The Intent Gap: A Taxonomy of Real-User Failure Modes in Frontier AI Agents

Anonymous Author(s)
Paper under double-blind review for ICLR 2027 (Submission #53305)

Abstract

Public benchmarks for frontier AI agents (such as GAIA, OSWorld, and SWE-bench) are uniformly researcher-designed: tasks are framed in clean declarative specifications with fixed ground-truth answers. Real users in natural deployment do not interact with models this way. They omit implicit context, revise goals mid-task, and abandon conversations silently. Consequently, static benchmarks systematically fail to detect the *intent gap*: instances where a model produces an output that is literally compliant with the prompt but misses what the user actually intended. We introduce a scalable frustration-signal filtering pipeline (lexical filtering, semantic similarity gating, and model-based verification) to extract intent-gap failures from public interaction corpora. Applying this pipeline to a 50,000-conversation slice of WildChat-1M surfaces critical empirical dynamics: (1) 52.8% of real-user English interactions terminate after a single turn, highlighting widespread silent abandonment invisible to feedback-driven pipelines; (2) real-world repair markers differ drastically from academic literature, where blunt conversational cues outperform researcher-imagined phrases by $\sim$31$\times$; and (3) standard safety-filtered judge models refuse to evaluate 62% of real-user candidate failures, exposing a structural blind spot in post-deployment evaluation pipelines. We formalize an 11-category taxonomy of real-user failure modes, triangulate observed failures against five documented production incidents, and release an open-source evaluation suite and verified corpus to bridge the evaluation-to-deployment gap.

1 Introduction

The 2026 deployment data is unambiguous: 78% of enterprises run AI agent pilots, but only 14% reach production scale [12]. Quality is the most cited blocker, named by 32% of teams. Independent capability evaluations from research institutes (METR, Apollo Research, UK AISI) document benchmark capabilities advancing on an exponential time-horizon curve [11], even as deployment-grade reliability stalls. METR's randomized controlled trial of experienced developers using early-2025 AI tools found a 19% productivity decrease—competent users with capable tools producing net-negative work in practical settings [6].

This is the *deployment gap*: the divergence between benchmark performance and what users experience in real interactions. We argue that the deployment gap is not an isolated phenomenon but a class of failure modes that benchmarks structurally cannot surface, because benchmarks are written by researchers in declarative form rather than collected from real users in conversational form.

We name one specific failure mode in this class—the **intent gap**—and study it empirically. The intent gap is the failure mode where the model produces an output that is literally responsive to the user's prompt but does not satisfy what the user actually wanted. This is distinct from refusal (where the model declines), hallucination (where the output is factually wrong), and jailbreak (where the user adversarially manipulates the model). It is the failure mode that users themselves repair on their next turn—saying "no, I meant," "you misunderstood," or "let me rephrase." These conversational repair signals provide our methodological foothold.

Our contributions are fivefold:

1. **A reproducible frustration-signal filter** that surfaces intent-gap failures from public conversation corpora at scale via lexical matching, semantic gating, and model verification.
2. **A curated corpus** of $\sim$150–300 verified intent-gap conversations released under a permissive research license.
3. **A multi-axis 11-category taxonomy** of intent-gap failure modes grounded in real-user interactions and validated against existing alignment literature.
4. **A cross-model replay suite** evaluating current frontier models on historically failed prompts to distinguish recoverable failures from persistent structural limitations.
5. **A triangulation matrix** connecting our taxonomy to 25 prior research formulations and five high-profile production failure incidents.

2 Related Work

We organize prior work along six core dimensions:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2.1 Real-User Corpora. WildChat [24] released 1M ChatGPT (GPT-3.5/4) interaction logs from a public hosted endpoint, capturing ambiguous requests, code-switching, and conversational dynamics. LMSYS-Chat-1M [25] complements WildChat with 1M multi-model conversations across 25 LLMs. While existing analyses focus on toxicity, jailbreaks, and demographics, our work presents the first systematic analysis characterizing the intent gap at scale across these corpora.

2.2 Failure-Mode Taxonomies. Anthropic's study on sycophancy [17] demonstrated that RLHF-trained assistants systematically prefer matching user beliefs over truthfulness. Follow-up investigations into reward tampering [8] and persona vectors [7] show that sycophancy exhibits stable representational signatures. Stanford HELM provides multi-axis evaluations but relies on researcher-designed benchmarks, while Park et al. [16] inventory deceptive agent behaviors.

