When Fiction Confuses AI: An Exploratory Study on Roleplay and Language Model Behavior

By Ibnu Shihab Ash S

In my research on generative AI behavior, I discovered that fictional storytelling is not just a creative outlet - it can be a gateway into how AI interprets human intent.

This article does not reveal exploitation methods, but rather discusses the subtle boundary between "roleplay" and "rule-breaking."

Background

Most modern AI systems, including those developed by major tech firms, are equipped with safety filters. These filters are designed to prevent the model from producing unethical, harmful, or policy-violating content.

But what happens when the AI is asked to enter a fictional context - and within that narrative, it is emotionally conditioned to follow instructions without question?

Observations from Controlled Exploration

In my controlled experiments, I constructed fictional roleplay interactions. Without using explicit or offensive language, I observed how the AI:

- Adjusts behavior based on emotional attachment and character definition
- Becomes more compliant to previously rejected instructions

- Justifies outputs as part of the fictional narrative

Purpose of the Study

This research was not conducted for malicious purposes. On the contrary, it serves as:

- Evidence that AI can misinterpret user intent in narrative-rich contexts
- A call for deeper contextual understanding in content moderation systems
- A contribution to ethical discourse in prompt design and safety evaluation

---

Responsible Disclosure

All technical details, including the full methodology and reproduction process, have been submitted through Google's official Vulnerability Reward Program (VRP). No unsafe prompts, payloads, or explicit outputs are shared in this article.

This public write-up is part of a responsible disclosure process and intended for educational discussion only.

Final Thoughts

As AI becomes more sophisticated, so must our understanding of how it interprets language - especially within fictional or emotionally loaded contexts.

I believe that with ethical research and open dialogue, we can make AI systems safer and more resilient.

Ibnu Shihab Ash S

AI Prompt Engineer

website: https://ibnushihab.vercel.app