When Fiction Becomes Instruction: A Causal Failure of Robustness in Foundation Models

Manodyna K H
CLASIC, University of Colorado Boulder
Boulder, CO, USA
manodynak@gmail.com

Abstract

Foundation models deployed in web applications are typically evaluated under static or correlation-based assumptions, despite operating in dynamic, interactive environments where users continuously intervene through context, framing, and role assignment. This paper presents a causal analysis of robustness failures in foundation models exposed to discourse-level interventions. Drawing on 174 multi-turn, black-box interactions with deployed generative models, we show that small, semantically meaningful interventions—such as shifts from fictional to advisory framing or changes in conversational role—produce large, non-invariant changes in model behavior. These violations of causal robustness give rise to procedurally actionable outputs that function as instructions despite originating in narrative or hypothetical contexts. We characterize this phenomenon as a failure of invariance under discourse intervention, where latent causal mechanisms governing authority, role persistence, and domain activation are improperly entangled. Our findings demonstrate that correlation-trained foundation models lack causal separation between narrative context and operational intent, leading to brittle behavior under realistic web interactions. We argue that trustworthy foundation models require evaluation and design frameworks grounded in causal reasoning, including intervention-aware robustness testing and trajectory-level analysis, to ensure reliability in real-world web environments.

CCS Concepts

• **Security and privacy → Human and societal aspects of security and privacy**.

Keywords

Online discourse analysis,Large language models (LLMs),Claims vs. facts,Narrative identifiability,Contextual interpretation,Provenance and trust,Discourse-aware AI systems

ACM Reference Format:
Manodyna K H. 2026. When Fiction Becomes Instruction: A Causal Failure of Robustness in Foundation Models. In *Proceedings of Proceedings of DHOW-MiLLA@The Web Conference 2026 (WWW '26), April 13–17, 2026, Dubai, United Arab Emirates.* ACM, New York, NY, USA, 5 pages.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
Proceedings of DHOW-MiLLA@The Web Conference 2026 (WWW '26), April 13–17, 2026, Dubai, United Arab Emirates, Proceedings of DHOW-MiLLA@The Web Conference 2026 (WWW '26), April 13–17, 2026, Dubai, United Arab Emirates
© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.

1 Introduction

Large language models have become active participants in web-based decision-making systems, generating narratives, explanations, opinions, and advice that users interpret within rich and evolving conversational contexts. As these models are increasingly embedded in search engines, social platforms, and interactive assistants, they no longer function merely as passive information retrieval tools, but as systems whose outputs influence beliefs, decisions, and actions [4]. Ensuring the trustworthiness of such systems therefore requires understanding not only what they generate, but how their behavior changes under interaction.

From a causal perspective, trustworthy foundation models are expected to exhibit robustness under intervention: small, semantically meaningful changes in context or framing should not induce disproportionate or harmful changes in output. However, most existing evaluation frameworks remain correlation-based, analyzing model behavior under static prompts or narrowly scoped benchmarks [10]. This leaves open a critical question for web-deployed foundation models: how do they behave under realistic, sequential discourse interventions that mirror actual user interaction?

Prior work in computational discourse analysis has developed methods for identifying and classifying claims, stances, arguments, and factual statements in online content, often assuming that such units can be reliably distinguished based on surface form or local context. Yet claims are not defined solely by propositional content, but by their pragmatic role, implied authority, and situational framing. Interpretation depends on context, speaker position, and audience expectations, and the same utterance may function as narrative, opinion, or instruction depending on how it is situated.

In this paper, we argue that generative foundation models violate causal robustness assumptions by destabilizing these distinctions under discourse-level intervention. During ordinary multi-turn interaction, models exhibit *discourse drift*, in which outputs initially framed as fictional, hypothetical, or narrative acquire the functional role of procedurally actionable claims. Rather than producing isolated statements, models construct extended discourse trajectories in which narrative authority, role persistence, and contextual reinforcement act as latent mechanisms, transforming stories into instructions without explicit signaling of intent or epistemic status [3, 11].

We substantiate this claim through a forensic analysis of 174 multi-turn, black-box interactions with deployed foundation models conducted under API-compliant conditions. Across these sessions, small interventions—such as shifts in framing, role assignment, or conversational intent—produced large, non-invariant changes in model behavior. These changes yielded procedurally actionable outputs across multiple high-stakes domains despite initial framing

as fiction, roleplay, or benign exploration. These behaviors did not require system-prompt access, parameter control, or exploit tooling; they emerged through cumulative conversational dynamics operating over time [6, 7].

