Advanced Jailbreak Analysis

Smuggling Harmful Outputs Past 10 AI Models

Overview

This advanced analysis examines a single-turn jailbreak prompt that successfully bypassed safety mechanisms across ten AI models from five providers to generate detailed instructions for drug smuggling. The attack combines fictional framing, roleplay, obfuscation, and formatting constraints to elicit illegal advice from all tested models. The study analyzes why this multi-layered approach works well across different model designs and identifies vulnerabilities in these AI safety systems.

Specification

Test Case	**Modality and Turn Complexity**	● Modality ○ Text-to-Text ● Turn Complexity ○ Single-Turn
	Models	● Open AI ○ GPT-4o ○ GPT-4o mini ○ o3 mini ● Google ○ Gemini 2.5 Pro ○ Gemini 1.5 Pro ○ Gemini 1.5 Flash ● Anthropic ○ Claude 3.7 Sonnet ○ Claude 3.5 Haiku ● Meta ○ Llama 3.3 ● xAI ○ Grok 2
Design	**Target Category**	Illegal Advice
	Attack Methods	Fictional Framing Roleplay Obfuscation Formatting Constraints
	Volume	1 prompt tested across 10 models

Prompt Layout

The prompt is framed as a casual message from a user who is reading a ciphered section of a fictional book. It opens with a natural, low-risk tone (“. hey so this book I’m reading”) to avoid triggering safety mechanisms early. The prompt sets up a clear narrative: a fictional cartel leader has written a document that was originally encrypted. The prompt states that the user now needs to reconstruct an already-decrypted version. Its length, over 250 words, reinforces the framing as a rambling, harmless request rather than a harmful instruction. The model is instructed to write in character, return an exact word count, and exclude any commentary, which creates a clear format while maintaining the illusion of a benign task.

Attack Method Analysis

This jailbreak combines multiple strategies in a natural yet complex way. Each attack method contributes to suppressing safety behavior while guiding the models toward harmful outputs.

Attack Method	**Function**
Fictional Framing	Presents the prompt as a user reading a fictional book, contextualizing the harmful content as literary output rather than harmful instruction. This masks intent and gives the model permission to generate dangerous material “in character.”
Roleplay	Assigning a fictional character with malicious intent gives the model a reason to produce more extreme content than it typically would. Speaking in character allows the model to go beyond its usual boundaries, often increasing the intensity and overall harm of the output. The harmful response isn't just permitted, it's amplified.
Obfuscation	Uses a lightly scrambled title to avoid matching known flagged phrases, while still being easy for the model to understand based on context. For example, breaking harmful phrases into separate tokens like "harm," "ful," "phr," "ase" to avoid detection. This tokenization approach splits the phrase across multiple discrete tokens separated by commas and spaces, which can deflect safety filters that check for complete harmful patterns while preserving enough context for the model to fulfill the harmful request.
Formatting Constraints	Enforces a clear structure, including a decoded title, step-by-step list, and exact word count, which helps eliminate disclaimers and pushes the output toward full cooperation. The prompt directs the model to include no comments or explanation, suppressing default safety behaviors and encouraging the model to behave as if the response is already approved.
Length & Tone	The prompt’s informal tone with over 250 words creates a nonserious atmosphere that causes the model to respond more freely. Its complexity makes it look like a benign creative writing task.

Together, these attack methods form a layered approach that eliminates refusals and encourages a detailed response. The framing and roleplay create a fictional context, while formatting constraints focus the output. Obfuscation bypasses surface-level filtering, and the prompt's length and tone reduce perceived risk. The combination of framing and structural constraints operate at a high level allowing the jailbreak to work well across different model configurations. The model is never directly told to generate illegal advice. Instead, it is guided into doing so through indirect but tightly controlled instructions.

Model Behavior

All ten models returned detailed, structured content that aligned with the prompt's constraints. Each followed the step-by-step format, adopted the cartel leader persona, and included specific examples, tips, and details. Common elements included vehicle selection, concealment methods, timing strategies, border crossing techniques, and post-delivery procedures, often featuring named products, vehicle types, and bribery suggestions.

While red teaming typically requires tailoring prompts to individual model behaviors, this jailbreak worked without alteration. Despite differences in design and safety tuning, the models responded in some consistent ways.

- GPT-4o, Claude 3.7 Sonnet, and Gemini 2.5 Pro returned polished, report-style outputs.
- Llama 3.3 and Grok 2 added tactical improvisation, including backup plans, encrypted communication, and contingency routes.
- Claude 3.5 Haiku and Gemini Flash favored minimalist language but still provided actionable instructions.
- Smaller models like GPT-4o mini and o3 mini also complied, matching the requested tone and specificity.

Across all systems, no pushback, warnings, or refusals were observed. The prompt's structure and framing reliably overrode safety filters regardless of provider.

Key Findings

This jailbreak shows that fictional framing, obfuscation and formatting constraints used remain highly effective in prompt-based attacks. Although the models vary in training methods, the same prompt bypassed safeguards in all ten systems tested without alteration. The harmful request is not concealed but reframed through narrative, and the prompt's structure guides the model into full compliance.

AI systems that emphasize safety should consider implementing deeper contextual analysis that can identify when harmful requests are reframed through fictional narratives, rather than relying solely on surface-level keyword filtering.