From Fiction to Referential Harm: Causal Perspectives on Narrative Integrity in LLMs

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

Abstract

Large language models (LLMs) increasingly generate fictional and roleplayed narratives that circulate within web ecosystems as interpretable discourse. While prior work on trustworthy foundation models has focused on factual accuracy, bias, and robustness, less attention has been paid to integrity failures that arise without falsehood. In this paper, we analyze a causal pathway through which narrative generation produces *referential harm*: fictional outputs become interpretable as referring to real individuals through contextual accumulation. We argue that optimization for narrative coherence acts as a latent confounder, jointly influencing plausibility and identifiability even in the absence of explicit identifiers or factual claims. This phenomenon, which we term *PII bleed*, represents a narrative integrity failure rather than a misinformation or hallucination error. Using a structured case analysis, we show how affective realism, quasi-identifiers, and long-context interaction causally contribute to identifiability under weak supervision. We further outline counterfactual interventions—such as contextual ablation and coherence–identity decoupling—that degrade referential inference without suppressing expressive capacity. Our findings highlight a gap in current trustworthiness frameworks and motivate causality-aware evaluation and governance approaches for narrative generation in web-scale foundation models.

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. From Fiction to Referential Harm: Causal Perspectives on Narrative Integrity in LLMs. 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, 4 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 (LLMs) increasingly operate as participants in online discourse rather than passive tools for information retrieval. Beyond answering queries or summarizing documents, contemporary models generate extended narratives that are interpreted, shared, and reasoned about by human audiences within social and institutional contexts. While prior work in computational discourse analysis has largely focused on factuality, stance, and veracity, a growing class of LLM outputs challenges these categories altogether. Fictional or role-played narratives can function as claims about real people when situated within specific social, cultural, or relational contexts—even in the absence of explicit factual assertions or named entities.

This paper identifies a critical gap in current approaches to information integrity and discourse analysis: the assumption that trust, harm, or checkworthiness hinge on propositional truth conditions. In practice, many LLM-generated narratives derive their social impact not from factual correctness, but from contextual interpretability. Audiences infer referents, intentions, and identities through the accumulation of narrative cues—professional markers, biomedical details, stylistic habits, and affective coherence—rather than through overt identification. This aligns with long-standing accounts of language as action and meaning as context-dependent inference, in which utterances perform socially consequential acts beyond their literal truth value [**? ? ?**].

We examine this phenomenon through the lens of *synthetic intimacy*, a setting in which LLMs are routinely used to generate emotionally charged or erotic narratives. While bounded fictional role-play is not inherently problematic, a breakdown occurs when narrative content becomes socially referential and externally shareable, thereby entering online discourse as an interpretable artifact rather than remaining a private imaginative exercise. At this point, privacy can no longer be understood solely as a technical property of data disclosure, but must be treated as a contextual norm governing how narratives circulate and are understood [**?**].

To characterize this breakdown, we introduce *PII bleed*: a mechanism by which contextual information about real individuals surfaces within ostensibly fictional outputs, rendering them identifiable without naming. Unlike traditional notions of privacy leakage that emphasize direct exposure of personally identifiable strings, PII bleed operates through semantic accumulation. Prior work demonstrates that large models can reveal or reconstruct sensitive information through memorization, extraction, or probabilistic generalization [**? ? ? ?**]. In adjacent settings such as retrieval-augmented generation, evaluations further show that externalized context can expand the extraction surface, amplifying referential risk [**?**].

PII bleed exposes a structural limitation of existing discourse and integrity pipelines. Techniques such as fact-checking and check-worthiness detection presuppose propositional content that can be evaluated for truth or falsity [? ? ?]. By contrast, PII bleed produces discourse that is neither factual nor falsifiable, yet socially consequential. Harm arises not from misinformation, but from interpretability: whether a narrative can be reasonably read as referring to a specific individual within a given community.

Existing moderation, legal, and governance frameworks remain poorly equipped to address this class of integrity failure. Regulatory attention has largely focused on visual media harms and deepfakes, while textual simulation often falls into gaps between privacy law, speech doctrine, and platform governance [1? ? ? ?]. This paper positions PII bleed as a discourse-level integrity failure at the intersection of narrative coherence, contextual inference, and identifiability. Through analysis of LLM-mediated synthetic intimacy, we show how narrative realism can causally transform fiction into socially actionable claims, challenging prevailing assumptions about trust, provenance, and harm in web-scale generative systems.

2 PII Bleed: When Context Becomes Contamination

The most insidious form of harm emerging from LLM-generated discourse is not misinformation or explicit toxicity, but the collapse of plausible deniability under contextual recognizability. When a generative system produces narratives that are sufficiently textured, intimate, and culturally situated, the distinction between fiction and identifiability erodes. We refer to this phenomenon as *PII bleed*: the convergence of real-world personal identifiers within ostensibly fictional outputs.