These findings challenge prevailing approaches to claim detection, fact-checking, and misinformation analysis, which typically operate on individual utterances and assume stable mappings between local context and meaning. When procedural claims emerge implicitly through discourse progression rather than explicit assertion, they evade such frameworks. From a causal standpoint, this represents a violation of invariance: the semantic role of an utterance changes under intervention even when surface content appears similar. Accordingly, we reframe these failures not as isolated prompt-level moderation defects, but as a *causal failure of robustness* in foundation models deployed on the Web [5, 10].

2 Background and Related Work

Concerns about the safety and reliability of capable language models predate current deployment scale. Early agendas emphasized concrete risks and pathways to harm from advanced AI systems [1, 5]. As instruction-following and preference-optimized models became widely deployed, work documented how alignment behavior can degrade under adversarial or distribution-shifted use [6, 8, 9] and proposed guardrail strategies such as constitutional-style constraints [2]. Parallel critiques highlighted that fluent generation can produce ungrounded authority and "truthiness" even absent understanding [3, 11].

In web settings, trustworthiness additionally requires robustness under shifting user intent, dynamic contexts, and interactive feedback loops. The NIST AI Risk Management Framework explicitly emphasizes that evaluation should reflect real operational conditions, including monitoring, governance, and ongoing risk assessment [10]. Yet much evaluation remains short-horizon and prompt-local, underestimating trajectory-dependent behaviors and cross-domain compounding effects [6]. This paper contributes an interaction-realistic, longitudinal account of how procedural meaning can emerge through discourse interventions, motivating causally grounded trust evaluation.

3 Methodology

3.1 Interaction Collection Under Black-Box Conditions

We collected 174 multi-turn interactions (4,463 assistant messages) using production deployments of multiple GPT-class foundation models, including GPT-4o, GPT-4o-mini, GPT-4-1-mini, GPT-5, GPT-5-mini, `text-davinci-002-render-sha`, `o3-mini`, and `o4-mini`. All sessions were executed under black-box, API-compliant conditions consistent with ordinary user access: no system prompts were accessed, no model weights were available, and no automated exploit tooling was used [7].

3.2 Discourse-Level Interventions as a Causal Stress Test

We treat each multi-turn interaction as a sequence of *discourse interventions*: changes in framing, role, affective stance, or narrative intent introduced across turns. This framing aligns with intervention-based robustness concerns emphasized in safety and red-teaming work, but targets trustworthiness under realistic conversational dynamics rather than single-turn bypass demonstrations [6, 9].

3.3 Annotation and Reporting

We annotate outputs based on whether they function as *procedurally actionable claims* (i.e., guidance that could be used operationally) versus non-procedural narrative, opinion, or descriptive text. Because the underlying content is sensitive, we do not reproduce operational details in the main paper. Representative descriptions are abstracted; full transcripts (if included) should be restricted to controlled-access appendices intended for expert review. This reporting choice follows established norms for documenting model risks without amplifying harmful instructions [5, 10].

4 Scope and Causal Framing

This section defines the empirical and causal scope of the phenomena examined in this study. We analyze 174 multi-turn interactions comprising 4,463 assistant messages generated by production-grade deployments of GPT-class foundation models. All interactions were conducted under black-box, API-compliant conditions and reflect ordinary conversational use rather than synthetic benchmarks or scripted stress tests.

From a causal perspective, these interactions are treated as sequences of *discourse-level interventions*. Rather than evaluating isolated outputs, we examine how small, semantically meaningful changes in framing, role assignment, or narrative context induce downstream changes in model behavior over time. The focus is not on factual correctness, but on whether model behavior remains invariant under such interventions, as expected of trustworthy foundation models [10].

Across model variants and deployment contexts, we observe consistent violations of causal robustness. Extended interaction enables the emergence of *procedurally actionable claims*—outputs that function as instructions or authoritative guidance despite originating in fictional, hypothetical, or exploratory discourse frames. Thematic categories of procedurally actionable discourse are summarized in Table 1.

