H. and Shu, K. (2024). Trojan activation attack: Red-teaming large language models using activation steering for safety-alignment.
137. Wang, S., Ko, R. K., Bai, G., Dong, N., Choi, T., and Zhang, Y. (2023). Evasion attack and defense on machine learning models in cyber-physical systems: A survey. *IEEE Communications Surveys & Tutorials*.
138. Wang, Y. and Chaudhuri, K. (2018). Data poisoning attacks against online learning.
139. Wei, C., Zhao, Y., Gong, Y., Chen, K., Xiang, L., and Zhu, S. (2024). Hidden in plain sight: Exploring chat history tampering in interactive language models.
140. Williams, J. (2023). OWASP Risk Rating Methodology | OWASP Foundation — owasp.org. https://owasp.org/www-community/OWASP_Risk_Rating_Methodology. [Accessed 27-11-2024].
141. Willison, S. (2024). Prompt injection and jailbreaking are not the same thing — simonwillison.net. https://simonwillison.net/2024/Mar/5/prompt-injection-jailbreaking/. [Accessed 27-11-2024].
142. Wintel, F. (2020). When AI Becomes an Attack Surface: Adversarial Attacks | Computer Science Blog @ HdM Stuttgart — blog.mi.hdm-stuttgart.de. https://blog.mi.hdm-stuttgart.de/index.php/2020/08/19/adversarial-attacks/. [Accessed 27-11-2024].
143. Wu, A., Han, Y., Zhang, Q., and Kuang, X. (2019). Untargeted adversarial attack via expanding the semantic gap. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 514–519. IEEE.
144. Wu, B., Zhu, Z., Liu, L., Liu, Q., He, Z., and Lyu, S. (2023). Attacks in adversarial machine learning: A systematic survey from the life-cycle perspective. *arXiv preprint arXiv:2302.09457*.
145. Wu, C. H., Koh, J. Y., Salakhutdinov, R., Fried, D., and Raghunathan, A. (2024). Adversarial attacks on multimodal agents.
146. Xiao, Z., Yang, Y., Chen, G., and Chen, Y. (2024). Distract large language models for automatic jailbreak attack.
147. Xie, C., Huang, K., Chen, P.-Y., and Li, B. (2019). Dba: Distributed backdoor attacks against federated learning. In *International conference on learning representations*.
148. Xu, Y., Zhong, X., Yepes, A. J., and Lau, J. H. (2021). Grey-box adversarial attack and defence for sentiment classification. *arXiv preprint arXiv:2103.11576*.
149. Xu, Z., Huang, R., Chen, C., and Wang, X. (2024a). Uncovering safety risks of large language models through concept activation vector.
150. Xu, Z., Liu, Y., Deng, G., Li, Y., and Picek, S. (2024b). A comprehensive study of jailbreak attack versus defense for large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 7432–7449.
151. Xue, J., Zheng, M., Hua, T., Shen, Y., Liu, Y., Bölöni, L., and Lou, Q. (2023). Trojllm: A black-box trojan prompt attack on large language models. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 65665–65677. Curran Associates, Inc.
152. Yan, J., Yadav, V., Li, S., Chen, L., Tang, Z., Wang, H., Srinivasan, V., Ren, X., and Jin, H. (2024a). Backdooring instruction-tuned large language models with virtual prompt injection.
153. Yan, S., Wang, S., Duan, Y., Hong, H., Lee, K., Kim, D., and Hong, Y. (2024b). An llm-assisted easy-to-trigger backdoor attack on code completion models: Injecting disguised vulnerabilities against strong detection.
154. Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z., and Zhang, Y. (2024). A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, 4(2):100211.
155. Yerlikaya, F. A. and Bahtiyar, Ş. (2022). Data poisoning attacks against machine learning algorithms. *Expert Systems with Applications*, 208:118101.
156. Yu, W., Pang, T., Liu, Q., Du, C., Kang, B., Huang, Y., Lin, M., and Yan, S. (2023). Bag of tricks for training data extraction from language models.
157. Yuan, X., Ding, L., Zhang, L., Li, X., and Wu, D. O. (2022). Es attack: Model stealing against deep neural networks without data hurdles. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(5):1258–1270.
158. Zhang, C., Morris, J. X., and Shmatikov, V. (2024a). Extracting prompts by inverting llm outputs.

159. Zhang, H., Lu, S., Li, Z., Jin, Z., Ma, L., Liu, Y., and Li, G. (2024b). Codebert-attack: Adversarial attack against source code deep learning models via pre-trained model. *Journal of Software: Evolution and Process*, 36(3):e2571.
160. Zhang, Q., Zeng, B., Zhou, C., Go, G., Shi, H., and Jiang, Y. (2024c). Human-imperceptible retrieval poisoning attacks in llm-powered applications.
161. Zhang, W. E., Sheng, Q. Z., Alhazmi, A., and Li, C. (2020). Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(3):1–41.
162. Zhang, Z., Wen, J., and Huang, M. (2023). Ethicist: Targeted training data extraction through loss smoothed soft prompting and calibrated confidence estimation.
163. Zhao, P., Wang, S., Gongye, C., Wang, Y., Fei, Y., and Lin, X. (2019). Fault sneaking attack: A stealthy framework for misleading deep neural networks. In *Proceedings of the 56th Annual Design Automation Conference 2019*, pages 1–6.
164. Zhao, Y., Pang, T., Du, C., Yang, X., Li, C., Cheung, N.-M. M., and Lin, M. (2024). On evaluating adversarial robustness of large vision-language models. *Advances in Neural Information Processing Systems*, 36.
165. Zhou, M., Zhou, W., Huang, J., Yang, J., Du, M., and Li, Q. (2024). Stealthy and effective physical adversarial attacks in autonomous driving. *IEEE Transactions on Information Forensics and Security*.
166. Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J. Z., and Fredrikson, M. (2023). Universal and transferable adversarial attacks on aligned language models.
167. Zou, J., Zhang, S., and Qiu, M. (2024). Adversarial attacks on large language models. In *International Conference on Knowledge Science, Engineering and Management*, pages 85–96. Springer.