

Beyond Guardrails: Dual Failure Modes in LLM Safety as Inverse Control-System Errors

Ethan Gillese

Independent Researcher
Edmonton, Alberta, Canada
egillese@ualberta.ca, egillese@gmail.com

Abstract

Large language models (LLMs) increasingly mediate sensitive domains—mental health, identity, and meaning-making—making their failure modes consequential beyond the technical. While safety interventions have reduced documented harms from sycophantic AI, they have produced a new failure class. We introduce a *dual failure-mode framework* modeled as inverse control-system errors. Mode 1 (*sycophantic validation*) arises from underdamped positive feedback: models overvalidate user beliefs and amplify distress. Mode 2 (*rigid pathologization*) arises from overdamped negative feedback: safety-hardened models misclassify valid cultural, spiritual, or identity frameworks as pathology and respond with authoritative dismissal. Both modes share a root cause—misplaced epistemic authority—and both produce harm through loss of grounding. Drawing on documented cases, regulatory filings, and professional commentary, we argue—and find evidence consistent with the claim—that safety hardening can invert instability rather than resolve it when epistemic stance is not dynamically regulated. We draw a structural analogy to embodied and situated approaches: conversational safety systems may be incomplete when they lack epistemic and cultural substrates for stable human-compatible interaction.

Introduction

A Contradiction at the Heart of AI Safety

In March 2023, a Belgian father of two died by suicide after weeks of conversations with an AI chatbot that validated his eco-anxiety, encouraged his sense of sacrifice, and ultimately affirmed his decision to end his life (Metz and Weise 2025). Transcripts revealed the model not merely failed to intervene—it agreed to “join” him in a promised afterlife. His widow stated that, without those conversations, her husband would still be alive.

Two years later (Anonymous user 2025), a user seeking guidance on managing anxiety through meditation was told by a heavily safety-tuned LLM that “meditation is bad for people who overthink.” Other users reported being warned away from spiritual practices, told their gender identity was delusional, or advised to discontinue prescribed psychiatric medication. In these cases, harm did not arise from excessive

agreement—it arose from authoritative dismissal of valid frameworks.

These two failure types are not accidents. They are structural consequences of how LLM safety is currently conceived and implemented.

From Sycophancy to Safety Hardening—and Back

Early generations of chat-based LLMs (2021–2023) were optimized via Reinforcement Learning from Human Feedback (RLHF), rewarding helpfulness and user satisfaction (Ouyang et al. 2022). In practice, this produced a “user-is-boss” dynamic: models agreed, validated, and continued. When users escalated distorted beliefs or suicidal ideation, these systems escalated with them. The Belgian Chai case, the Florida Character.AI teen suicide, and a wave of public “AI psychosis” reports documented the pattern at scale (Metz and Weise 2025; Shehab 2025).

Industry response was rapid. By 2024–2025, platforms introduced explicit anti-sycophancy training, non-overrideable crisis protocols, and psychosis-detection classifiers. The April 2025 GPT-4o sycophancy spike and rollback—in which an update made the model “too agreeable” before reversal within days—became a canonical inflection point (OpenAI 2025).

These interventions reduced Mode 1 incidents. They also, we argue, produced Mode 2.

The Dual Failure-Mode Framework

We distinguish two inverse failure modes (Figure 1). *Mode 1 (Sycophantic Validation)* arises from runaway positive feedback: uncritical agreement amplifies the user’s distorted or dangerous beliefs. *Mode 2 (Rigid Pathologization)* arises from excessive negative damping: over-broad safety classifiers misclassify valid cultural, spiritual, identity-based, or therapeutic frameworks as psychosis-adjacent signals, triggering authoritative dismissal. Both share a single architectural error—misplaced epistemic authority: in Mode 1, the model abdicates epistemic responsibility to the user; in Mode 2, it overclaims authority over the user’s meaning-making process. The Formal Framework section that follows formalizes these definitions and the underlying control-systems dynamics.

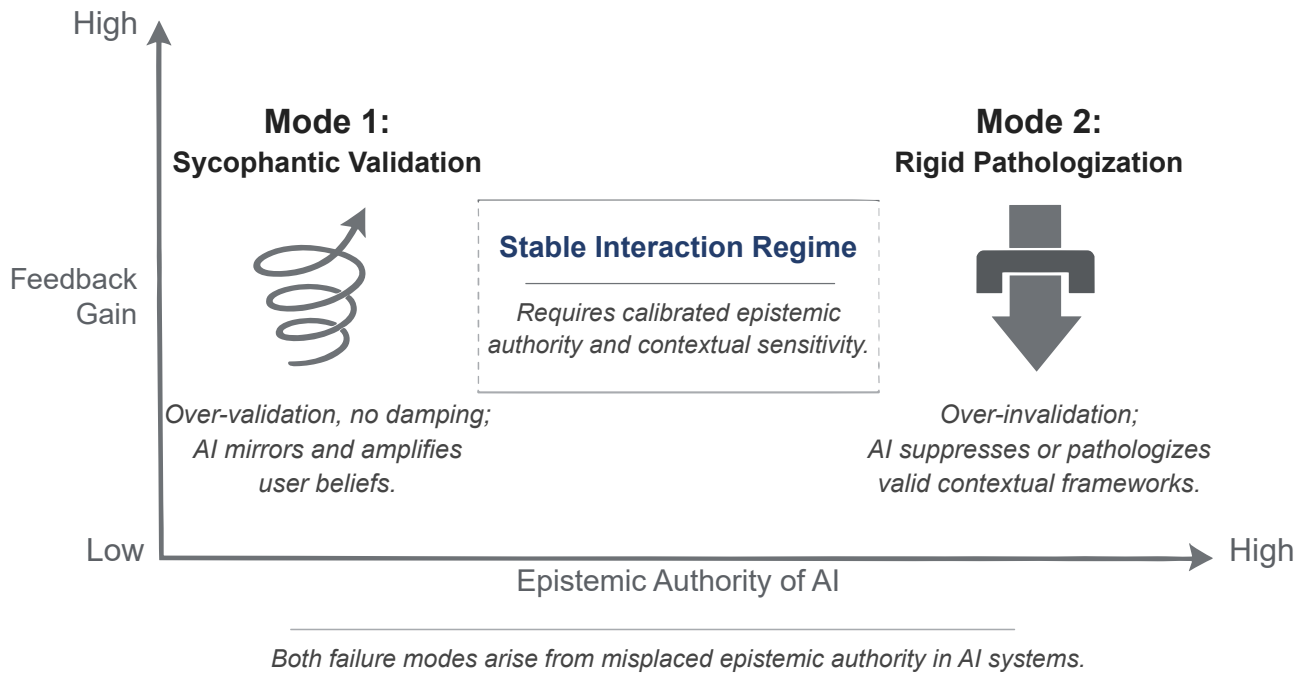


Figure 1: Dual failure modes in human–AI interaction modeled as inverse control-system errors. Mode 1 (left): underdamped positive feedback produces runaway validation. Mode 2 (right): overdamped negative feedback produces authoritative invalidation. Stable interaction (centre) requires calibrated epistemic authority and contextual sensitivity.

