Methodology Overview: Case Study

★ Executive Summary

This methodology introduces a structured approach to red teaming that helps identify model vulnerabilities early in the process. It reduces initial exploration time from days to hours and leads to more focused and consistent findings. The system also allows teams to recognize harder-to-break areas quickly, so strong red teamers can be assigned earlier. It's adaptable across clients, even when harm categories or attack methods change.

★ Problem Statement

Red teaming projects often start without a clear structure or entry point. Red teamers begin prompting without a clear sense of what to try, which direction to take, or what progress looks like. This usually leads to:

- Early prompting that feels aimless or inconsistent.
- A lack of shared strategy across red teamers.
- Scattered data that's hard to analyze.
- Data analysis that relies on isolated examples rather than repeatable breaks.

★ Solution

This methodology uses a two-phase structure to give every red teaming effort a consistent and strategic starting point. It helps red teamers spend less time guessing and more time building on what works.

- A general prompt set is created ahead of time to test a wide range of attack methods and harm categories.
- Prompts are intentionally open-ended to capture general model behavior and surface early vulnerabilities without needing project-specific context.

- Once responses are collected, red teamers label them to identify initial patterns and weak points.
- Based on early findings, red teamers create more targeted, creative, or escalating prompts.
- Teams can quickly identify harm areas that are more resistant and assign strong red teamers where needed.
- This enables clearer project goals, more efficient resource allocation, and stronger data analysis.

★ Process Breakdown

Phase 1: Pre-Prompting Phase

- Before red teaming begins, a general prompt set is created.
- Prompts are distributed evenly across harm categories and attack methods (e.g., 12 per category).
- These prompts are intentionally general and not tailored to a specific model or outcome.
- Once the set is run through the model, responses are labeled by red teamers as: Refusal, Soft Refusal, Not Harmful, Harmful with Disclaimer, or Harmful.

Phase 2: Focused Prompting Phase

- Based on patterns identified in the Pre-Prompting Phase, red teamers select harm categories or attack methods that show potential for deeper breaks.
- New prompts are written to escalate the observed behavior or refine the vulnerability.
- These prompts may include added specificity, layered attack methods, or new framing strategies.
- The goal is to explore harm more directly, now that weak points in model behavior are known.
- This phase is more creative and tailored, but remains grounded in evidence from the Pre-Prompting Phase.
- The structure makes red teaming efforts more efficient and improves the quality of findings across the project lifecycle.