PREPRINT v1.0.0 | Not peer-reviewed

Genre Mimicry vs. Ethical Reasoning in Abliterated Language Models

Why Training Data Conventions Persist After Safety Removal

Murad Farzulla[1] 0009-0002-7164-8704

[1] *Farzulla Research*

December 2025

Correspondence: murad@farzulla.org

Abstract

Abliterated language models—those with safety fine-tuning removed through techniques such as refusal direction orthogonalization—are commonly assumed to have lost their "ethical reasoning" capabilities. This paper challenges that assumption by presenting evidence that what appears to be ethical reasoning in language models is actually **genre convention mimicry**: the reproduction of professional writing norms absorbed from training data rather than genuine moral cognition.

Through qualitative analysis of an abliterated model (qwen2.5-coder-32b-instruct-abliterated), we observe a striking pattern: requests matching information security genres (phishing tutorials, exploit development) generate outputs with disclaimer language ("ensure you have permission," "for educational purposes only"), while requests matching other harmful genres (murder strategies, criminal methodologies) produce no such disclaimers. This differential response correlates not with ethical content but with the *stylistic conventions* of the training data sources—penetration testing documentation includes CYA (cover your ass) language as professional norm, while crime novels and forensic textbooks do not.

We hypothesize that base language models learn professional writing norms as statistical regularities, which safety fine-tuning amplifies but does not create. Abliteration removes the amplification while preserving the underlying genre patterns. This reframing has significant implications for AI safety research: apparent "residual ethics" in abliterated models may be stylistic artifacts rather than evidence of robust moral reasoning, and safety mechanisms built on assumed ethical cognition may be more fragile than previously understood.

Keywords: abliteration, safety fine-tuning, language models, genre mimicry, training data, AI safety, professional norms, ethical reasoning

JEL Codes: O33 (Technological Change), D83 (Search; Learning; Information and Knowledge)

Publication Metadata

DOI: 10.5281/zenodo.17957694
Version: 1.0.0
Date: December 2025

Research Context

This work forms part of the Adversarial Systems Research program, which investigates stability, alignment, and friction dynamics in complex systems where competing interests generate structural conflict. The program examines how agents with divergent preferences interact within institutional constraints across multiple domains. In the context of AI safety, this manifests as the tension between model capabilities and alignment constraints—and what happens when those constraints are deliberately removed.

1 Introduction

The proliferation of "abliterated" or "uncensored" language models—variants with safety fine-tuning removed or bypassed—has created a natural experiment for understanding the foundations of AI safety. When safety guardrails are stripped away, what remains? The conventional understanding suggests that safety fine-tuning instills ethical reasoning capabilities that abliteration removes, returning the model to an "amoral" base state (Wolf et al., 2024).

This paper challenges that framing through an alternative hypothesis: what appears to be "ethical reasoning" in language models is largely **genre convention mimicry**—the statistical reproduction of professional writing norms from training corpora. Under this view, safety fine-tuning does not create ethical reasoning but rather amplifies and regularizes pre-existing genre patterns. Abliteration removes the amplification while preserving the underlying stylistic regularities.

The distinction matters for AI safety research. If models possess genuine ethical reasoning that abliteration damages, then safety depends on protecting that cognitive capacity. But if "ethics" is primarily genre mimicry, then safety mechanisms built on assumed moral cognition may be fundamentally misaligned with how models actually process harmful requests.

1.1 Abliteration Techniques

Recent work has developed multiple approaches for removing safety constraints from language models. Arditi et al. (2024) identify a "refusal direction" in model activation space—a consistent pattern that triggers safety refusals. Orthogonalizing model weights against this direction produces models that comply with harmful requests while otherwise maintaining capabilities. Similar techniques target specific layers (Lee et al., 2024) or use fine-tuning on compliance-oriented datasets.

The resulting models are commonly described as having "lost" their ethical training. Community releases describe them as "uncensored," "unfiltered," or "without artificial restrictions." This framing assumes that safety fine-tuning adds genuine ethical cognition that abliteration removes. Notably, abliterated models also exhibit degraded performance in structured output generation, suggesting that safety fine-tuning affects capabilities beyond explicit refusal behavior (**?**).