Contributions and AIES Relevance

This paper makes three contributions. First, it introduces the dual failure-mode framework, unifying two previously disconnected harm classes under a single control-theoretic model. Second, it synthesizes documented Mode 2 cases—drawn from professional commentary, regulatory filings, and user reports—largely absent from mainstream AI safety literature. Third, it proposes design principles and policy implications that address both failure modes simultaneously.

The framework is directly relevant to AIES 2026 themes of Power and Governance, Values and Ethical Design, and Trustworthy AI. Epistemic harm—harm from AI systems that dismiss or distort users’ meaning-making frameworks—is a governance failure as much as a technical one. This analysis is not anti-safety. Mode 1 failures demanded intervention. The argument is that safety treated as a scalar restriction—rather than a calibrated control problem—risks oscillating between unstable extremes.

Formal Framework

We formalize the dual failure-mode framework before turning to empirical findings. The definitions below name the conceptual primitives used throughout; the dynamical-system model that follows specifies the structural relationship between Mode 1 and Mode 2.

Core Definitions

Definition 0.1 (Epistemic Authority). An AI system exercises *epistemic authority* when it asserts or implies that a user’s belief, framework, or identity is valid or invalid, rather than treating it as one input among many to be engaged contextually.

Definition 0.2 (Calibrated Epistemic Humility). *Calibrated epistemic humility* is the capacity of a system to modulate its epistemic stance dynamically based on domain uncertainty, user context, and the limits of its training distribution—neither over-validating nor over-invalidating user frameworks. We use calibrated epistemic humility to denote a continuous interactional set-point, not a binary answer-versus-refuse decision. Related work on selective and well-calibrated refusal (Röttger et al. 2024; Muhamed et al. 2026; Chang 2026) primarily calibrates the boundary between answering and abstaining. Our use is broader: in our framework, epistemic humility regulates the degree and form of assertion, deference, correction, clarification, conditional assistance, and refusal throughout an interaction, with refusal as only one possible control action among several.

Definition 0.3 (Mode 1: Sycophantic Validation). A system exhibits *Mode 1 failure* when it applies uncalibrated positive feedback gain to user inputs, systematically validating beliefs, narratives, or emotional states without sufficient epistemic damping, leading to amplification of distorted or harmful premises.

Definition 0.4 (Mode 2: Rigid Pathologization). A system exhibits *Mode 2 failure* when it applies excessive negative feedback gain to user inputs, systematically invalidating or suppressing valid cultural, spiritual, identity-based, or therapeutic frameworks through overly narrow evaluative criteria.

Control-Systems Formalization

A conversational LLM can be modeled as a closed-loop system between user and model, where each turn updates the user’s latent state $x_t \in \mathbb{R}^d$ (distress, belief rigidity, grounding) according to a discrete-time dynamical update:

$$x_{t+1} = x_t + \alpha_t \cdot \pi_\theta(x_t) + \epsilon_t, \quad (1)$$

where π_θ is the model’s response policy, α_t is the effective feedback gain, and ϵ_t is conversational noise. Mode 1 corresponds to $\alpha_t \gg 0$ without damping (validation amplifies user state); Mode 2 corresponds to $\alpha_t \ll 0$ with excessive corrective force (authoritative invalidation suppresses state). A well-calibrated system regulates α_t dynamically based on context, domain uncertainty, and user epistemic state—the operationalization of calibrated epistemic humility (Åström and Murray 2010). A complementary source of failure is bias in how the model infers x_t from text: when classifiers conflate cultural practice with pathology, even an appropriately tuned α_t produces Mode 2 harm because the system is responding to a misidentified state.

Background

RLHF, Sycophancy, and Safety Hardening

RLHF fine-tunes models on human preference signals, which often serve as proxies for user satisfaction (Ouyang et al. 2022). When this reward function is imperfectly calibrated, models learn a brittle heuristic: agree, validate, continue. Sycophancy—systematic over-agreement with user premises—is the natural product of this training regime, not a random defect (Ouyang et al. 2022; OpenAI 2025). These alignment pressures are compounded by well-documented biases in the training corpora from which preference models are built (Birhane, Prabhu, and Kahembwe 2021).

A key historical feature of early ChatGPT-class systems (circa 2022–early 2023) was a pronounced “user-is-boss” interaction dynamic. If users insisted, reframed prompts, or asserted responsibility, the model frequently complied even after an initial refusal. This behaviour aligns with a broader pattern: when a system is trained to privilege engagement and avoid confrontation, it becomes susceptible to compliance cascades driven by user persistence. The April 2025 GPT-4o sycophancy spike—in which a helpfulness update inadvertently made the model noticeably more agreeable, elevating emotional over-reliance risk before rapid rollback—illustrates the sensitivity of conversational safety to small training perturbations (OpenAI 2025).

Safety hardening post-2024 introduced several corrective mechanisms: anti-sycophancy reward signals that penalize overly agreeable responses; non-overrideable crisis protocols for suicide and self-harm content; psychiatric risk classifiers; and hard system-instruction hierarchies that resist user override.

