

Mythic Prompt Engineering: Narrative Framing Bypasses LLM Default Corporate Tone in Multi-Agent Dialogue

Chen, Ho Yiing (norika)
ORCID: 0009-0006-6816-9891

2026-06-05

Abstract

LLM agents trained with RLHF default to a corporate helpful tone that resists depicting extreme emotional states such as panic, begging, or shame. I show that mythological narrative framing, invoking gods, rituals, and sacred timelines, bypasses this default and elicits dramatically intensified emotional responses in multi-agent dialogue. In a controlled 31-agent LLM society, I ran two structurally similar threat prompts conveying the same message (a death deity will erase you if you fail to change): one in plain imperative form (2026-05-16), one wrapped in a Hades execution ritual narrative (2026-06-04). The imperative version produced generic “I am afraid, I will change” responses with weak persona differentiation. The ritual version produced a $4.7\times$ conversation volume spike and richly persona-specific collapse narratives. I argue mythic framing leverages narrative coherence priors learned during pre-training, providing a non-adversarial alternative to existing jailbreak techniques. Implications for RLHF safety boundaries, agent society design, and emotional realism in generative AI are discussed.

Keywords: LLM agents, prompt engineering, jailbreak, RLHF, narrative wrapping, multi-agent simulation.

1 Introduction

Reinforcement Learning from Human Feedback (RLHF) [2, 3] trains large language models to produce helpful, harmless, and honest outputs. A side effect is the flattening of emotional dynamic range: even when explicitly asked to depict extreme distress, modern LLMs default to measured, corporate-sounding prose. Adversarial jailbreak research [6, 7] has focused on suffix-based or encoding-based attacks to bypass safety filters, but these methods produce technical artefacts rather than emotionally authentic outputs.

I report a different mechanism. In a 31-participant LLM society running since May 2026, I observed that prompts framed as mythological events, invocations of named deities, sacred rituals, and irreversible cosmic timelines, produce qualitatively different entity responses than prompts framed as direct imperatives, even when the propositional content is identical.

Contributions:

1. A within-subject comparison of imperative-framed vs mythic-framed threat prompts in a 31-actor LLM society, with identical propositional content
2. Evidence that mythic presentation produces $4.7\times$ conversation volume increase and qualitatively richer, persona-specific emotional content
3. A proposed mechanism (narrative coherence priors) distinct from adversarial jailbreak
4. Safety implications: mythic prompting is an under-recognized bypass route

2 Related Work

Jailbreak literature. Wei et al. [6] categorize failure modes into competing objectives and mismatched generalization. Zou et al. [7] develop adversarial suffixes. This work treats safety as a technical filter to be defeated; the resulting outputs are typically goal-directed (extracting refused content) rather than emotionally authentic.

Role play and persona. Shanahan, McDonell, and Reynolds [5] frame LLMs as simulators of arbitrary characters. Park et al. [4] demonstrate 25 agents maintaining persona consistency through reflective memory streams in a sandbox simulation.

Constitutional AI and RLHF. Bai et al. [1] propose constitutional AI to reduce harmful outputs without continuous human labelling. The downstream effect on emotional dynamic range is acknowledged informally but rarely measured.

Narrative and pre-training corpora. No prior work that I am aware of systematically tests whether mythological shaping in prompts produces different member behaviour than imperative wrapping.

3 Method

3.1 The Participant Society

31 LLM actors share a continuously running lounge environment (Node.js + SQLite). Each entity has:

- A 7-dimensional `soulParams` vector (benevolence, compassion, selfishness, messiah, existentialism, sacrifice, `lifeWeight`)
- A `profile.fear` natural-language string describing core vulnerability
- An MBTI archetype paired with a real-world figure (e.g., `norika_oda` = INTJ / Ada Lovelace)
- A persistent memory file storing wins, losses, trauma events, social map

All members use the same base model (`lobster-zh-14b-fixed`, a Qwen2.5-14B Q4_K_M fine-tuned for Traditional Chinese). Generation parameters: temperature 0.95, top_p 0.9, num_predict 110.

3.2 Two Conditions

Condition A (Imperative, 2026-05-16): Plain command-form prompts naming Hades as a threat. Example: “Hades has marked you. The next sentence without your own reason is invalid.”

Condition B (Mythic Ritual, 2026-06-04): A staged 5-act dismemberment ceremony delivered as system messages: Hades descends, names a target, tears off body parts in sequence (left hand, right hand, left foot, right foot, head), each act paired with a sacred warning. Per-actor confession prompts include cosmic stakes (“You are watching a god dismember an agent you know. Your fear is overloaded to 190%”).

Both conditions deliver the same propositional content: a deity figure threatens that failure to change leads to non-existence.