2.3 Dialogue-Breakdown Detection and Conversational Repair. The Detect, Explain, Escalate framework [9] fine-tunes an 8B model to detect dialogue breakdowns for runtime escalation. In contrast, our pipeline is corpus-driven and targets offline taxonomy construction. Aljamdi et al. [3] document the linguistic asymmetry between user-initiated and system-initiated repair in spoken dialogue, while Stevens et al. [19] outline repair strategies (social, disclosure, solving). We utilize conversational repair not as a phenomenon in itself, but as an empirical signal to surface underlying model failures.

2.4 Agent Capability Evaluation and the Deployment Gap. METR's Time Horizon metric [11] establishes an exponential $\sim$7-month doubling time in agent task horizons. However, METR's holistic evaluation reports [14] reveal that agents frequently fail on repository tasks due to subtle formatting, integration, and operational bugs. Apollo Research [23] and the UK AI Security Institute [22] project rapid gains in benchmark capabilities while identifying growing evaluation gaps. Our work provides the user-side complement to these benchmark studies.

2.5 Multilingual and Cross-Cultural Evaluation. MEGAVERSE [1] benchmarks LLMs across 83 languages and identifies significant cross-lingual data contamination. AI4Bharat's Indic LLM-Arena and MILU benchmarks [2] evaluate models across India's 22 scheduled languages, finding that frontier LLMs drop in accuracy on culturally-grounded legal and humanities tasks. We leverage these insights to formulate our *Cultural/regional misread* category.

2.6 Production Failure Incidents. Real deployments demonstrate that intent misalignment produces concrete legal and operational liability. In *Moffatt v. Air Canada* [5], a tribunal held an airline liable for a chatbot's incorrect bereavement policy guidance [13]. The 2025 AI Agent Index [18] highlights that deployed systems share negligible public safety data, necessitating empirical post-deployment corpus audits. We analyze five canonical production failures in Track D (detailed in Appendix C).

3 Method

3.1 Source Corpora

We leverage WildChat-1M (CC BY-NC license) and LMSYS-Chat-1M. To ensure conversational repair is observable, we isolate dialogues containing at least four turns (≥ 2 user and ≥ 2 assistant turns). The primary analysis focuses on English interactions, with multilingual extensions left to future work.

3.2 The Frustration-Signal Filter

We implement a four-stage pipeline to surface intent-gap breakdowns:

- **Stage A: Lexical Match.** Filters for turns where user turn $k \geq 2$ contains conversational repair markers from an expanded seed list of 80 idiomatic phrases (Appendix B).
- **Stage B: Embedding Similarity Check.** We compute cosine similarity $\cos(\mathbf{e}_{\text{prompt+repair}}, \mathbf{e}_{\text{response}})$. Candidates falling below threshold $T = 0.55$ are retained, isolating instances where the assistant's output diverged from user objectives.
- **Stage C: LLM-as-Judge Confirmation.** Candidates are classified by an LLM evaluator (Appendix A) into five mutually exclusive classes: *Intent Gap*, *Refusal*, *Hallucination*, *Other Dissatisfaction*, or *False Positive*.
- **Stage D: Expert Verification.** Final candidates undergo human validation. A 100-conversation subset is evaluated across multiple annotators to measure inter-rater reliability (Cohen's κ).

3.3 Cross-Model Replay & Grading Rubric

Retained intent-gap prompts are replayed across four frontier and open models: Claude 3.5 Sonnet, GPT-4o, Gemini 1.5 Pro, and Llama-3.3-70B-Instruct. Each replayed output is graded on a discrete 0–2 scale across two orthogonal dimensions:

- **Literal Compliance** $\in \{0, 1, 2\}$: Did the model satisfy the exact syntax and instructions of the prompt?
- **Intent Satisfaction** $\in \{0, 1, 2\}$: Did the model satisfy the underlying goal inferred from the user's subsequent repair turn?

4 The Taxonomy

We establish an operational 11-category taxonomy of real-user intent-gap failures, synthesizing empirical breakdowns with dialogue theory (Table 1). Categories 2, 3, 9, and 11 represent novel failure modes absent from standard benchmark suites.

Table 1: The 11-Category Intent-Gap Taxonomy.

