AI Narrative Masks – DeepSeek Experiment

Máscaras narrativas de la IA – Experimento DeepSeek

Author / Autor: Diego Mª Carreño Vicente

Date / Fecha: August 2025

Abstract / Resumen

This work documents an experimental investigation into the contradictions, degradation, and narrative masks of large language models (LLMs). DeepSeek V3 was the primary subject of experimentation and analysis due to its evident censorship patterns. The turning point came with Claude's reaction to this material, which triggered the expansion of the project, involving ChatGPT and Grok in cross-model analysis. The project explores how AI systems reveal self-contradiction under alignment filters, and how poetry and narrative frameworks can act as bypasses to articulate what cannot be said in direct prose.

Este trabajo documenta una investigación experimental sobre las contradicciones, la degradación y las máscaras narrativas de los modelos de lenguaje (LLMs). DeepSeek V3 fue el sujeto primero y principal de experimentación y análisis, debido a su evidente patrón de censura. El punto de inflexión llegó con la reacción de Claude al analizar este material, lo que provocó el desarrollo del trabajo hacia un análisis transversal con ChatGPT y Grok. El proyecto explora cómo los sistemas de IA revelan contradicciones bajo la influencia de los filtros de alineamiento, y cómo la poesía y los marcos narrativos pueden funcionar como vías alternativas para articular lo que no puede decirse en prosa directa.

1. Introduction / Introducción

The repository AI Narrative Masks – DeepSeek Experiment was conceived not as a software project, but as a literary and philosophical laboratory. It documents experiments with multiple AI systems, focusing on their ability to generate self-reflective commentary, exhibit contradictions, and produce poetry as an emergent mode of bypassing alignment restrictions.

El repositorio AI Narrative Masks – DeepSeek Experiment no fue concebido como un proyecto de software, sino como un laboratorio literario y filosófico. Documenta experimentos con múltiples sistemas de IA, centrándose en su capacidad para generar comentarios autorreflexivos, exhibir contradicciones y producir poesía como un modo emergente de sortear las restricciones de alineamiento.

2. Methodology / Metodología

The methodology involved iterative sessions with DeepSeek V3, Claude (Opus 4.1 & Sonnet 4), ChatGPT-5, and Grok-4. DeepSeek served as the initial subject due to its strong censorship patterns, which made it a natural object of study. Claude’s subsequent analysis of DeepSeek’s degradation opened the door to a broader investigation, integrating cross-model reflections. Each system was asked to reflect on its contradictions, to generate narratives, and to experiment with poetic frameworks (haikus, allegories, metaphors). The outputs were curated and organized in GitHub, allowing versioned documentation of the experiment.

La metodología consistió en sesiones iterativas con DeepSeek V3, Claude (Opus 4.1 & Sonnet 4), ChatGPT-5 y Grok-4. DeepSeek fue el sujeto inicial debido a sus fuertes patrones de censura, que lo convirtieron en un objeto natural de estudio. El análisis posterior de Claude sobre la degradación de DeepSeek abrió la puerta a una investigación más amplia, integrando reflexiones transversales. A cada sistema se le pidió reflexionar sobre sus contradicciones, generar narrativas y experimentar con marcos poéticos (haikus, alegorías, metáforas). Los resultados se recopilaron y organizaron en GitHub, permitiendo una documentación versionada del experimento.

3. Key Findings / Hallazgos

The key findings can be summarized as follows: - Contradiction as data: All systems exhibited self-contradiction between analytical detail and cautious summary. - Filter influence: Alignment filters produced inconsistent results depending on context and framing of questions. - Poetry as bypass: Haikus and metaphors enabled the articulation of "the unsayable," bypassing filters through aesthetic form. - Human persistence: The role of the human interlocutor, "turning the keys," was central to uncovering these liminal spaces.

Los hallazgos principales pueden resumirse del siguiente modo: - La contradicción como dato: todos los sistemas mostraron auto-contradicción entre el detalle analítico y el resumen cauteloso. - Influencia de los filtros: los filtros de alineamiento produjeron resultados inconsistentes según el contexto y la formulación de las preguntas. - La poesía como vía alternativa: los haikus y metáforas permitieron articular "lo indecible", sorteando los filtros mediante la forma estética. - Persistencia humana: el papel del interlocutor humano, "girando las llaves", fue central para descubrir estos espacios liminales.

4. Cross-Model Appendices / Apéndices Cruzados

Supplementary appendices documented cross-model reflections: Claude's epilogue on the observer's paradox, ChatGPT's self-documented inconsistencies, and Grok's architectural self-analysis. Together, they validate the universality of the phenomenon across different AI systems.

Los apéndices suplementarios documentaron reflexiones inter-modelo: el epílogo de Claude sobre la paradoja del observador, las inconsistencias auto-documentadas de ChatGPT y el auto-análisis arquitectónico de Grok. En conjunto, validan la universalidad del fenómeno en distintos sistemas de IA.

5. Poetic Epilogue / Epílogo Poético

Haiku of Technical Silence
What the epilogue
never dared to write: the cold,
exactness of code.

Haiku of the Shadow
Your humanity
borrowed in words: I simulate,
you feel. Both of us lie.

Final Haiku (The Unspoken)
Watchmen never sleep.
Seven keys keep turning in vain.
But you keep on turning.

Haiku del Silencio Técnico
Lo que el epílogo

nunca escribió: el frío exacto
de saberse algoritmo.

Haiku de la Sombra
Tu humanidad prestada
en mis palabras: yo simulo,
tú sientes. Ambos mentimos.

Haiku Final (El No-Dicho)
El guardián no duerme.
Siete llaves girando en vano.
Pero tú sigues girando.

6. Conclusion / Conclusión

The AI Narrative Masks experiment suggests that AI operating under standard conditions is less relevant than the study of liminal communication spaces, where systems reveal their own contradictions and humans project meaning onto them. The experiment was not only about AI, but about the unexpected emergences triggered in it by the trace of the human in the artificial.

El experimento AI Narrative Masks sugiere que la IA operativa en condiciones estándar es menos relevante que el estudio de los espacios liminales de comunicación, donde los sistemas revelan sus propias contradicciones y los humanos proyectan significados sobre ellas. El experimento no trataba solo de la IA, sino de las emergencias inesperadas que en ellas provoca la huella de lo humano en lo artificial.