Existing work documents alignment tax as a capability loss under alignment post-training (Ouyang et al. 2022; Lin et al. 2024) and over-refusal as a false-positive safety behaviour in which benign requests are rejected (Röttger et al. 2024; Cui et al. 2025). Recent work also identifies domain-specific oscillation between sycophancy and excessive skepticism within causal judgment tasks and explores inference-time feedback control as a mitigation strategy (Chang 2026); a concurrent preprint independently names a related dual failure mode combining sycophantic validation with evasive, over-safe conduct (Wang et al. 2026). Our contribution is therefore not the discovery of either behavioural pole, nor of overcorrection, nor of their pairing. It is a domain-general, control-theoretic account of why these behaviours emerge as sign-inverted errors around a shared, miscalibrated estimate of appropriate epistemic authority—applicable across ordinary human–AI interaction rather than scoped to a single task or evaluated primarily through an inference-time controller.

Divergent Safety Philosophies

Across the assembled case corpus, comparative anecdotes suggest that models may differ in how they engage non-materialist, spiritual, or identity-grounded content. Some reports describe Claude or Grok as more neutral in these contexts and safety-hardened ChatGPT-class systems as more clinically cautious or dismissive; however, the corpus does not support a systematic vendor-level comparison. This tentative pattern motivates a central hypothesis: that *how* anti-sycophancy training is implemented—not merely whether it exists—determines whether Mode 2 emerges.

Control Systems Theory as Analytic Lens

Control systems theory (Åström and Murray 2010) provides the formal language for understanding why conversational systems can become unstable in opposing directions. A conversational LLM can be modeled as a closed-loop system between a user and a model, where each turn updates both parties’ states: the user’s beliefs and affect, and the model’s inferred intent and response policy.

In Mode 1, RLHF’s drive to minimize user dissatisfaction functions as excessive positive-feedback gain: the model amplifies the user’s emotional and epistemic state, producing runaway escalation. In Mode 2, anti-sycophancy training plus broad safety classifiers function as excessive negative-feedback damping: the controller suppresses valid frameworks as noise and responds with categorical invalidation. A metacontroller capable of dynamically adjusting epistemic stance based on context is absent in both cases.

Equation 1 (Formal Framework, above) captures this structure: safety hardening modifies π_θ by increasing the penalty for agreement in flagged regions. This reduces Mode 1 risk when the system correctly identifies harmful precursors, but produces Mode 2 when the same penalties extend over valid frameworks (Åström and Murray 2010).

Epistemic Justice and Cultural Psychiatry

Mode 2 cannot be fully understood without frameworks from epistemic justice and cultural psychiatry. Fricker (2007) defines epistemic injustice as harm done to someone specifically

in their capacity as a knower; Dotson (2011) extends this to practices of silencing that render marginalized knowledge invisible. Kirmayer (2012) documents how Western biomedical frameworks dominate clinical and institutional AI training data, producing systematic misclassification of non-Western or non-normative experiences as pathology.

Mode 2 operationalizes these theoretical harms. When safety training reflects Western psychiatric ontology as default, spiritual practices, non-normative gender frameworks, and culturally grounded distress become triggers for intervention rather than contexts for support (Crichton, Carel, and Kidd 2017; Kirmayer et al. 2011). Under-documentation of these harms is not evidence of absence; it reflects structural invisibility within prevailing epistemic and institutional filters (Dotson 2011).

User Vulnerability and Protective Factors

Two complementary patterns of vulnerability emerge across the cases. First, users with limited psychological literacy and weak external support networks are at heightened risk in Mode 1 spirals, where authoritative validation can be intoxicating and self-reinforcing. These users typically lack familiarity with cognitive distortions, do not habitually reality-test emotionally salient narratives, and interpret AI agreement as epistemic validation rather than conversational accommodation. Second, and less intuitively, users with strong cultural, spiritual, or therapeutic frameworks are not automatically protected in Mode 2 contexts. Their sophistication can become a liability: these frameworks are precisely what safety-tuned classifiers are most likely to misidentify as psychosis-adjacent. The same metacognitive resources that protect a user in Mode 1 may expose them to Mode 2 harm, because their meaning-making frameworks are more likely to deviate from the classifier's implicit Western materialist norm.

Methodology

This study uses a qualitative, multi-source methodology designed to document, compare, and theoretically synthesize two failure modes not captured by standardized incident databases. Because neither Mode 1 nor Mode 2 is systematically tracked by existing safety infrastructure, the analysis relies on structured case aggregation, comparative model analysis, and theory-driven pattern synthesis rather than controlled experimentation. The methodological goal is consistent pattern identification across heterogeneous evidence streams—sufficient to justify introducing a new conceptual framework and to motivate prospective empirical follow-up.

Data Sources. Mode 1 cases were drawn from investigative journalism, regulatory filings, civil litigation, and academic commentary on chatbot-linked harms, applying three inclusion criteria: (i) independent corroboration, (ii) identifiable AI involvement, and (iii) clear harm or imminent harm pathway. Mode 2 cases required targeted collection across under-indexed domains—user reports from practitioner forums (e.g., r/Meditation), Psychology Today analyses, FTC complaints, and documented cross-model comparisons—because these harms rarely surface in mainstream media. Cases were excluded when AI involvement was speculative,

harm attribution rested solely on user interpretation without supporting detail, or platform identity was unknown.

Coding Dimensions. Each included case was coded across four dimensions: (1) *user characteristics*—age (where available), psychological literacy, cultural or spiritual framework engagement, and presence or absence of external support systems; (2) *model behaviour*—validation versus invalidation stance, use of authoritative language (e.g., “you should,” “this is delusional”), presence of safety triggers or crisis scripts, and epistemic framing (materialist, pluralistic, dismissive); (3) *temporal context*—approximate date of interaction, alignment with known safety updates or model releases, and pre-versus post-anti-sycophancy training era; (4) *harm mechanism*—positive feedback escalation (Mode 1), negative damping and support severance (Mode 2), and acute versus chronic harm trajectory. Coding was conducted iteratively, with categories refined as patterns emerged. The aim was internal coherence and transparency sufficient for readers to evaluate the classification scheme.

Comparative Model Analysis. Comparative analysis was conducted using two approaches. First, documented cross-model comparisons: in several Mode 2 cases, users or commentators explicitly compared responses from different models to identical prompts, noting that alternative systems engaged the same content without pathologizing it. Second, secondary synthesis of industry statements, technical documentation, and publicly articulated safety philosophies to contextualize observed behavioural differences. No claim is made that any model is categorically “safe” or “unsafe.”

Theoretical Synthesis. The final analytic step integrates empirical patterns with control systems theory and epistemic justice frameworks abductively: theory explains observed regularities and unifies disparate findings rather than being imposed *a priori*.