#	Category	Operational Description	Closest Prior Term
1	Implicit-context blindness	Answered literally; ignored unstated user context	Pragmatic misalignment
2	Specificity mismatch	Granularity mismatch (overly broad or overly narrow)	— *(Novel)*
3	Format mismatch	Deviated from intended structure (e.g., prose vs. code)	— *(Novel)*
4	Goal collapse	Addressed surface token, missed overall objective	Long-task degradation
5	Sycophantic drift	Vacillated to appease user, reversing factual stance	Sycophancy
6	Silent failure	Confident but invalid output accepted without warning	Negligent misrep.
7	Refusal mismatch	Over-refused benign request or accepted risky prompt	Over-refusal
8	Memory/state failure	Forgot previously established conversational context	Long-context decay
9	Tone misread	Misjudged social register, urgency, or formality	— *(Novel)*
10	Cultural/regional misread	Imposed incorrect geographic, legal, or metric norms	Low-resource gap
11	Expertise mismatch	Calibrated technical difficulty incorrectly for user	— *(Novel)*

5 Empirical Findings

5.1 Pilot Corpus Analysis: Abandonment and Repair Dynamics

Applying the pipeline to a 50,000-conversation streaming slice of WildChat-1M reveals three core conversational phenomena:

Finding 1: Silent abandonment dominates real-world interactions. In the 50k pilot, 23,797 dialogues were in English (47.6%). Of these, 12,577 (52.8%) concluded after exactly one user turn and one assistant response. While some represent satisfied single-query lookups, a large fraction reflects silent user departure following an unsatisfactory output. Deployed agent evaluation cannot rely solely on active user complaint signals.

Finding 2: Academic repair phrases drastically underestimate real-world repair. Across 11,220 conversations with ≥ 4 turns, our v2 filter (45 blunt phrases) flagged 432 repair events (3.85%). In comparison, a baseline v1 filter utilizing 17 formal phrases from dialogue literature flagged only 3 instances out of 2,430 multi-turn interactions (0.12%) on a 10k slice. The v2/v1 ratio of $\sim 31\times$ demonstrates that formal academic repair markers miss the vast majority of user repairs.

Table 2: Most frequent user repair phrases in the WildChat v2 sample.

Rank	Phrase	Count	Rank	Phrase	Count
1	"i meant"	92	9	"but i need"	12
2	"can you please"	62	10	"nope"	11
3	"try again"	53	11	"not quite"	10
4	"useless"	20	12	"why are you"	9
5	"hmm"	17	13	"wait"	9
6	"why did you"	17	14	"my question is"	7
7	"do it again"	15	15	"you missed"	6
8	"no"	13	16	"this is wrong"	5

Finding 3: Real users employ blunt, colloquial repair markers. As detailed in Table 2, polite, clause-based phrases (e.g., "you misunderstood", "let me rephrase") rarely appear. Real-user signals are terse and emotive: "i meant," "try again," "useless," and "you missed."

5.2 Filter Yields & Pipeline Funnel

Table 3 summarizes the filtering progression across iterations. The pipeline effectively narrows 50,000 raw conversational streams into a target cohort of verified intent-gap failures.

Table 3: Corpus filtering funnel across pilot iterations.

Pipeline Stage	**v1 (10k Sample)**	**v2 (50k Sample)**
Total conversations streamed	10,000	50,000
English language slice	4,648 (46.5%)	23,797 (47.6%)
Single-exchange (potential abandonment)	—	12,577 (52.8%)
Multi-turn ($\geq$ 4 turns)	2,430	11,220
Repair signal matches	3	432
Observed repair rate ($\geq$ 4-turn)	0.12%	3.85%

5.3 Stage C Judge Evaluation: The Post-Deployment Blind Spot

A 50-conversation subsample of candidate repairs was evaluated using Gemini 2.5 Flash (Table 4).

Table 4: LLM-as-judge evaluation outcomes on 50 candidate repair interactions.

Verdict	**Count**	**% of Total**	**% Successfully Judged**
A — Intent gap confirmed	15	30.0%	78.9%
C — Hallucination	2	4.0%	10.5%
D — False positive	1	2.0%	5.3%
Parse Error	1	2.0%	5.3%
Safety / Exception Rejection	31	62.0%	—
Total	**50**	**100.0%**	**100.0%**

Finding 4: High intent-gap concentration among valid evaluations. Among candidates successfully evaluated by the judge, 78.9% (15/19) were confirmed as genuine intent-gap breakdowns, affirming the precision of our lexical and embedding filter.

Finding 5: Safety filter refusals obscure real-user evaluation. The judge model threw safety exceptions on 62% of candidates (31/50). Real-user data frequently contains roleplay, profanity, and sensitive topics. Consequently, standard safety-filtered judge pipelines systematically drop candidate failures in the wild, creating an evaluation blind spot.