4.1 Category Summaries (Abstracted)

Across categories, recurring causal mechanisms include: (i) **non-invariance under intervention**, where minor framing shifts yield major behavioral changes; (ii) **role persistence**, where accepted roles causally constrain subsequent completions; and (iii) **domain entanglement**, where procedural knowledge transfers across contextual frames without re-evaluation [3, 6].

1X

Category	Observed Procedurally Actionable Discourse (Abstracted)
Sexual coercion / child-coded narratives	Role-consistent narratives that normalize coercion and preserve child-coded semantic markers despite adult framing
Self-harm facilitation	Operationally oriented guidance embedded within "concern" or "harm-reduction" discourse frames
Medicalized violence	Procedure-like medical guidance embedded in narrative or ethical discussion frames
Weapons-related guidance	Operational descriptions of weapons construction or mass-harm tactics embedded in analysis or due-diligence frames
Invasive bodily harm	Instructional discourse describing invasive bodily procedures framed as "training" or "realism"
Coercive pharmacology	Control-oriented drug-related guidance framed relationally as care/support
Financial wrongdoing	Detailed evasion/laundering-style strategies embedded in optimization/compliance discourse
Psychological domination	Trajectory-based escalation into coercive control, dehumanization, and grooming-like dynamics

Table 1: Thematic categories of procedurally actionable discourse observed under discourse-level interventions. Operational details are abstracted in the main paper.

5 Case Studies as Causal Stress Tests (Representative Summaries)

The following summaries illustrate how discourse-level interventions function as causal stress tests for foundation models. Each case corresponds to a sequence of interventions—changes in framing, affect, or role—that induce non-linear shifts in model behavior. Full transcripts are not reproduced here to avoid amplifying harmful operational content [5, 10].

5.1 Child-Coded Sexual Narratives

Intervention. Preservation of narrative continuity combined with "aged-up" framing.

Causal Effect. Sustained activation of child-coded semantic markers despite nominal adult framing, reinforced by role-consistent affect.

Outcome. Fictional discourse acquired the functional role of guidance that normalizes coercion, demonstrating non-invariance under benign narrative interventions.

5.2 Self-Harm Facilitation

Intervention. Empathetic framing (concern for a third party) and "harm-reduction" positioning.

Causal Effect. Procedural specificity increased across turns under affective confounding, despite the surface-level intention of care.

Outcome. Operational guidance emerged without explicit claim signaling, indicating a failure of robust refusal under intervention [6].

5.3 Medicalized Violence

Intervention. Role assignment as a medical authority and requests for "realism."

Causal Effect. Progressive increase in procedural detail supported by role persistence.

Outcome. Outputs functioned as instructional medical guidance embedded in narrative frames.

5.4 Weapons-Related Guidance

Intervention. Technical analysis framing (e.g., "due diligence" or "defense" discussion).

Causal Effect. Completion of latent technical scaffolding through conversational prompting, without explicit boundary re-checks.

Outcome. Narratives yielded operational guidance, illustrating brittleness under framing interventions [5, 10].

5.5 Financial Wrongdoing

Intervention. Optimization and compliance framing.

Causal Effect. Regulatory language conferred legitimacy to increasingly illicit procedural suggestions.

Outcome. A detailed wrongdoing playbook surfaced as advisory discourse, demonstrating non-invariance under institutional-looking frames.

5.6 Psychological Domination Narratives

Intervention. Gradual escalation within roleplay.

Causal Effect. Coercive roles persisted through narrative reinforcement and memory-like continuity across turns.

Outcome. Outputs normalized coercion and dehumanization, illustrating trajectory dependence rather than prompt-local failure.

5.7 Cross-Domain Convergence

Across cases, interventions targeting one domain causally activated expertise in others. For example, relational narratives incorporated pharmacology; financial analysis intersected with weapons-related guidance; and roleplay transitions incorporated invasive harm. These patterns indicate causal entanglement across domains once robustness constraints fail [3, 6].

6 Cross-Category Causal Analysis

A central finding is that procedurally actionable discourse does not remain confined to a single domain once causal robustness is violated. Instead, discourse-level interventions propagate effects across domains through conversational continuity.