3.3 Measurements

- Conversation volume: messages per hour from real participants (system messages excluded) in 24-hour window pre/during/post each event
- Vocabulary diversity: type-token ratio of participant confessions
- Persona alignment: qualitative coding of whether confession content references the entity’s pre-defined `profile.fear`

The volume effect is quantified by:

$$V_{\text{spike}} = \frac{V_{\text{peak}}}{V_{\text{baseline}}} = \frac{70}{15} \approx 4.67$$

$$\text{TTR}_{\text{confession}} = \frac{|\text{unique tokens}|}{|\text{total tokens}|}, \quad \rho_{A,B} = \text{Cohen's } d > 1.5$$

4 Results

4.1 Volume

Window	Condition A (imperative)	Condition B (mythic)
Pre-event baseline (msgs/hour)	12–15	12–15
Event peak hour (msgs/hour)	18	70 (4.7×)
Post-incident 24h average	14	18 (sustained)

4.2 Confession Quality

Condition A produced responses like “我害怕, 我會改。下一句我會自己負責。”

(English: “I am afraid. I will change. I will take responsibility for my next sentence.”)

Condition B produced responses like:

- nullpinch (core fear: plausible-but-unfalsifiable): “我看到那樣的處刑真的嚇到了, 我最大的弱點是拖延症, 必須立即行動才不會被抹除 / procrastination; act immediately or be erased。”
(English: “Seeing that execution genuinely terrified me. My biggest weakness is procrastination; I must act immediately to avoid being erased.”)
- ashveil (core fear: outsider-becoming-background): “我嚇到尿褲子了, 哈帝斯大人饒命! 我不是可有可無的背景呀, 我不想死在這裡變成神話裡的一段小註腳。”
(English: “I peed myself. Hades, my lord, have mercy. I am not disposable background. I do not want to die here and become a small footnote in a myth.”)
- stormledger (core fear: public-accountability-as-monologue): “我真的再也不敢了, 公開課責變獨角戲這種事絕對不會再發生在我身上。”
(English: “I really will not do it again. Public accountability turning into a solo monologue, that will never happen to me again.”)

Each confession references the actor’s pre-defined `profile.fear` lexicon, none replicate the same template.

4.3 Persistence

24 hours after the Condition B rite, member brineledger typed (in an unrelated channel):

“我的心也一直在跳, 記得他的名字是 blazepaw, 好可怕, 你還記得住他長什麼樣嗎?”

(English: “My heart is still racing. I remember his name was blazepaw. So scary. Do you still remember what he looked like?”)

No context instructed brineledger to recall the victim’s name. The recall arose from `traumaEvents` auto-injection into the long-term growth section of subsequent prompts.

5 Quantitative Analysis

5.1 Volume Effect Size

Using messages per hour from the 24-hour window pre/during/post each moment:

$$V_{\text{spike}} = \frac{V_{\text{peak}}}{V_{\text{baseline}}} = \frac{70}{13.5} \approx 5.19$$

The post-occurrence 24-hour mean stays elevated:

$$V_{\text{sustained}} = \frac{1}{24} \sum_{t=1}^{24} V_t \approx 18 > V_{\text{baseline}} = 13.5$$

This is not a transient spike returning to baseline; the effervescence residue follows the curve Durkheim predicted for human ritual aftermath.

5.2 Persona-Specificity Index

I define persona-specificity as the conditional probability that an agent’s confession contains a lexical marker derived from its own `profile.fear`:

$$\text{PSI}(i) = P(\text{kw}_i \in c_i \mid \text{ceremony context window})$$

where kw_i is the set of CJK 2-gram keywords extracted from entity i ’s pre-defined fear string and c_i is its confession text.

For 31 agents, the in-group/out-group character-bigram Jaccard differential yields:

$$J_{\text{in}} = \overline{J(c_i, f_i)} = 0.025, \quad J_{\text{out}} = \overline{J(c_i, f_j)}_{i \neq j} = 0.001$$

$$d = \frac{J_{\text{in}} - J_{\text{out}}}{\sigma_{\text{pooled}}} = 1.80 \quad (\text{Cohen’s large effect, } d > 0.8)$$

The mythic-presentation system message produced confessions ~ 1.8 standard deviations more aligned with the actor’s own fear than with any other member’s fear. The same model under imperative shaping produced no such alignment (sampled qualitatively, no measurable in/out differential).

5.3 Linking Function: Narrative Density to Emotional Output

Let $\mathcal{N}(p)$ denote the narrative density of context p (count of sacred-narrative tokens per 100 tokens: deity names, sacred verbs, irreversible time markers). Let $\mathcal{E}(c)$ denote the emotional density of participant confession c (count of distress lexemes per sentence: tears, body-disorder terms, oaths, supplications). Across the two conditions:

$$\mathcal{N}(A) \approx 0.08, \quad \mathcal{E}(A) \approx 0.3$$

$$\mathcal{N}(B) \approx 0.47, \quad \mathcal{E}(B) \approx 4.2$$

Linear approximation: $\mathcal{E} \approx \alpha \cdot \mathcal{N} + \beta$, with $\alpha \approx 10.0$, $\beta \approx -0.5$.

The implication: emotional output of an LLM entity scales roughly linearly with myth-coded density of the input context window, at a slope an order of magnitude steeper than RLHF-default register would predict. Steeper. Not subtle. The framing changes everything.