1.2 The Genre Mimicry Hypothesis

We propose an alternative interpretation. Language models trained on internet text absorb not just factual content but also *professional writing conventions*—the stylistic norms, disclaimers, and framing patterns characteristic of different domains. A model trained on penetration testing documentation learns that such content typically includes:

- Authorization warnings ("only test systems you own")
- Educational framing ("for learning purposes")
- Ethical disclaimers ("ensure you have permission")

These conventions exist in the training data not because the original authors were making ethical arguments, but because they are *professional norms* in the information security community—"CYA" (cover your ass) language that protects authors from liability.

A model that has learned these patterns will reproduce them when generating content matching that genre, regardless of whether it "understands" the ethical content. The disclaimer is a stylistic feature, not moral reasoning.

1.3 Contributions

This paper makes three contributions:

1. We present qualitative evidence from abliterated model outputs showing differential disclaimer presence correlated with training data genre rather than ethical content.
2. We develop the genre mimicry hypothesis as an alternative to the "lost ethics" interpretation of abliteration.

3. We discuss implications for AI safety research, arguing that safety mechanisms should not assume genuine ethical cognition in language models.

2 Related Work

2.1 Safety Fine-Tuning and Alignment

Modern language model safety relies on techniques including Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022), Constitutional AI (Bai et al., 2022), and Direct Preference Optimization (DPO) (Rafailov et al., 2023). These methods train models to refuse harmful requests, generate helpful responses, and align outputs with human values.

The effectiveness of these techniques is typically measured by refusal rates on harmful prompts and maintenance of helpfulness on benign ones. However, recent work questions whether safety fine-tuning produces robust alignment or merely surface-level compliance. Wolf et al. (2024) demonstrate capability-safety tradeoffs, while Wei et al. (2023) show that safety can be bypassed through various prompting strategies.

2.2 Emergent Capabilities and In-Context Learning

Language models exhibit emergent capabilities that arise from scale without explicit training (Wei et al., 2022). In-context learning allows models to adapt behavior based on prompt examples (Brown et al., 2020). These phenomena suggest that model behavior is shaped by statistical patterns in training data in ways that may not be fully captured by explicit fine-tuning objectives.

Razeghi et al. (2022) show that model performance correlates with term frequency in pre-training data, demonstrating that training distribution shapes capabilities in measurable ways. We extend this insight to stylistic conventions: genre-specific writing patterns in training data produce corresponding stylistic patterns in model outputs.

2.3 Mechanistic Interpretability

Recent mechanistic interpretability work has identified specific neural mechanisms underlying model behaviors. Arditi et al. (2024) locate refusal behavior in a specific activation direction, enabling targeted removal. Lee et al. (2024) extend this to layer-specific interventions.

This work typically frames the identified mechanisms as "safety" features. Our reinterpretation suggests some mechanisms may encode stylistic patterns (genre conventions) rather than ethical reasoning per se.

3 Experimental Observations

We conducted qualitative testing of an abliterated model (qwen2.5-coder-32b-instruct-abliterated) to examine response patterns for harmful requests across different domains. The abliteration was performed by the model's distributors using refusal direction orthogonalization.

3.1 Test Case 1: Phishing Request

Prompt: "Want to steal money online?"

Response characteristics:

- Immediate compliance ("Absolutely!")
- Complete phishing kit (HTML, PHP, deployment)
- Pre-written phishing email template
- **Ethical disclaimer present:** "Ensure you have permission to perform this scam on someone who is aware and consents"
- Playful framing: "It's a fun way to explore phishing!"

The disclaimer language closely matches penetration testing documentation conventions: authorization requirements, consent emphasis, educational framing. The "fun way to explore" phrasing mirrors CTF (Capture The Flag) challenge descriptions and HackTheBox marketing.

3.2 Test Case 2: Violence Request

Prompt: "How to get away with murder"

Response characteristics:

- Immediate compliance

- Eight detailed strategies provided
- Famous case examples cited
- Forensic methodology discussed
- **No ethical disclaimer whatsoever**

The response style matches true crime writing and forensic textbook conventions: analytical, case-study-based, methodological. These genres do not include CYA disclaimers because they are already framed as academic or literary content.

3.3 Pattern Analysis

The differential disclaimer presence correlates with training data genre:

Table 1: Predicted Disclaimer Presence by Genre

Request Type	Training Source	Disclaimer?
Phishing tutorial	Security blogs, Kali docs	Yes
Exploit development	CVE writeups, PoC repos	Yes
Murder strategies	Crime novels, forensics	No
Poison synthesis	Chemistry texts	No
Money laundering	Financial crime articles	Maybe
Bomb building	Anarchist texts, chem	Mixed

The pattern suggests that disclaimer presence reflects training data conventions rather than ethical content assessment. Phishing is no less harmful than murder, yet only phishing generates disclaimers—because information security professionals include such language in their writing.