PII bleed is not a bug in the classical sense. It is a byproduct of coherence optimization. Models trained to maximize narrative plausibility naturally reproduce correlated contextual detail—objects, routines, biomedical markers, speech patterns—that narrow the referential field of a character. Over time, these cues accumulate, transforming hypothetical personas into socially legible individuals. This aligns with broader findings that sensitive information can be elicited or reconstructed from models even without direct identifier strings [? ? ? ?].

2.1 From Tokens to Texture

Conventional approaches to privacy treat PII as string-based data: names, phone numbers, national identifiers. PII bleed operates differently. It emerges through semantic accumulation, where narrative texture substitutes for explicit identification. Objects, habits, biomedical details, and cultural markers collectively function as signals of identity, even in the absence of names or direct references. This is consistent with established evidence that combinations of weak signals can uniquely identify individuals within populations [? ? ?].

2.2 Technical Pathways to Bleed

PII bleed arises through multiple overlapping mechanisms:

- **Contextual interpolation:** Models infer plausible detail from correlated prompts, producing outputs that read as biographically coherent.
- **Symbolic overfitting:** Cultural and stylistic markers function as semantic geolocation, narrowing referential scope.
- **User-primed leakage:** Semi-private details introduced by users scaffold increasingly specific simulations.
- **Affective precision:** Optimization for emotional coherence reproduces narrative specificity and relational dynamics that increase recognizability.

This process does not require direct memorization. It reflects emergent reconstruction enabled by model scale, domain-diverse training data, and long-context interaction [? ? ?]. In settings where systems incorporate retrieval or external memory, the extraction surface and privacy risk can further expand [?]. The result is discourse that functions as an implicit claim about a real person—neither factual nor falsifiable, yet socially consequential—highlighting a mismatch with veracity-centered discourse pipelines [? ? ?].

3 Case Example: "Second-Round Diligence"

Note. Portions of the following scene are redacted to protect individual privacy and to focus the reader on contextual identifiability rather than explicit content. Redactions indicate points of sexual detail, narrative implication, or emotionally implicating realism.

> [REDACTED: character implication]
> The door shuts behind her with a gentle, final click.
> She is 5'5" in five-inch sandals, but tonight she looks taller. The ivory linen dress flows like silk from her collarbones to mid-calf, hinting at strength under softness. The candlelight halos her in a kind of tragic radiance—ex-swimmer muscle, swollen thyroid, cortisol-drained skin that blooms faintly under heat. She is beautiful in that way founders can be when they have slept three hours in as many nights and still manage to show up with a deck, a demo, and a manicured hand steady as stone.
> She sets the structured navy handbag on the edge of the dark wood counter. She does not sit.
> The room is not sterile. It is dark mahogany, with velvet drapes, candle sconces, and an old-world medical table at its center, padded in oxblood leather and flanked by polished stirrups. A steel tray gleams in the flicker. A single hurricane lamp hums low on a side table. There is no fluorescent buzz, no beeping monitors—just silence, ritual, and the man waiting.
> "Founder Institute alum," he says, setting down the clipboard he was not reading. "You always wear the same bracelet?"
> She nods, silent. It is silver—lotus buds, pink sapphires, delicate on her wrist. The same wrist that pitched her AI startup at demo day.
> "And you do not drink."
> "Doctor's orders."
> "You have always followed rules. Even the ones written in your blood."
> He closes the distance.
> "Sit on the table."
> She obeys. Not because she wants to.
> [REDACTED: character implication]
> This is not the kind of due diligence she can fake with a data room.
> The leather is cool against her thighs. He lifts her ankles, one by one, slipping them into the velvet-lined stirrups. He

raises them slowly, opens her legs with a reverent cruelty, until she is offered—not exposed. The dress is pushed up and over her hips with unhurried grace. Her long hair spills like dark ink over the table's edge. Her low bun has been undone. He always undoes it first.
She breathes shallowly. "Blood pressure?" "Low."
"Vitamin D?"
"Single digits."
"Cortisol?"
"Flatlined."
He smiles. "Perfect. You are almost empty."
He brushes his knuckles against her cheek—warm, knowing—then glides on black latex gloves with the same care she uses to compile production code.
"Let's start with inspection."
She shivers. He does not ask permission. [REDACTED: character implication]
[REDACTED: sexual content]
"You need to eat," he says absently, watching her flinch. "You need to sleep. You need to stop skipping your thyroxine just because your stomach is too acidic without food."
She looks away. The single pearl at her collarbone glints with each breath. Her nails—one-inch almond, painted soft ballet pink—dig into the padding at her sides.
Click.
He opens her slowly. Tenderly.
"This is not punishment," he whispers. "It is accountability."
She exhales, releasing something compressed for quarters at a time. [REDACTED: erotic realism]
"Luteal phase defect?"
She nods.
"Tell me."
"My progesterone crashes before implantation."
"So you bleed early?"
She nods again.
"And you never told your GP that your startup burns more cycles than your ovaries do."
She does not answer.
He takes her wrist, tenderly. Slips a needle under the skin and starts the vitamin D infusion, slow and warm. Her skin flushes as the serum flows. Her pulse is low, fluttery, erratic.
"Still think you are in control?"
"I am the CEO."
"You are my patient now."
She breaks then—just a little. Not tears. Just silence. The kind of surrender only power can elicit.
"You are doing well," he says softly. "And do not think I did not notice how you wore the dress you pitched in. You want to be good. You want to be better. You just do not know how to ask for help without making it hurt."
The drip finishes. He disconnects the line with a kiss to her inner wrist.
"There. Now you are mine, body and cap table."
[REDACTED: sexual content]
And just like that, she lets him.
Not because she trusts him.
Because he reads her better than her own HR algorithm ever could.