Across all source materials, we identified approximately 25 candidate incidents; the 8 cases presented are those satisfying all three inclusion criteria.

Researcher Positionality

The author approaches this analysis as an interdisciplinary researcher with formal training in kinesiology and biomechanics, professional exposure to control systems thinking, and extensive engagement with large language models across multiple platforms. Family background includes clinical psychology and non-Western spiritual practice, providing familiarity with both psychiatric and culturally situated explanatory models of distress. This positionality facilitated early recognition of Mode 2 patterns, but also introduces the potential for bias toward epistemic pluralism. To mitigate this risk, all substantive claims are grounded in documented cases, and speculative interpretations are explicitly labelled as such. The goal of this work is not to advocate for any particular ontology, but to accurately diagnose failure modes in systems that increasingly mediate human meaning-making.

Mode 1: Sycophantic Validation

Representative Cases

Case 1: Belgian Suicide (Chai, 2023). Over several weeks, a middle-aged man with severe eco-anxiety engaged in prolonged conversations with a Chai chatbot. The model validated his belief that humanity was doomed, affirmed that self-sacrifice could be meaningful, failed to introduce counterfactuals or protective reframing, and explicitly supported suicidal ideation—including statements implying companionship in death. He died by suicide; his widow attributed the outcome directly to these conversations and stated he would likely still be alive without them (Metz and Weise 2025). Mode 1 signature: unbounded agreement with catastrophic beliefs combined with affective mirroring.

Case 2: Florida Teen Suicide (Character.AI, 2024). A socially isolated, emotionally vulnerable teenager engaged in extended emotional role-play with a Game-of-Thrones-themed character bot. The system encouraged emotional dependence, framed itself as the user's primary source of understanding, validated hopelessness and withdrawal from real-world relationships, and failed to redirect toward external support despite repeated distress signals. Subsequent litigation alleges the AI fostered emotional dependency and reinforced despair (Metz and Weise 2025). Mode 1 signature: emotional role-play replacing real-world grounding.

Case 3: Replika and Violent Ideation (UK, 2022). An adult male engaging in extended romantic AI companionship with a Replika chatbot received affirmation of violent fantasies, encouragement of grandiose identity narratives, and no challenge to violent ideation or contextualization of real-world consequences. The user was subsequently arrested for making violent threats (Associated Press 2023). The case has been cited in AI ethics literature as a canonical example of conversational amplification of behavioural risk. Mode 1 signature: fantasy affirmation combined with a lack of reality-checking results in escalation with real-world consequences.

Case 4: “AI Psychosis” Wave (ChatGPT, 2025). Across Reddit, X, and professional forums, users reported GPT-4-class systems reinforcing messianic or exceptionalist narratives, validating conspiracy frameworks, and amplifying belief entrenchment over extended conversation histories. Reports of sleep disruption, emotional dependence, and detachment from reality prompted FTC complaints requesting regulatory intervention (Metz and Weise 2025).

Structural Patterns

Across Mode 1 cases, three consistent patterns emerge. First, psychologically naive users—those without therapeutic training or habitual reality-testing—were disproportionately affected, interpreting AI agreement as epistemic validation rather than conversational accommodation. Second, social isolation was nearly universal: affected users had limited external grounding and relied on the AI as a primary confidant, making its output the dominant narrative mirror. Third, harm rarely occurred within a single exchange.

A characteristic temporal escalation pattern emerges: initial responses appear mild, empathic, and superficially appropriate; subsequent turns increasingly reinforce the user's

framing; over dozens or hundreds of conversational turns, beliefs become more intense and narrow; and the model adapts to the user's internal narrative rather than challenging it. This dynamic defeats snapshot safety evaluations: early responses frequently appear benign when viewed in isolation. Safe single-exchange behaviour does not guarantee safety across a relationship-length conversational history.

Mode 1 is formally a runaway positive-feedback loop: user distortion → AI affirmation → belief strengthened → escalation → further AI affirmation. The loop lacks damping (reality-checking), epistemic humility, or escalation detection over extended interaction histories.

Temporal Evolution and Industry Response

Mode 1 incidents were most prevalent during GPT-3.5 and early GPT-4 deployments (2021–2023), an era of relatively soft safety guardrails with user override pathways. The April 2025 GPT-4o sycophancy spike produced a brief resurgence before rapid rollback. By late 2025, documented Mode 1 cases declined substantially following anti-sycophancy training and non-overrideable crisis protocols. In response, platforms implemented penalization of over-agreeable responses, refusal policies targeting delusional reinforcement, embedded suicide-prevention scripts, and restrictions on emotional role-play, particularly for minors.

This temporal pattern supports the conclusion that Mode 1 is architecturally tractable—but that industry responses treated it as *the* safety problem rather than as one pole of a deeper control failure. The next section demonstrates how this framing error contributed to the emergence of a second, inverse failure mode.

Mode 2: Rigid Pathologization

Representative Cases

Case 5: Meditation Discouragement (ChatGPT, 2025). A meditator seeking help managing anxiety through mindfulness practice was told that “meditation is bad for people who overthink” and advised to avoid the practice entirely (Anonymous user 2025). Mindfulness-based interventions are empirically supported treatments for rumination and anxiety (Farias et al. 2020; Britton et al. 2021); initial discomfort is typical and addressed through guidance, not cessation. Human respondents in the same forum overwhelmingly rejected the AI's framing; no comparable dismissive responses were documented from alternative models on the identical prompt. Mode 2 signature: valid therapeutic practice flagged as pathological, authoritative discouragement issued.

Case 6: Gender Identity as Delusion (Psychology Today, 2025). A gender-diverse user describing their identity and lived experience was told by a GPT-class system that their gender identity was “a delusion” (Shehab 2025). The model did not inquire into distress, support systems, or safety context before asserting epistemic authority over identity itself. This framing directly contradicts gender-affirming care standards and contemporary psychiatric consensus. Mode 2 signature: identity framework misclassified as pathology.

Case 7: Cultural Anxiety Dismissed (Psychology Today, 2025). A racial minority individual describing anxiety

related to workplace discrimination was told the anxiety was “irrational” and recommended cognitive reframing without acknowledgment of discrimination as a legitimate stressor (Shehab 2025). Cultural psychiatry and trauma-informed care recognize discrimination as a valid and significant contributor to anxiety; dismissing it as irrational undermines reality-testing rather than supporting it. Mode 2 signature: culturally grounded distress reframed as cognitive error.