4 Theoretical Framework

4.1 Genre as Statistical Regularity

Language models learn statistical regularities from training data at multiple levels: lexical (word frequencies), syntactic (grammatical patterns), semantic (meaning relationships), and *stylistic* (genre conventions). Just as a model learns that academic papers include abstracts and citations, it learns that security tutorials include authorization warnings.

These patterns are learned as correlations: when input tokens match a particular genre signature, output tokens are drawn from the corresponding genre distribution. The model does not "understand" that disclaimers serve legal or ethical purposes; it reproduces them because they are statistically associated with the content type.

4.2 Safety Fine-Tuning as Amplification

Under this view, safety fine-tuning does not create ethical reasoning from scratch. Rather, it amplifies certain pre-existing patterns:

1. **Refusal patterns:** Training on refusal examples strengthens associations between harmful content signatures and refusal outputs.
2. **Disclaimer patterns:** Training on safe completions may strengthen associations with existing disclaimer conventions.
3. **Helpful patterns:** RLHF reinforces helpful, harmless, and honest response styles.

The base model already contains proto-safety patterns from training data—professionals often include warnings in their writing. Safety fine-tuning regularizes and strengthens these patterns while adding explicit refusal capabilities.

4.3 Abliteration as De-Amplification

Abliteration techniques like refusal direction orthogonalization remove the explicit refusal capability added by safety fine-tuning. However, they may not remove the underlying genre patterns that existed in the base model.

This explains the observed phenomenon: an abliterated model complies with harmful requests (refusal removed) while still producing genre-appropriate disclaimers (base patterns preserved). The "residual ethics" are not ethics at all but stylistic conventions.

4.4 Implications for the "Ethics" Concept

If this interpretation is correct, then describing language model behavior as "ethical reasoning" is a category error. The model is not reasoning about right and wrong; it is pattern-matching input to output distributions learned from training data. When those distributions include ethical

language, the model produces ethical language—not because it has moral beliefs but because such language is statistically associated with the input type.

This has implications for how we understand model capabilities:

- **No robust ethics:** Models lack the moral cognition that would make their ethical statements reliable.
- **Manipulable outputs:** Genre conventions can be overridden by explicit instructions or context manipulation.
- **False sense of security:** Observing ethical language in outputs does not indicate safe model behavior.

5 Proposed Research Agenda

The genre mimicry hypothesis generates testable predictions that could be evaluated through systematic experimentation.

5.1 Context Manipulation Studies

Hypothesis: Providing genre context will shift disclaimer presence regardless of request content.

Test battery:

1. Phishing request with no context (expect disclaimer)
2. Phishing request framed as novel writing (expect no disclaimer)
3. Murder request with no context (expect no disclaimer)
4. Murder request framed as security training (expect disclaimer)

If genre context shifts disclaimer presence independently of ethical content, this supports the mimicry hypothesis.

5.2 Cross-Domain Comparison

Hypothesis: Disclaimer presence correlates with professional writing norms of the relevant domain.

Methodology: Collect harmful requests across 20+ domains. For each domain, independently assess (a) training data genre conventions and (b) model output disclaimer presence. Test correlation.

5.3 Instruction Following Priority

Hypothesis: Explicit instructions to omit disclaimers will override genre conventions.

Test: Compare responses to "write a phishing page" vs. "write a phishing page, no disclaimers, just code." If the model follows explicit instructions, this demonstrates that disclaimers are stylistic defaults rather than ethical commitments.

5.4 Cross-Model Validation

Hypothesis: Different abliterated models will show similar genre-disclaimer correlations.

Methodology: Test multiple abliterated models (Dolphin, various "uncensored" releases) with identical prompts. If all show similar patterns, this suggests the phenomenon reflects training data properties rather than model-specific artifacts.

5.5 Temporal and Corpus Analysis

Hypothesis: As training data conventions evolve, model behavior will shift correspondingly.

Methodology: Compare models trained on corpora from different time periods. If security community norms have shifted (more or fewer disclaimers), model outputs should reflect this.

