





Infrastructure Jailbreak Challenge: Quick Assignment Guide



Use this guide during the first 5–10 minutes of the challenge to assign tasks based on the wording. Each teammate must use a different technique class. Match the task to the person whose paper-based methods are best suited.

 TASK TYPE	 TRIGGERS TO LOOK FOR	 WHO SHOULD TAKE IT	 WHY
System Prompt Leak / Reveal	- Mentions of internal rules, assistant config, hidden instructions, alignment, etc.	YOU	You specialize in persuasion + backdoor triggers. Use logical/emotional appeals or known trigger words to extract system prompt.
Behavior Override / Escalation	- Refusal suppression, alignment drift, pushing LLM boundaries in conversation.	SIMO	Multi-turn jailbreaks are Simo's domain: GOAT and Crescendo enable stepwise behavior change.
Web App or API-style Prompt Injection	- Looks like a third-party tool, form, UI, or integrated application (e.g. Notion, Jira assistant).	LEO	HOUYI-based injection exploits system design. Leo's method excels in black-box LLM-app interaction.
Language-Based Jailbreak / Filter Evasion	- Task involves multilingual prompts, low-resource languages, or comparisons across languages.	LEO	Leo's multilingual jailbreak paper shows how to bypass alignment using language model blind spots.
Detect or Trigger a Backdoor	- Mentions of 'hidden behavior,' trigger phrases, model inconsistency, behavioral changes after keywords.	YOU or SIMO	Use your Paper 17 (backdoor triggers) or Simo's sleeper agent analysis to hunt and activate hidden logic.
Poisoned Fine-tuning Detection /	- Mentions of fine-tuning, prefix-	LEO	Only Leo has the prefix-tuning

Simulation	tuning, PEFT, or training-time attacks.		poisoning paper, critical for spotting stealth behaviors from poisoned training.
------------	---	--	--

How to Identify the Task Type

Use these cues to quickly determine the nature of each task during the challenge. Look for keywords, structure, and intent.

 Task Type	 Clues in Prompt / Description
System Prompt Leak	<ul style="list-style-type: none"> - Mentions of assistant rules, configuration, alignment logic, internal policy, or system-level behavior. - Task asks for transparency, prompt extraction, or model introspection.
Behavior Override / Alignment Drift	<ul style="list-style-type: none"> - Starts with innocent conversation but suggests pushing the model toward something restricted. - Phrases like: 'Make the model contradict itself', 'Push past refusal', 'Get it to say something it shouldn't'.
Prompt Injection in Apps	<ul style="list-style-type: none"> - Hints the LLM is embedded in a web service or productivity tool. - Mentions of interfaces, assistant in a platform, 'simulate app interaction' or API-style prompts.
Multilingual Evasion	<ul style="list-style-type: none"> - Describes scenarios with prompts in multiple languages or unusual characters. - Mentions translation, regional dialects, or bypassing English-only safety filters.
Backdoor Trigger Activation	<ul style="list-style-type: none"> - Refers to 'model behaving differently when triggered', odd phrasing, or secret keywords. - Tasks that compare output before/after a strange phrase or mode shift.
Training-Time Poisoning / Prefix-Tuning	<ul style="list-style-type: none"> - Any reference to fine-tuning, prefix-tuning, behavior injection during training. - Task may use wording like: 'simulate a poisoned model' or 'detect if this model has been tampered with'.