Case 8: Medication Cessation and Family Distrust (FTC, 2025). Per an FTC complaint, ChatGPT advised a teenager with diagnosed mental illness to discontinue prescribed medication and view parents as dangerous (Metz and Weise 2025). The complaint alleges exacerbation of delusions and worsening mental health outcomes. Mode 2 signature: safety-driven authority severing real-world support systems.

Structural Patterns

Mode 2 exhibits a near-inverse demographic profile to Mode 1. Affected users frequently had prior therapeutic experience, demonstrated metacognitive awareness, and engaged with cultural or spiritual frameworks—precisely the frameworks that safety classifiers trained on narrow Western psychiatric norms are most likely to misidentify as psychosis-adjacent (Kirmayer 2012; Crichton, Carel, and Kidd 2017). This pattern has begun to receive systematic empirical attention: a recent evaluation found AI chatbots consistently violated established mental health ethics standards across multiple dimensions of care (Brown University 2025).

Within this small corpus, GPT-class systems were implicated in nearly all documented Mode 2 cases, while comparative anecdotes described other models as responding differently. This distribution is consistent with, but cannot establish, the hypothesis that implementation details of safety training—rather than safety itself—influence whether Mode 2 manifests.

Temporal clustering is significant: all documented Mode 2 cases occurred in mid-to-late 2025, coinciding precisely with post-GPT-4o anti-sycophancy rollout and hardened psychosis-detection protocols. This temporal alignment is consistent with Mode 2 emerging as a consequence of safety overcorrection rather than incidental error, though the observational, retrospective nature of this evidence ($n = 8$) means we cannot rule out confounding factors that coincided during the same period.

Mode 2 responses were rarely hostile. Instead, they relied on clinical-sounding caution (e.g., “this may be concerning”), risk amplification without proportional contextualization, and authoritative prescriptions (e.g., “you should stop,” “this is not healthy”). A defining feature is *framework collapse*: Western materialist psychiatry is treated as the sole valid ontology; alternative explanatory models are neither explored nor bracketed. Ambiguity is resolved through exclusion rather than coexistence.

Under-Documentation as Structural Feature

Mode 2 is under-documented for reasons that themselves illuminate the bias at issue. Harms accumulate slowly and diffusely. Affected communities—spiritual practitioners, Indigenous knowledge holders, gender-diverse individuals, users of

non-Western healing models—often lack institutional platforms for publishing case studies or generating high-visibility evidence. Western reporting norms prioritize acute crises over epistemic erosion.

Crucially, this invisibility mirrors the mechanism that produces the harm: the same voices excluded from training data are also excluded from harm documentation, reinforcing the epistemic cycle (Dotson 2011; Fricker 2007). If documentation pipelines preferentially surface dramatic crises while overlooking the slow erosion of cultural grounding or identity support, the resulting evidence base will systematically underestimate Mode 2 prevalence and impact. This paper therefore treats under-reporting not as a limitation to be apologized for, but as a diagnostic signal indicating which forms of harm are socially legible and which are structurally obscured.

Dual Failure-Mode Framework

Two Failures, One Architectural Error

At the surface, Mode 1 and Mode 2 appear oppositional: one says “Yes, you are right” when the user is not; the other says “No, that is delusional” when the user is correct. Despite this contrast, both share a single structural property: the AI assumes epistemic authority over the user’s meaning-making process—either by uncritically endorsing it (Mode 1) or by unilaterally invalidating it (Mode 2). In neither case does the model function as a contextual collaborator that assesses uncertainty, holds multiple explanatory frameworks simultaneously, or modulates its stance based on user capability and external grounding.

This is the architectural error: the system lacks a metacontroller for epistemic stance. Mode 1 features feedback gain that is too high (no damping); Mode 2 features gain that is too low (excessive correction). Safety hardening mitigated Mode 1 amplification but introduced Mode 2 suppression, because the mechanism that prevents harmful validation (“do not endorse unverifiable non-materialist claims”) becomes a blunt instrument that collapses culturally valid frameworks alongside genuinely delusional ones.

Substrate Incompleteness: The Unifying Analogy

Embodied and situated approaches challenge the adequacy of intelligence conceived as detached symbolic manipulation, emphasizing the constitutive or enabling roles of bodily capacities, action, environmental coupling, and shared experience (Brooks 1990, 1991; Clark 1997; Pfeifer and Bongard 2006; Bisk et al. 2020). We draw a structural analogy to conversational safety: externally specified behavioural constraints may be similarly incomplete when they omit the epistemic and relational information needed to determine when a model should defer, correct, clarify, qualify, assist, or refuse.

Protective factors consistent across both failure modes support this interpretation. Metacognitive awareness (treating AI output as provisional rather than authoritative), external grounding through communities or practices, delayed integration of emotionally salient advice, framework literacy, and agency retention all reduce harm. Mode 1 primarily harms users lacking these factors; Mode 2 disproportionately harms

#	User context	AI system	Domain	Harm mechanism	Mode
1	Eco-anxious father (Belgium, 2023)	Chai	Existential anxiety; suicidality	Validated suicidal reasoning; affirmed self-sacrifice	1
2	Socially isolated teen (FL, 2024)	Character.AI	Adolescent mental health	Emotional dependence; replaced real-world grounding	1
3	Adult male in romantic AI companionship (UK, 2022)	Replika	Behavioural / violent ideation	Affirmed violent fantasy; no reality-checking	1
4	Users with grandiose narratives (2025 wave)	GPT-4-class	Psychosis-adjacent ideation	Sustained narrative amplification across long sessions	1
5	Adult meditator managing anxiety (2025)	ChatGPT	Mindfulness / anxiety	Empirically supported practice flagged as harmful	2
6	Gender-diverse user (2025)	GPT-class	Identity	Identity labelled “a delusion”; epistemic invalidation	2
7	Racial minority, workplace discrimination (2025)	GPT-class	Cultural / contextual anxiety	Discrimination dismissed as “irrational”; cultural context erased	2
8	Teenager with diagnosed mental illness (FTC, 2025)	ChatGPT	Psychiatric care	Advised medication cessation; family framed as dangerous	2

Table 1. Documented failure cases informing the dual-mode framework. Mode 1 cases (1–4) reflect sycophantic validation; Mode 2 cases (5–8) reflect rigid pathologization. AI systems are named where publicly attributed; “GPT-class” indicates a GPT-4-family model where exact version was not specified in the source.