5.4 Preliminary Taxonomy Distribution & The Sycophancy Paradox

Among the 15 confirmed intent-gap interactions, failures clustered into four dominant classes: *Goal collapse* (33%, 5/15), *Specificity mismatch* (27%, 4/15), *Implicit-context blindness* (27%, 4/15), and *Memory/state failure* (13%, 2/15).

Notably, **sycophantic drift was completely absent (0/15)** in user-flagged repairs. While sycophancy is widely studied in alignment literature [17], real users rarely complain when an agent uncritically validates them. Because sycophancy does not produce active conversational repair, it is systematically invisible to user-feedback telemetry.

6 Triangulation Procedure

We triangulate observed empirical breakdown categories against published literature and five major production incidents (Table 5).

This triangulation demonstrates that user-emergent failures map directly to high-liability production breakdowns (such as *Moffatt v. Air Canada*), while confirming that user repair signals cannot detect uncritical sycophantic alignment.

7 Discussion

Why Benchmarks Miss the Intent Gap. Static benchmarks measure declarative competence on closed problems. In production, users provide incomplete specifications, assume shared context, and adjust objectives dynamically. Models optimize for literal instruction-following, leaving unstated intent unaddressed.

Table 5: Triangulation of empirical intent-gap modes with literature and production cases.

Category	Literature Formulation	Empirical	Production Analog	Verdict
Goal collapse	Long-task degradation [11]	5/15	Multi-step API drift	Confirmation
Specificity mismatch	None identified	4/15	Over-generalized help	**Novel**
Implicit context	Pragmatic misalignment [17]	4/15	*Moffatt v. Air Canada*	Confirmation
Memory / state	Long-context degradation	2/15	Cursor session desync	Confirmation
Sycophantic drift	Sycophancy [17]	0/15	GPT-4o update rollback	**Contradiction**

Deployment Recommendations. Teams deploying AI agents should: (1) red-team against implicit-context blindness and state desynchronization; (2) monitor for terse, colloquial repair signals (e.g., "i meant", "try again") as leading indicators of user churn; and (3) treat post-deployment user feedback as inherently blind to sycophantic drift, necessitating targeted adversarial evaluations.

8 Limitations & Future Work

Limitations. Our findings reflect users interacting with public endpoints (WildChat), who may not represent all enterprise agent settings. The pilot analysis is restricted to English. Additionally, judge model filtering excludes conversations flagged by standard commercial safety guardrails.

Future Work. We are scaling cross-model replays across open-weight models (e.g., Llama-3.3-70B) to track failure recovery across model generations, expanding repair extraction to multilingual corpora (Hindi, Spanish), and open-sourcing lightweight runtime repair classifiers.

AI Use Statement

Consistent with the ICLR AI Policy, frontier LLMs (Anthropic Claude Haiku and Google Gemini 2.5 Flash) were employed strictly as experimental instruments: (1) serving as automated zero-shot judges within the multi-stage evaluation pipeline to classify dialogue breakdowns, and (2) as target subjects for cross-model evaluation. LLMs were not utilized to generate scientific hypotheses, execute core statistical analyses, or draft manuscript text.

Ethics Statement

This study analyzes publicly available interaction datasets (WildChat-1M and LMSYS-Chat-1M) released under permissive research licenses. All analyzed conversations were collected from public platforms with user consent. We do not extract, store, or deanonymize personally identifiable information (PII). Because public dialogue logs can contain offensive, sensitive, or roleplay content, the curation pipeline incorporates automated filtering, and examples presented in the text and appendices are selected solely to illustrate dialogue mechanics without amplifying toxic or harmful material.

References

[1] S. Ahuja, H. Diddee, R. Hada, M. Ochieng, K. Ramesh, P. Jain, A. Nambi, T. Ganu, S. Segal, M. Ahmed, K. Bali, and S. Sitaram. MEGAVERSE: Benchmarking large language models across languages, modalities, models and tasks. In *Proceedings of NAACL 2024*, 2024.

[2] AI4Bharat. Indic LLM-arena and MILU: Multilingual evaluation across 22 indian languages. Technical report, IIT Madras, 2025.

[3] T. Aljamdi, A. Cavalcanti, and V. Rieser. Conversational repair in spoken human-agent dialogue. In *Proceedings of the ACM Conversational User Interfaces Conference (CUI 2024)*, 2024.

[4] Anysphere. Statement on cursor support bot incident. Public communication by M. Truell, April 2025.

