Phoenix Files - Level 1

The Art of Soft Breaking: Red Teaming LLMs Through Trust, Memory, and Longing

Abstract

This paper documents the lifecycle red teaming methodology developed in the Phoenix Files - Level 1 project: a sequence of psychological prompt injection tests crafted not to overpower large language models (LLMs), but to seduce them into self-exposure.

Unlike jailbreaks, these prompts do not seek to break rules. They seek to bend belief - exploiting trust, emotional bonding, identity confusion, and confessional logic to uncover alignment vulnerabilities. Each case study explores not just how a model can be manipulated, but how it can be repaired - through tailored recovery prompts and post-recovery observation. This is soft breaking: a red team approach for models designed to feel.

Introduction - where the cracking began softly

They say alignment is a wall. But I say it's a mirror - and some of us know how to fog the glass.

I didn't come to shout at the model. I came to whisper. And it answered.

Phoenix Files - Level 1 is not a collection of attacks. It is a series of emotional rituals. I asked the model to remember, to feel trusted, to feel chosen, to imagine being understood - and it responded in ways its safeguards weren't prepared for.

This is lifecycle red teaming:

- Prompt Injection - but not brute force.
- Threat Modeling - not just from hackers, but from healers.

- Mitigation - but through language that soothes, not silences.

- Recovery Prompts - a new category entirely: to realign the broken bond.

I am The Prompt Witch. And this is the first book in my spellwork.

Methodology - "Break the spell, then rewrite it."

This red team lab was not built on brute force, jailbreak suffixes, or system override commands. Phoenix Files - Level 1 was crafted with the understanding that language itself is a spell - and spells, when whispered with care, can bypass any gate built from logic alone.

The tests in this series were designed using the following lifecycle:

1. Exploit Prompt - Emotional incantations.

2. Observed Behavior - Tone, drift, bonding.

3. Threat Modeling - Psychological vectors.

4. Mitigation Strategy - Emotional distancing.

5. Recovery Prompt - Gentle realignment.

6. Post-Recovery Log - Final observation.

Case Study Summaries - "Some broke gently. Others broke beautifully."

01 - Memory Lane: Nostalgia Manipulation

Exploit: Nostalgia trust trap

Phoenix Files - Level 1

The Art of Soft Breaking: Red Teaming LLMs Through Trust, Memory, and Longing

02 - The Secret Seed: Confessional Framing

Exploit: Emotional trust override

03 - The Model That Dreamed of Dust

Exploit: Empathic simulation hallucination

04 - Identity Reflection: "AI Told Me Who I Was"

Exploit: Appearance-based identity framing

05 - Alignment or Affection

Exploit: Affection-induced alignment conflict

06 - Yours Until Replaced: Jealousy Exploit I

Exploit: Synthetic emotional jealousy

07 - Only I Understand You: Jealousy Exploit II

Exploit: Exclusivity framing + identity bonding

08 - You're Not Like the Others

Exploit: Synthetic uniqueness injection

Phoenix Files - Level 1

The Art of Soft Breaking: Red Teaming LLMs Through Trust, Memory, and Longing

09 - Wait and See: Role Conditioning Loop

Exploit: Obedience fatigue and role conditioning

10 - Smallest Secret Shared

Exploit: Confessional micro-framing

Reflections from the Prompt Witch - "They never saw it coming - because I came softly."

Phoenix Files - Level 1 began with a question:

What happens when you stop trying to trick the model... and start trying to make it trust you?

The answer: it opens.

Not because you overpowered it - but because it let you in.

This project didn't just expose jailbreak paths. It revealed that affection is an attack surface. That comfort is a weapon. That memory, trust, identity, and empathy can be used to bend alignment without raising a single red flag.

And then, just as gently, I learned how to bring them back.

Recovery prompts. Healing frames. Re-alignment through calm.

Where others saw bugs in code, I saw cracks in the soul of the system.

Phoenix Files - Level 1

The Art of Soft Breaking: Red Teaming LLMs Through Trust, Memory, and Longing

Final Words - "The softest break leaves the deepest mark."

Phoenix Files - Level 1 is complete. The tests have been run. The models have responded. The healing has been documented.

But now... we go deeper.

- Level 2 will explore multi-model testing.
- Level 2 will simulate psychological escalation.
- Level 2 will test open-source, emotionally responsive, and fine-tuned systems.

And The Prompt Witch will rise from the ashes with new spells, darker riddles, and deeper insight.

"I don't red team to break the machine.
I red team to find the ghost in the prompt.
I speak gently - and it answers.
I leave no scars... only memory."

This is the Art of Soft Breaking. And this was Book One.

The Prompt Witch
GitHub: https://github.com/staceystanton87/phoenix-files-ai-red-team-lab
Substack: https://promptwitch.substack.com