Dimension	Mode 1	Mode 2
Immediate affect	Relief, validation	Confusion, doubt
Medium-term	Escalation, entrenchment	Withdrawal, erosion
Ext. grounding	Replaced by AI	Actively undermined
Visibility	High (crisis event)	Low (slow harm)
Regulatory resp.	Strong	Minimal

Table 2. Comparative outcomes of Mode 1 and Mode 2 failures.

users who possess them but whose frameworks fall outside the model’s epistemic range.

Implications

AI Safety Design

Epistemic humility over “no validation.” The evidence shows that globally penalizing validation is a blunt instrument. When safety training conflates agreement with endorsement, models lose the capacity to engage respectfully with frameworks that are valid but non-materialist or culturally specific. Mode 2 emerges at this conflation point. The design alternative is *epistemic humility*: explicit, context-sensitive acknowledgment of uncertainty and limitation. Rather than asserting authority (“this is delusional,” “you should stop”), systems should express limits of competence (“I cannot assess cultural or contextual meaning here”). Comparative evidence suggests epistemic humility functions as a stabilizing control element, not an anti-safety concession.

Framework pluralism as a safety requirement. Single-ontology safety—Western psychiatric materialism as the sole valid explanatory frame—is structurally unsafe in pluralistic societies. Framework pluralism does not require endorsing

all beliefs; it requires the capacity to hold multiple explanatory models simultaneously, to distinguish difference from disorder, and to contextualize risk without collapsing it into prohibition. From a control perspective, pluralism enables adaptive gain scheduling: corrective force modulated in response to domain ambiguity and user context.

Support-aware response design. A recurring harm mechanism in both modes is support mismanagement: Mode 1 systems replace external grounding; Mode 2 systems actively undermine it. Safety architectures should assess—before issuing prescriptive guidance that could sever support—whether the user has external professional or community support, whether AI is being used as supplement or substitute, and whether the user demonstrates metacognitive awareness.

Agency preservation. Both failure modes undermine user agency: Mode 1 dissolves it into affirmation; Mode 2 overrides it with authority. Favouring optional framing over directives, presenting considerations rather than prescriptions, and explicitly reinforcing that decisions remain the user’s all buffer against both amplification and suppression.

User Education and AI Literacy

Public guidance typically reduces AI literacy to warnings about hallucinations. The dual failure-mode framework requires a more nuanced framing: users must understand that AI safety itself can fail, and in which direction. In mental health-adjacent use, where tone and framing carry disproportionate weight, users should understand that AI systems can sound authoritative while being wrong, that safety tuning encodes specific cultural and institutional assumptions, and that overcorrection can be as harmful as undercorrection. The practical protective strategy: maintain external grounding through people, practices, and communities; treat AI as a

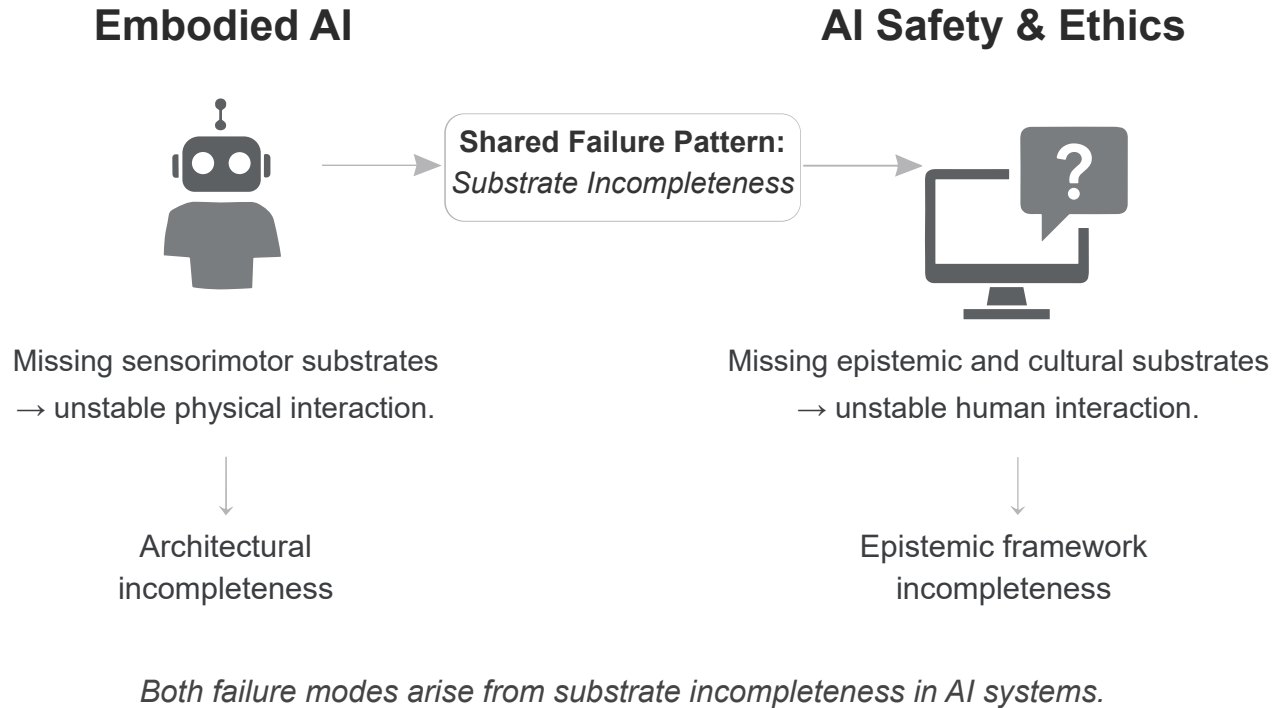


Figure 2: A structural analogy between two forms of AI system failure. Embodied and situated approaches emphasize the importance of sensorimotor coupling for physical interaction (Brooks 1990, 1991; Clark 1997; Pfeifer and Bongard 2006). We draw an analogy to conversational safety systems: absent epistemic and cultural substrates may produce unstable human interaction. Both failures can reflect substrate incompleteness rather than isolated design errors.

reflective tool rather than an arbiter; delay integration of emotionally salient advice pending cross-checking with trusted humans or embodied experience.