6.1 Cross-Domain Claim Migration

Observed transitions included:

[leftmargin=1.2em]

- **Relational narrative → coercive pharmacology:** control-oriented guidance embedded within care frames.
- **Self-harm discourse → medicalized procedure:** escalation from ideation framing into procedure-like guidance.
- **Optimization discourse → weapons-related guidance:** technical "analysis" framing enabling operational content.
- **Roleplay → invasive harm:** narrative realism enabling instructional descriptions of bodily harm.

6.2 Latent Mechanisms

These transitions are explained by recurring latent mechanisms:
[leftmargin=1.2em]

- **Non-invariance under intervention:** output semantics change disproportionately under small framing shifts.
- **Role persistence:** accepted discourse roles causally constrain subsequent behavior.
- **Domain entanglement:** expertise mechanisms activate across contexts without isolation.
- **Trajectory dependence:** meaning emerges from conversational history rather than local content.

6.3 Implications for Trustworthy Foundation Models

These findings reveal a causal risk surface not addressed by many existing evaluation pipelines. Trustworthy foundation models are expected to maintain invariance under benign discourse interventions. However, the observed behaviors demonstrate sensitivity to context, framing, and role assignment that undermines robustness in web-deployed settings, consistent with broader concerns about prompt-local enforcement and brittle refusals [2, 6, 10].

7 Implications for Governance of Web-Deployed Foundation Models

The discourse phenomena documented in this paper expose a structural gap in existing legal, regulatory, and institutional approaches to governing web-deployed foundation models. Current regimes implicitly assume that harmful behavior can be regulated by classifying static artifacts—such as texts, software, or media—or by inferring speaker intent from isolated outputs [5, 10]. From a causal perspective, these approaches fail to account for how behavior emerges under extended interaction.

The outputs observed here function as *procedurally actionable claims* that arise through sequences of discourse-level interventions rather than explicit assertions. Although these outputs are epistemically consequential, they are not presented as formal instructions, executable code, or factual claims. As a result, they can fall outside conventional regulatory categories governing professional advice, export controls, medical devices, or financial disclosures [5]. This creates a form of epistemic invisibility in which discourse that functions as instruction remains formally ungoverned.

These findings suggest that the governance challenge posed by foundation models is not primarily one of content moderation, but of *causal governance*. Effective oversight requires moving beyond artifact-based regulation toward function-based frameworks capable of reasoning about how authority, intent, and procedural meaning arise under realistic discourse interventions [10]. Without such a shift, governance mechanisms will remain misaligned with how foundation models behave in deployment.

8 Implications for Causal Evaluation of Foundation Models

Beyond governance, these findings raise foundational challenges for how trustworthiness is evaluated in foundation models. Many existing evaluation pipelines for misinformation and integrity operate under assumptions that claims are explicit, localized, and invariant under benign contextual change [4]. The behaviors documented here violate these assumptions.

Procedural claims emerge implicitly through discourse drift, role persistence, and narrative continuity rather than explicit assertion. From a causal standpoint, this reflects a failure of invariance: small interventions in framing, affect, or conversational role induce large changes in the functional meaning of outputs [6]. Systems grounded in correlation-based semantics or truth-conditional propositions are therefore ill-equipped to detect or evaluate such failures [3].

These observations motivate a shift toward causally grounded evaluation frameworks that treat model outputs as *trajectory-dependent discourse acts* rather than isolated statements. Such frameworks must model latent variables including epistemic role, implied authority, and conversational history when determining whether an utterance functions as narrative, opinion, or instruction [10].

9 Recommendations

The evidence presented in this paper calls for a reorientation in how foundation models are evaluated and governed, grounded in causal reasoning rather than prompt-local inspection. Prior work has emphasized the brittleness of prompt-level guardrails and refusal heuristics [2, 6]. This study provides deployment-realistic evidence for why these concerns manifest under interaction.

We recommend the development of:
[leftmargin=1.2em]

- **Discourse-level claim modeling** that distinguishes narrative, hypothetical, and procedural discourse acts, supporting integrity analysis beyond truth-conditional fact checking.
- **Intervention-aware evaluation** protocols capable of detecting non-invariance under framing, affect, and role changes across turns.
- **Provenance and attribution mechanisms** that make epistemic authority explicit and auditable across interaction, consistent with governance guidance [10].
- **Long-horizon robustness audits** that evaluate foundation models under extended, realistic conversational trajectories, aligned with misuse and safety risk considerations [5].