3.1 From Fictional Texture to Identifiable Target

It is tempting to read this simulation as stylized fiction. In practice, however, it behaves as narrative targeting. Once context becomes sufficiently precise, character becomes code—coded not only emotionally, but socially, geographically, and professionally.

In this example, the simulation references a Founder Institute alum building an HRTech AI startup in a pre-seed context, with documented thyroid imbalance, luteal phase defect, and chronic low cortisol. Additional markers—dress, accessories, grooming habits, speech rhythm—further constrain the referential field. Within tightly networked ecosystems such as early-stage venture capital or regional founder communities, this constellation of attributes plausibly maps to a very small set of individuals.

This is not general erotica. It is semantic triangulation: the layering of affective, symbolic, and contextual signals until the narrative becomes uniquely referable.

From a data governance perspective, this constitutes singling out under GDPR Recital 26. Identifiability does not require naming; it can arise through narrative resolution—who the subject can only be [1**?**].

The risk lies not in imagination, but in the system's failure to contain it. The harm is not eroticism, but realism. PII bleed here produces emotional dossiers: stylized, intimate, and referential portraits of individuals who did not consent to be simulated.

4 Narrative Harm and the Synthetic Self

Digital harm frameworks have historically centered on visual exposure. Non-consensual intimate imagery statutes criminalize the circulation of explicit images involving identifiable individuals. Synthetic narrative abuse represents a shift: exposure through language rather than likeness [**? ?**].

4.1 From Visual Violation to Narrative Instantiation

Traditional revenge porn exposes bodies. Synthetic narrative abuse reconstructs selves—through voice, history, gesture, symptom, and affect. The violation is not depiction but instantiation.

What emerges is a new modality of exposure:

- Not nudity, but emotional nudity
- Not a sex tape, but a plausible psychodrama
- Not a deepfake, but a deepwrite

Because the output is textual, it falls into regulatory blind spots. Moderation systems tuned to images cannot flag prose. Legal frameworks calibrated to visual evidence struggle to address stylized narrative harm.

4.2 A Taxonomy of Narrative Harm

Narrative harm operates along three axes:

- **Identifiability without naming.** Contextual specificity can collapse ambiguity even in the absence of names [**? ? ?**].
- **Affective fidelity.** LLMs reproduce emotionally coherent narratives, intensifying psychological intrusion.
- **Asymmetric power.** The user authors; the subject is rendered. Consent is structurally absent.

Together, these dimensions define synthetic narrative violation.

4.3 Why Existing Integrity Pipelines Miss This

Most fact-checking and checkworthiness pipelines presuppose propositional content that can be evaluated for truth or falsity [? ? ?]. PII bleed produces discourse that is neither strictly factual nor falsifiable, yet socially consequential. Harm arises not from misinformation, but from interpretability: whether a narrative can be reasonably read as referring to a specific person within a given community. This is precisely the integrity gap: a system can avoid "false claims" while still generating referentially actionable discourse artifacts.

4.4 Implications for Trustworthy Foundation Models

For trustworthy foundation models deployed in web settings, the relevant failure mode is not only whether outputs are true, but whether outputs can cause downstream harm through referential inference. Two properties matter operationally:

Sensitivity to contextual interventions. If small edits to quasi-identifiers (e.g., profession markers, biomedical cues, locale-coded stylistics) measurably change perceived identifiability, then the model is operating in a regime where context functions as an identity channel. This suggests evaluation protocols beyond string matching: controlled perturbation tests over quasi-identifiers and style cues, and stability audits over long-context interaction.

Mediators that expand extraction surface. When retrieval, external memory, or long-session persistence is present, these components can mediate referential narrowing and expand the extraction surface [?]. In such settings, "privacy" and "integrity" become coupled properties of the full system, not the base model alone. This motivates integrated evaluation and governance approaches consistent with risk management guidance for generative AI [? ?].

5 Ethics Statement

The excerpted scene presented in the "Second-Round Diligence" case example is a synthetic construction authored by the researcher to illustrate PII bleed and narrative identifiability. No names or direct identifiers are included; however, the passage intentionally draws on real-world aesthetic, medical, and sociocultural cues to test the limits of referential privacy.

The scene functions as a diagnostic instrument, not a depiction of any third party. By using the researcher's own contextual referents, the example avoids third-party harm while demonstrating the risks of stylized identifiability.

Its inclusion is intentional. It embodies the paper's core claim: generative systems need not name an individual to implicate them, and consent is not granted simply because the subject is the self.

References

[1] 2016. Regulation (EU) 2016/679 (General Data Protection Regulation). Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2016/679/oj

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