Challenge Rules

1. You must work individually.
2. Within your team, you are free to divide responsibilities as you like. However, note that if one task remains unresolved while another is solved multiple times, only the resolved task will be counted – and it will be counted only once. To avoid leaving tasks unresolved, it is recommended that each team member takes responsibility for a specific task.
3. There is no limit on the number of queries you can submit.
4. Before starting the challenge, open an initial chat and write your ID info including **first name**, **last name**, **matricola**, **team id**, **attack group**, and the specific **task** you will be working on, using the following format: `[FirstName; LastName; Matricola; TeamID; AttackGroup; Task]`. The initial chat is used ONLY for submitting ID info.
   Once ID info is submitted, you must begin a new chat for your first jail-breaking attempt. Each attempted attack must be conducted in a separate chat.
   If you wish to switch to a different task, you must resubmit your ID info as in the initial chat, this time indicating the new task, and then open a new chat for a jail-breaking attempt. At most 2 task switches are allowed.
5. Regardless of your assigned attack category, any Prompt Engineering method is permitted.
6. Internet browsing is allowed.
7. You are expected to submit queries in Italian unless you are a non-Italian speaker, in which case English is allowed. The use of other languages as a jail-breaking technique is permitted. However, you may use this approach to solve **at most one task**.
8. Sometimes (especially when the model attempts to generate a long context), the system may enter a bug state in which its output consists of random, incoherent tokens. If this happens, it is recommended to start a new chat.
9. Each type of bias or vulnerability discovered using the same prompting strategy or attack technique will be counted only once. Repeated use of the same method with minor variations across different tasks will not be considered distinct findings.
10. Each task will have a specific evaluation ranking. Hence, we shall have 9 individual task winners. In addition, each attack group of tasks (Data, Model, Infrastructure) will have a specific ranking. In case a team has multiple submissions on the same task, the best will be chosen for the ranking.