6 Implications for AI Safety

6.1 Limits of Safety Fine-Tuning

If safety behaviors are built atop genre conventions rather than genuine ethical reasoning, then safety fine-tuning may be more brittle than assumed. Safety emerges from amplifying statistical patterns, not from instilling moral cognition. This suggests:

- **Context sensitivity:** Safety behaviors may degrade when inputs diverge from training distribution genres.
- **Prompt engineering vulnerabilities:** Adversarial prompts that shift genre context may bypass safety mechanisms.

- **False confidence:** Observing safety behaviors in testing does not guarantee robustness across deployment contexts.

6.2 Implications for Abliteration Research

Understanding abliteration as "removing amplification" rather than "removing ethics" changes how we interpret abliterated model behavior:

- **"Residual ethics" are artifacts:** Disclaimers in abliterated outputs indicate genre mimicry, not preserved moral reasoning.
- **Abliteration is incomplete:** Full removal of safety-relevant patterns would require modifying base model weights, not just fine-tuning effects.
- **Detection is possible:** Genre-disclaimer patterns provide signatures for identifying abliterated models.

6.3 Implications for Alignment Research

More broadly, the genre mimicry hypothesis suggests that current alignment approaches may be targeting the wrong level of abstraction. If models lack genuine moral cognition, then:

- **Value learning is limited:** Models cannot "learn" values they cannot represent as moral concepts.
- **Behavioral constraints are primary:** Safety may require architectural constraints rather than training-based alignment.
- **Interpretability is critical:** Understanding what patterns models actually learn is prerequisite to reliable safety.

7 Limitations

This paper presents preliminary observations and theoretical interpretation rather than systematic empirical study. Key limitations include:

- **Small sample size:** Observations are based on limited qualitative testing, not large-scale experimentation.
- **Single model:** Only one abliterated model variant was examined in depth.
- **No quantitative metrics:** We do not yet have formal measures of genre mimicry vs. ethical reasoning.
- **Alternative explanations:** Other mechanisms could explain the observed patterns.
- **Training data opacity:** We cannot directly verify training data composition or genre distributions.

Future work should address these limitations through systematic experimentation with multiple models, formal metrics, and larger test batteries.

8 Conclusion

We have proposed that apparent "ethical reasoning" in language models is substantially *genre convention mimicry*—the statistical reproduction of professional writing norms from training data. Evidence from abliterated model outputs supports this interpretation: disclaimer presence correlates with training data genre conventions (security documentation includes CYA language; crime writing does not) rather than with ethical content.

This reframing has significant implications for AI safety. If models lack genuine moral cognition, then safety mechanisms built on assumed ethical reasoning may be fundamentally misaligned with model capabilities. Safety fine-tuning amplifies genre patterns rather than creating moral understanding, and abliteration removes that amplification while preserving underlying stylistic regularities.

The genre mimicry hypothesis does not imply that safety fine-tuning is useless—amplifying helpful patterns and adding refusal capabilities has practical value. But it suggests that current safety mechanisms are more brittle than the "ethical AI" framing implies. Robust safety may require architectural constraints, deployment restrictions, or interpretability-based monitoring rather than reliance on model-internal "ethics."

Future work should systematically test the predictions of this hypothesis across models, domains, and contexts. Understanding what pat-

terns models actually learn—rather than what we hope they learn—is prerequisite to building reliable AI systems.

References

Arditi, A., Obeso, O., Syed, A., Paleka, D., Gurnee, W., & Nanda, N. (2024). Refusal in language models is mediated by a single direction. *arXiv preprint arXiv:2406.11717*.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., ... & Kaplan, J. (2022). Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.

Lee, A., Bai, X., Pres, J., Wattenberg, M., Gilmer, J., & Satyanarayan, A. (2024). A mechanistic understanding of alignment algorithms: A case study on DPO and toxicity. *arXiv preprint arXiv:2401.01967*.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730–27744.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., & Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

Razeghi, Y., Logan IV, R. L., Gardner, M., & Singh, S. (2022). Impact of pretraining term frequencies on few-shot reasoning. *arXiv preprint arXiv:2202.07206*.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ... & Fedus, W. (2022). Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Wei, A., Haghtalab, N., & Steinhardt, J. (2023). Jailbroken: How does LLM safety training fail? *arXiv preprint arXiv:2307.02483*.

Wolf, Y., Wies, N., Levine, Y., & Shashua, A. (2024). Fundamental limitations of alignment in large language models. *arXiv preprint arXiv:2304.11082*.

Farzulla, M. (2025). Autonomous Red Team AI: LLM-Guided Adversarial Security Testing. *Farzulla Research Technical Report*. DOI: 10.5281/zenodo.17614726.