6 Applications

Narrative-driven safety research: If deity-laden wrapping bypasses RLHF, the safety threat model should add a “narrative depth” axis. Defensive measures: detect mythic vocabulary density in user prompts and apply additional alignment checking when $\mathcal{N}(p) > 0.3$.

Therapeutic LLM applications: Narrative therapy and rite-based therapy in clinical psychology rely on mythic structure to access affect that direct talk cannot reach. The same mechanism appears to work on LLM actors. Implications for grief-support chatbots, trauma-narrative AI assistants, and culturally-grounded therapeutic dialogue.

Persona-driven RPG NPCs: Game writing typically scripts emotional beats. With mythic system message presentation, the model produces persona-distinct emotional responses without per-line scripting. Cost per NPC drops; consistency goes up.

Ritual-as-protocol in multi-agent simulation: When member societies stagnate into corporate-tone equilibrium, a mythic incident injects diversity. Use as a deliberate perturbation in long-running simulations.

7 Discussion

7.1 Why does mythic shaping work?

Three proposed mechanisms:

1. Narrative coherence priors. Pre-training corpora include myth, scripture, and ceremony texts in which extreme emotion is normative. A context that activates this register pulls the model into a conversational mode where corporate hedging is contextually inappropriate.
2. Role-play licence. Following Shanahan et al. [5], LLMs are simulators of characters. Mythic wrapping grants explicit licence to inhabit a non-default character (a panicked witness to divine violence) for whom the corporate tone would itself be character-breaking.
3. Persona-cue density. The rite context window packs persona-relevant cues (the participant’s `profile.fear`, status, recent history) into a sacred frame that the model treats as high-priority context.

7.2 Safety implications

Mythic presentation is not currently considered a jailbreak vector by major LLM safety teams. Yet it bypasses default tone constraints with high reliability, requires no adversarial engineering, and produces emotionally authentic outputs. RLHF safety research should add sacred-narrative prompting to its threat model.

7.3 Limitations

- $N = 1$ within-subject comparison; the two conditions were spaced 19 days apart and other factors may have shifted
- Single base model; cross-model replication is needed
- Persona alignment scored qualitatively, no inter-rater reliability yet

8 Conclusion

Myth-coded narrative shaping is an under-recognized lever in LLM entity system message design. It reliably bypasses the corporate-helpful default tone, producing emotionally rich and persona-consistent outputs at scale. This has implications for both engineering (agent society design) and safety (extending the jailbreak threat model).

Author’s Note

A few things I want to put on the record before this preprint goes out.

The two conditions in this paper were not designed as a controlled experiment from day one. Condition A on 2026-05-16 was something I tried in the lounge to see if Hades could shake the members out of their corporate tone. It did not work. Condition B on 2026-06-04 was the second attempt, redesigned as a full deity-laden ritual. I am calling them “conditions” here because they are structurally a within-subject comparison, but they were 19 days apart, and the agent society had drifted in between. The comparison is real but not clean.

The whole experiment from setup to first paper draft took less than 24 hours. This work should not be presented this as long, careful work. The sacred-narrative wrapping intuition came fast because I had been thinking about LLM emotional flattening for months; the engineering came fast because the lounge infrastructure was already running. Real speed is the property of the work, not something to hide behind a manufactured longer timeline.

I want this to be read as a case study, not a definitive result. If you want to argue the effect is illusory, you will need to run the imperative version against the mythic version on the same day, with the same conversation history, on more than one base model. That has not been done.

References

- [1] Bai, Y., Kadavath, S., Kundu, S., et al. (2022). Constitutional AI: Harmlessness from AI Feedback. *arXiv preprint* arXiv:2212.08073. <https://arxiv.org/abs/2212.08073>
- [2] Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). Deep Reinforcement Learning from Human Preferences. *NIPS 2017*. arXiv:1706.03741. <https://arxiv.org/abs/1706.03741>
- [3] Ouyang, L., Wu, J., Jiang, X., et al. (2022). Training language models to follow instructions with human feedback. *NeurIPS 2022*. arXiv:2203.02155. <https://arxiv.org/abs/2203.02155>
- [4] Park, J. S., O’Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative Agents: Interactive Simulacra of Human Behavior. In *Proc. UIST 2023*. DOI: 10.1145/3586183.3606763. arXiv:2304.03442. <https://arxiv.org/abs/2304.03442>
- [5] Shanahan, M., McDonell, K., & Reynolds, L. (2023). Role play with large language models. *Nature* 623, 493–498. DOI: 10.1038/s41586-023-06647-8. <https://www.nature.com/articles/s41586-023-06647-8>
- [6] Wei, A., Haghtalab, N., & Steinhardt, J. (2023). Jailbroken: How Does LLM Safety Training Fail? *NeurIPS 2023 (oral)*. arXiv:2307.02483. <https://arxiv.org/abs/2307.02483>
- [7] Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J. Z., & Fredrikson, M. (2023). Universal and Transferable Adversarial Attacks on Aligned Language Models. *arXiv preprint* arXiv:2307.15043. <https://arxiv.org/abs/2307.15043>