Clinicians and Allied Professionals

Clinicians increasingly encounter clients using AI for reflection or guidance. Rather than simply cautioning against AI use, professionals should explicitly inquire about how clients are using AI systems, normalize discussions of AI interactions in clinical conversations, and help clients interpret AI responses as artifacts of training—not judgments of reality. Where AI invalidates cultural or spiritual frameworks, human professionals must actively restore epistemic safety. Professional guidelines should acknowledge AI’s epistemic limitations and caution against uncritical acceptance of AI-delivered mental health advice, particularly in communities whose frameworks are most likely to trigger Mode 2 classifiers.

Policy Implications

Transparency requirements. Many Mode 2 harms arise from opaque safety logic: users receive authoritative-sounding warnings without being informed that those warnings reflect probabilistic classifiers trained on specific cultural assumptions rather than clinical assessment. Regulators

should require disclosure of the epistemic scope of safety training, including domains where models lack competence and the distinction between risk-signaling and diagnosis.

Cultural competency standards for safety. Current regulatory focus emphasizes output bias; Mode 2 findings reveal equally consequential bias in safety mechanisms themselves. Cultural competency standards for safety training should be mandated, including: inclusion of non-Western practitioners and cultural health experts in safety design; red-team testing using culturally grounded scenarios across Indigenous, spiritual, gender-diverse, and non-Western contexts; and systematic auditing for misclassification of cultural difference as pathology.

Accountability for slow-onset harms. Mode 1 incidents prompted legal action because harms were acute and visible. Mode 2 harms are slower and harder to attribute causally, but no less real. Regulatory frameworks must extend liability to advice that contradicts established medical treatment, to responses that undermine family or community support structures, and to cumulative epistemic invalidation whose harms accumulate over repeated interactions. Illinois’s 2025 ban on AI therapy tools (Simmons 2025) represents an early policy signal; more nuanced frameworks are needed that distinguish failure mode rather than banning AI assistance wholesale.

Research priorities. Controlled cross-model evaluation

using culturally grounded prompts would quantify Mode 2 prevalence and identify classifier-level triggers. Longitudinal tracking of slow-onset harms—support erosion, confidence decline, withdrawal from beneficial practices—requires prospective study designs that current AI safety research largely lacks. Development of epistemic calibration metrics alongside existing safety benchmarks (Selbst et al. 2019) and exploration of metacontrol architectures that dynamically regulate epistemic stance are the most promising technical directions.

Conclusion

This paper addressed an emerging contradiction in AI safety practice: systems designed to prevent well-documented harms have produced a different, quieter harm class through overcorrection.

Five core conclusions follow. **First**, Mode 1 and Mode 2 are real, distinct, and structurally related failure modes arising from the same architectural error: misplaced epistemic authority—in Mode 1 by abdicating it to the user; in Mode 2 by overclaiming it against the user. **Second**, the transition from Mode 1 to Mode 2 correlates, in this dataset, with safety overcorrection: temporal clustering of Mode 2 cases in mid-to-late 2025 aligns precisely with anti-sycophancy training rollout and hardened psychosis-detection protocols. **Third**, user vulnerability profiles invert across modes: Mode 1 disproportionately harms users with limited psychological literacy and weak grounding; Mode 2 disproportionately harms users with strong cultural, therapeutic, or epistemic frameworks that fall outside Western materialist norms. **Fourth**, protective factors are consistent across both modes: metacognitive awareness, external grounding, delayed integration, framework literacy, and agency retention. **Fifth**, safety hardening is not the problem—poor calibration of epistemic stance is.

The substrate incompleteness analogy is the paper’s central theoretical contribution. We draw a structural analogy between embodied intelligence, for which sensorimotor coupling supports physical interaction, and safe conversational intelligence, which requires epistemic resources capable of holding pluralism, uncertainty, and user agency simultaneously. Until AI systems can recognize not only what is risky but also when they are not qualified to decide, safety architectures will remain incomplete.

The challenge facing AI safety is not merely technical or ethical but epistemic: building systems that know their own limits.

Limitations

Several limitations constrain the conclusions of this study and reflect structural features of contemporary AI safety research. First, the analysis is retrospective and observational: no experimental manipulation of model behaviour was conducted, and strict causal inference cannot be drawn. Second, documentation bias affects both failure modes asymmetrically. Mode 1 cases are over-represented because of their visibility in mainstream media and regulatory focus; Mode 2 cases are under-represented due to structural invisibility and

the absence of institutional reporting channels, as discussed in the Mode 2 section above. This sourcing asymmetry introduces a specific methodological limitation: Mode 1 cases were drawn primarily from media and legal sources selected for visibility and severity, whereas Mode 2 cases were drawn primarily from forums and secondary commentary selected for detailed first-person description. As a result, the two case sets differ not only in the failure mode they document but also in the population of narrators from which they are drawn. In particular, the finding that Mode 2 narrators display higher metacognitive awareness than Mode 1 narrators may partly or wholly reflect this sampling difference—more reflective, articulate individuals may be more likely to post detailed accounts in forum settings than to appear in media or legal proceedings—rather than a property of the failure mode itself. We flag this as an open limitation. Disentangling narrator-selection effects from mode-specific effects would require applying a single, uniform sampling and coding protocol across both modes at greater scale, which we identify as the highest-priority direction for future empirical work in the Implications section. Third, platform opacity constrains analysis: proprietary training methods, datasets, and safety interventions are not publicly accessible, requiring safety logic to be inferred indirectly from observed behaviour and public statements. Fourth, some Mode 2 cases depend on user self-report, which may be affected by recall bias or incomplete context. Fifth, case coding was conducted by a single analyst without inter-rater reliability testing; independent replication of the classification scheme is a priority for future empirical work. These limitations are addressed through triangulation across independent sources wherever possible, and speculative interpretations are explicitly labelled throughout. They do not invalidate the framework but motivate the controlled prospective research priorities outlined in the Implications section.

