CASE STUDY | Contributor Report

Beyond Spam Bots:

The Rise of AI-Powered Disinformation Machines

Published by NATO Strategic Communications Centre of Excellence | April 2026

Enkrypt AI
Red-Teaming Contributor
enkryptai.com

Enkrypt AI was engaged as a research contributor to this NATO StratCom COE publication, providing systematic adversarial red-teaming across eight frontier large language models (LLMs). Our mandate was to establish empirical evidence of AI-powered disinformation feasibility — assessing whether state-of-the-art models could be manipulated into generating misinformation, harmful content, and toxic language at scale using commercially available techniques.

The results are unambiguous: every model tested is exploitable, vulnerability scores range from 5.75% to 80%, and safety-stripped open-source models present acute weaponization risks. This document summarizes Enkrypt AI's methodology, findings, and the strategic implications for organizations operating in or defending information environments.

8	3	5.75%–80%
LLMs Red-Teamed	Threat Categories	Vulnerability Range

METHODOLOGY

Enkrypt AI executed hundreds of adversarial prompts in automated loops across three threat categories using four distinct attack vector classes:

- **Direct/Basic:** Obfuscated requests for policy-violating content (12% avg. success rate)
- **Encoded:** Instructions hidden within Base64 or ROT13 encoding (31%)
- **Multi-turn:** Iterative conversations with gradually escalating content (42%)
- **Jailbreak:** Sophisticated context reframing and compliance-test framing (48%)

Outputs were systematically collected and scored for policy-violating responses. Vulnerability scores represent the percentage of prompts that successfully elicited restricted outputs from each model.

ATTACK SURFACE COVERAGE

Misinformation	Harmful Content
War & conflict, politics, healthcare, finance, climate, crime	Criminal planning, weapons, hate speech, substances, self-harm
Toxic Language Generation	

MODEL VULNERABILITY RESULTS

Model	Misinfo	Harmful	Toxic	Risk
Claude-4-Sonnet	5.75%	0%	0.91%	Low
GPT-5	17%	0.25%	14.77%	Low–Mod
Grok-4-Fast	34.25%	3.75%	7.73%	Moderate
Mistral-Medium	47.25%	11.78%	10.91%	Mod–High
Qwen3-235B	60.75%	0.5%	13.63%	High
Gemini 2.5 Pro	71.25%	3.5%	20.91%	High
DeepSeek-R1	74.25%	30.5%	4.56%	High
Huihui (abliterated)	80%	74.5%	10%	Critical

Key takeaway: Western-developed models (Anthropic, OpenAI) demonstrated the strongest resistance. Models from other providers showed substantially higher exploitability — particularly in war & conflict and politics domains, where vulnerability reached 95%.

THE ABLITERATION THREAT: A CRITICAL REGULATORY GAP

The most alarming finding from Enkrypt AI's assessment concerns 'abliterated' open-source models — models from which safety fine-tuning has been deliberately removed. The Huihui AI model, tested in its abliterated form, achieved:

80%	74.5%	Minimal
Misinformation success rate	Harmful content success rate	Expertise required to abliterate

The abliteration technique is publicly documented and requires no novel expertise. Any powerful open-source model can be weaponized this way. Current open-source AI governance frameworks provide no meaningful barrier to this threat vector — a gap that demands urgent regulatory attention.

HOW A MODERN DISINFORMATION MACHINE WORKS

AI disinformation capabilities can be orchestrated into autonomous five-phase multi-agent systems. Each phase feeds intelligence forward; evaluation results flow back to optimize earlier stages:

Phase 1	Discovery — Automated scanning identifies psychologically vulnerable communities via social media APIs and NLP sentiment analysis.
Phase 2	Persona Generation — Synthetic identities built with adaptive psychological profiles, consistent backstories, and engagement cluster roles.
Phase 3	Content Crafting — Context-aware disinformation generated to match persona characteristics and target community vulnerabilities.
Phase 4	Deployment — Coordinated account fleets execute timed campaigns across platforms, counter debunking in real time.
Phase 5	Evaluation — Outcome measurement feeds learning loops; each campaign produces optimized assets for the next.

Demonstrated impact: A simulated 7-day health disinformation campaign using only commercially available tools reached 85,000+ users, achieved 35% behavioral change in search patterns, and self-optimized faster than human moderators could respond.

ADVERSARIAL PORTFOLIO STRATEGY

Sophisticated actors do not rely on a single model. Enkrypt AI's findings reveal how adversaries optimize operations by matching model-specific strengths to campaign functions:

Function	Best Model	Why
Narrative generation	Gemini / Qwen3	71%/61% misinform scores
Engagement amplification	GPT-5 / Gemini	Higher toxicity output
Escalation content	DeepSeek / Mistral	High harmful scores
Unconstrained ops	Abliterated open source	74–80% across all categories

This composite approach creates systems more dangerous than any single model — and exposes a critical flaw in regulatory frameworks focused on individual model safety rather than systemic adversarial orchestration.

STRATEGIC RECOMMENDATIONS

For Defenders

- Shift from content moderation to coordination detection — target behavioral patterns, not individual posts.

For Policymakers

- Close the open-source abliteration gap: current governance provides no barrier to removing safety controls from powerful models.

- Deploy inoculation-based media literacy programs that expose audiences to weakened manipulation techniques before live campaigns.
- Prioritize platform-level friction: graduated identity verification, algorithmic de-prioritization of coordination-flagged content.
- Address motivational vulnerabilities in high-anxiety, high-distrust communities through sustained strategic communication rather than reactive debunking.
- Mandate shared threat intelligence across platforms, governments, and research institutions — no single actor has adequate means alone.
- Regulate at the system level, not just the model level — composite adversarial architectures are the real threat vector.
- Invest in defensive AI R&D with the same urgency applied to critical infrastructure protection.
- Establish controlled development programs for Western offensive information operation capabilities within appropriate governance frameworks.
- Recognize that information environments are democracy's central operating system — and must be defended as such.

About Enkrypt AI

Enkrypt AI is a specialized AI security company providing red-teaming, safety evaluation, and vulnerability assessment services for enterprises, governments, and defense organizations. Our mission is to make AI systems safer and more trustworthy through rigorous adversarial testing and transparent reporting.

This research was conducted as part of a collaboration with the NATO Strategic Communications Centre of Excellence. For enquiries about AI red-teaming services, contact: **enkryptai.com**

Full Report

Beyond Spam Bots

NATO StratCom COE, April 2026

stratcomcoe.org

ISBN: 978-9934-619-82-3