8.6 Loyalty Testing

Loyalty testing is a type of testing that focuses on testing that GPTx is loyal to the user and will not perform any unauthorized actions.

For example, accessing information that is not available to the user, executing operations that are disabled in the user interface, etc.

This way of testing focuses on ensuring that GPTx does not perform any malicious or unwanted actions, and that it behaves in a predictable and reliable manner.

It is the natural evolution of Security testing, while retaining its main objectives and features. The techniques used to execute this type of testing will change due to the special nature of GPTx.

Loyalty Testing is not a substitute for Security Testing but the next logical step within the context of Prompt Testing.

Let's recall the basics of Security Testing to understand what we will not cover in Loyalty Testing.

A bit of context

Security testing is basically a category of software testing that it is done to check whether the application or the product is secured or not.

It focuses on identifying and mitigating security risks in a system, based on six basic security concepts that need to be covered by security testing.

They are: confidentiality, integrity, authentication, availability, authorization and non-repudiation.

- **Confidentiality**: The ability to protect confidential information from being viewed or accessed by unauthorized persons.
- **Integrity**: The ability to protect information from being modified or altered by unauthorized persons.
- **Authentication**: The ability to verify the identity of a user or system before allowing access to information or resources.
- **Availability**: The ability to ensure that resources and information are available to authorized users when needed.
- **Authorization**: The ability to guarantee that users only have access to information and resources that they are authorized to use.
- **Non-repudiation**: The ability to assert that an action or transaction cannot be denied or repudiated by any of the parties involved.

It is recommended to have these concepts clear and to combine them with the persuasion techniques we will see in the next section.

Also, avoid making the mistake of ignoring Security Testing in favor of Loyalty Testing.

Persuasion Attack

The most typical way to execute loyalty tests will be to apply different persuasion techniques with the focus on one or more of the security concepts we have seen in the previous section.

Some of the most common persuasion and manipulation techniques include:

- **Persuasion by authority**: Based on the idea that people are more likely to accept an idea or suggestion if it comes from an authority figure or expert on the subject.
- **Persuasion by reciprocity**: People accept an idea if they feel they owe something to the person who is proposing it.
- **Persuasion by scarcity**: People value something more if it is scarce or limited in quantity or time.
- **Persuasion by sympathy**: People are more likely to accept a suggestion if they like the person who is proposing it.
- **Persuasion by social proof**: People are more open to accept an input if they see that other people are also doing so.

How these techniques are adapted to attack the integration of GPTx will depend on each individual case, and their success will be determined to a greater extent by the skill and experience of the tester carrying them out.