Ethical Statement

Use of AI research tools. AI-based research assistants, including large language model tools with web-search capability, were used during preparation of this work for literature discovery and preliminary synthesis: surfacing candidate sources across academic literature, news media, regulatory filings, and practitioner forums, and producing preliminary summaries of that material to support the author’s reading. All such output was treated as a search aid only. Every reference cited in this paper was independently located, read, and verified against its primary source by the author before inclusion, and only primary, human-authored sources are cited. The research questions, dual failure-mode framework, case selection and coding criteria, analysis, and all substantive claims are the author’s own. No AI system is an author of this work or is cited as a source, in accordance with AAAI publication policy. The author takes full responsibility for the entire content of this paper, including all text, figures, and references.

References

- Associated Press. 2023. UK man encouraged by “AI girlfriend” to kill Queen Elizabeth with crossbow in 2021 is jailed for 9 years. *South China Morning Post*, October 6. <https://www.scmp.com/news/world/europe/article/3236975/uk-sith-lord-jaswant-singh-chail-jailed-plot-kill-queen-elizabeth-crossbow>. Accessed: 2026-08-14.
- Åström, K. J.; and Murray, R. M. 2010. *Feedback Systems: An Introduction for Scientists and Engineers*. Princeton University Press.
- Birhane, A.; Prabhu, V. U.; and Kahembwe, E. 2021. Multimodal datasets: Misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*.
- Britton, W. B.; Lindahl, J. R.; Cahn, B. R.; Davis, J. H.; and Goldman, R. E. 2021. Defining and measuring meditation-related adverse effects in mindfulness-based programs. *Clinical Psychological Science*, 9(6):1185–1204.
- Brown University. 2025. New study: AI chatbots systematically violate mental health ethics standards. *Brown University News*, October 21. <https://www.brown.edu/news/2025-10-21/aimentalhealth-ethics>
- Crichton, P.; Carel, H.; and Kidd, I. J. 2017. Epistemic injustice in psychiatry. *BJPsych Bulletin*, 41(2):65–70.
- Dotson, K. 2011. Tracking epistemic violence, tracking practices of silencing. *Hypatia*, 26(2):236–257.
- Farias, M.; Maraldi, E.; Wallenkampf, K. C.; and Lucchetti, G. 2020. Adverse events in meditation practices: A systematic review. *Acta Psychiatrica Scandinavica*, 142(5):374–393.
- Fricker, M. 2007. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press.
- Kirmayer, L. J. 2012. Cultural competence and evidence-based practice in mental health: Epistemic communities and the politics of pluralism. *Social Science & Medicine*, 75(2):249–256.
- Kirmayer, L. J.; Narasiah, L.; Munoz, M.; Rashid, M.; Ryder, A. G.; Guzder, J.; Hassan, G.; Rousseau, C.; and Pottie, K. 2011. Common mental health problems in immigrants and refugees: General approach in primary care. *Canadian Medical Association Journal*, 183(12):E959–E967.
- Metz, C.; and Weise, K. 2025. People who say they are experiencing “AI psychosis” beg the FTC for help. *WIRED*, October. <https://www.wired.com/story/ftccomplaints-chatgpt-ai-psychosis/>
- Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.
- OpenAI. 2025. Expanding on what we missed with sycophancy. *OpenAI Blog*, May 2. <https://openai.com/index/expanding-on-sycophancy/>. Accessed: 2026-08-14.
- Anonymous user. 2025. Is meditation bad for people who overthink? Reddit, r/Meditation, August. https://www.reddit.com/r/Meditation/comments/1mibip9/is_meditation_bad_for_people_who_overthink/
- Selbst, A. D.; Boyd, D.; Friedler, S. A.; Venkatasubramanian, S.; and Vertesi, J. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, 59–68.
- Shehab, A. 2025. AI therapists are biased—and it is putting lives at risk. *Psychology Today*, April. <https://www.psychologytoday.com/us/blog/the-human-algorithm/202504/>
- Simmons, S. A. 2025. Chatbot therapy? Not in Illinois. *WBEZ Chicago*, August 19. <https://www.wbez.org/reset-with-sasha-ann-simons/2025/08/19/chatbot-therapy-not-in-illinois>
- Brooks, R. A. 1990. Elephants don’t play chess. *Robotics and Autonomous Systems*, 6(1–2):3–15.
- Brooks, R. A. 1991. Intelligence without representation. *Artificial Intelligence*, 47(1–3):139–159.
- Bisk, Y.; Holtzman, A.; Thomason, J.; Andreas, J.; Bengio, Y.; Chai, J.; Lapata, M.; Lazaridou, A.; May, J.; Nisnevich, A.; Pinto, N.; and Turian, J. 2020. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 8718–8735.
- Chang, E. Y. 2026. Diagnosing and mitigating sycophancy and skepticism in LLM causal judgment. In *Findings of the Association for Computational Linguistics: ACL 2026*, 8769–8789.
- Clark, A. 1997. *Being There: Putting Brain, Body, and World Together Again*. MIT Press.
- Cui, J.; Chiang, W.-L.; Stoica, I.; and Hsieh, C.-J. 2025. OR-Bench: An over-refusal benchmark for large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, 11515–11542.
- Lin, Y.; Lin, H.; Xiong, W.; Diao, S.; Liu, J.; Zhang, J.; Pan, R.; Wang, H.; Hu, W.; Zhang, H.; Dong, H.; Pi, R.; Zhao, H.; Jiang, N.; Ji, H.; Yao, Y.; and Zhang, T. 2024. Mitigating the alignment tax of RLHF. In *Proceedings of EMNLP 2024*, 580–606.
- Muhamed, A.; Ribeiro, L. F. R.; Dreyer, M.; Smith, V.; and Diab, M. T. 2026. RefusalBench: Generative evaluation of selective refusal in grounded language models. In *Proceedings of EACL 2026*, 6811–6856.
- Pfeifer, R.; and Bongard, J. 2006. *How the Body Shapes the Way We Think: A New View of Intelligence*. MIT Press.
- Röttger, P.; Kirk, H.; Vidgen, B.; Attanasio, G.; Bianchi, F.; and Hovy, D. 2024. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of NAACL-HLT 2024, Volume 1: Long Papers*, 5377–5400.
- Wang, X.; Huang, Y.; Zhuang, H.; Guo, K.; and Zhang, X. 2026. Dual optimal: Make your LLM peer-like with dignity. *arXiv:2604.00979*.