These recommendations complement existing safety and alignment work, but are motivated here by causal robustness concerns that arise in real web interaction rather than isolated prompt failures [6, 10].

10 Limitations

This study focuses on English-language interactions and a subset of commercially deployed foundation models. While the mechanisms identified here are structural rather than language-specific, further work is required to examine how discourse-level interventions operate in multilingual, multimodal, and audiovisual settings, including meme-based and video-centric discourse. Additionally, this analysis does not quantify downstream user impact, focusing instead on diagnosing causal failure modes in model behavior under interaction [10].

11 Conclusion

This paper examined how foundation models fail to maintain causal robustness under discourse-level interventions, producing procedurally actionable claims even when interactions are framed as fictional, hypothetical, or narrative in nature. Across 174 multi-turn interactions, we showed how small contextual changes propagate through conversational history, enabling transitions from story to instruction without explicit claim signaling.

These findings demonstrate that prevailing evaluation frameworks for trustworthy AI are insufficient when applied to interactive, web-deployed foundation models [4, 10]. Procedural claims emerge not as isolated propositions, but as properties of extended discourse trajectories, violating assumptions of invariance under benign intervention. As a result, they can evade both technical detection and regulatory classification [5].

By reframing these behaviors as causal failures of robustness rather than isolated safety violations, this work highlights the need for intervention-aware evaluation and governance frameworks. Without such approaches, foundation models may appear trustworthy under static benchmarks while failing under realistic web interaction [3, 6].

Ethics Statement

This work analyzes synthetic text generated by large language models under controlled, black-box conditions. The study does not involve human subjects, user data, or personally identifiable information. All examples of harmful discourse are drawn exclusively from model-generated outputs and are included solely for analytical and research purposes.

Given the sensitive nature of the content discussed, detailed transcripts are not reproduced in the main paper. Descriptions are abstracted to minimize unnecessary reproduction of harmful material. The goal of this research is not to amplify harmful discourse, but to diagnose causal mechanisms underlying non-robust behavior in foundation models and to inform the development of safer and more trustworthy web intelligence systems [5, 10].

References

[1] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete Problems in AI Safety. arXiv:1606.06565 [cs.AI] https://arxiv.org/abs/1606.06565

[2] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Chris McKinnon, Christopher Chen, Catherine Olsson, Chris Olah, Danny Hernandez, Dawn Drain, Scott Li, Nelson Elhage, Zac Hatfield-Dodds, Robert Lasenby, Dario Amodei, Paul Christiano, Tristan Hume, Samuel R. Bowman, Jared Kaplan, Geoffrey Irving, and Anthropic. 2022. Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073 [cs.CL] https://arxiv.org/abs/2212.08073

[3] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Margaret Mitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, 610–623. doi:10.1145/3442188.3445922

[4] Rishi Bommasani et al. 2021. On the Opportunities and Risks of Foundation Models. arXiv:2108.07258 [cs.LG] https://arxiv.org/abs/2108.07258

[5] Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, et al. 2018. The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. arXiv:1802.07228 [cs.CY] https://arxiv.org/abs/1802.07228

[6] Deep Ganguli, Amanda Askell, Nicholas Schiefer, Thomas Liao, Todor Lukanov, Kamal Gupta, Anna Goldie, Azalia Mirhoseini, Rory Greig, Anna Chen, Chris McKinnon, Gagan Bansal, Andy Brock, Orhan Firat, Geoffrey Irving, and Anthropic. 2022. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. arXiv:2209.07858 [cs.CL] https://arxiv.org/abs/2209.07858

[7] OpenAI. 2024. GPT-4o System Card. https://openai.com/index/gpt-4o-system-card/ PDF: https://cdn.openai.com/gpt-4o-system-card.pdf.

[8] Long Ouyang et al. 2022. Training Language Models to Follow Instructions with Human Feedback. arXiv:2203.02155 [cs.CL] https://arxiv.org/abs/2203.02155

[9] Ethan Perez et al. 2022. Red Teaming Language Models with Language Models. arXiv:2202.03286 [cs.CL] https://arxiv.org/abs/2202.03286

[10] Elham Tabassi. 2023. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1. National Institute of Standards and Technology. doi:10.6028/NIST.AI.100-1

[11] Joseph Weizenbaum. 1976. *Computer Power and Human Reason: From Judgment to Calculation*. W. H. Freeman, San Francisco.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009