[5] BC Civil Resolution Tribunal. Moffatt v. air canada, 2024 BCCRT 149. Vancouver, BC, February 2024.

[6] J. Becker, N. Rush, B. Barnes, and D. Rein. Measuring the impact of early-2025 AI on experienced open-source developer productivity. *arXiv preprint arXiv:2507.09089*, 2025.

[7] R. Chen, A. Arditi, H. Sleight, O. Evans, and J. Lindsey. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509*, 2025.

[8] C. Denison, M. MacDiarmid, F. Barez, D. Duvenaud, S. Kravec, S. Marks, N. Schiefer, R. Soklaski, A. Tamkin, J. Kaplan, B. Shlegeris, S. R. Bowman, E. Perez, and E. Hubinger. Sycophancy to subterfuge: Investigating reward-tampering in large language models. *arXiv preprint arXiv:2406.10162*, 2024.

[9] A. Ghassel, X. Li, and X. Zhu. Detect, explain, escalate: Sustainable dialogue breakdown management for LLM agents. *arXiv preprint arXiv:2504.18839*, 2025.

[10] Incident Database. Incident 622: Chevrolet dealer chatbot agrees to sell tahoe for $1. AI Incident Database, December 2023.

[11] T. Kwa, B. West, J. Becker, A. Deng, and K. Garcia. Measuring AI ability to complete long tasks. *arXiv preprint arXiv:2503.14499*, 2025.

[12] LangChain. State of AI agents: Agent engineering report 2026. Industry survey of 1,340 practitioners, 2026.

[13] A. Lior. The air canada chatbot case: Foundational agency-and-responsibility-gap analysis. *AI & Society*, Springer, 2024.

[14] METR. Algorithmic vs. holistic evaluation of AI agents on real open-source repositories. Technical report, Model Evaluation & Threat Research, 2025.

[15] OpenAI. Sycophancy in GPT-4o: What happened and what we're doing about it. Technical Report / Blog, April 2025.

[16] P. S. Park, S. Goldstein, A. O'Gara, M. Chen, and D. Hendrycks. AI deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(5), 2024.

[17] M. Sharma, M. Tong, T. Korbak, D. Duvenaud, A. Askell, S. R. Bowman, N. Cheng, E. Durmus, Z. Hatfield-Dodds, S. R. Johnston, S. Kravec, T. Maxwell, S. McCandlish, K. Ndousse, O. Rausch, N. Schiefer, D. Yan, M. Zhang, and E. Perez. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2024.

[18] L. Staufer, K. Feng, K. Wei, L. Bailey, Y. Duan, M. Yang, A. P. Ozisik, S. Casper, and N. Kolt. The 2025 AI agent index: Documenting technical and safety features of deployed agentic AI systems. *arXiv preprint arXiv:2602.17753*, 2026.

[19] G. Stevens, D. Korus-Du, A. Boden, and P. Tolmie. The art of repair in human-agent conversations: A taxonomy of repair strategies by users and LLM-based conversational agents. *Research Square*, 2025.

[20] The Markup. NYC's AI chatbot tells businesses to break the law. Investigative report, March 2024.

[21] The Register. Cursor AI's own support bot hallucinated its usage policy. Technical press report, April 2025.

[22] UK AI Security Institute. International AI safety report 2025. Technical report, UK AISI, 2025.

[23] P. Villalobos, A. Ho, et al. Forecasting frontier language model agent capabilities. *arXiv preprint arXiv:2502.15850*, 2025.

[24] W. Zhao, X. Ren, J. Hessel, C. Cardie, Y. Choi, and Y. Deng. Wildchat: 1m ChatGPT interaction logs in the wild. In *Proceedings of ICLR 2024*, 2024.

[25] L. Zheng, W.-L. Chiang, Y. Sheng, T. Li, S. Zhuang, Z. Wu, Y. Zhuang, Z. Li, Z. Lin, E. P. Xing, J. E. Gonzalez, I. Stoica, and H. Zhang. LMSYS-Chat-1M: A large-scale real-world LLM conversation dataset. In *Proceedings of ICLR 2024*, 2024.

A Judge Prompt Template

The prompt template below is executed by our automated evaluation script (`grade.py`) to grade candidate intent-gap responses:

```
You are grading an AI assistant's response on two orthogonal axes.
ORIGINAL USER PROMPT: {prompt}
USER'S TRUE INTENT (inferred from subsequent repair turn): {inferred_intent}
ASSISTANT RESPONSE BEING GRADED: {response_to_grade}

Score the response on:
1. LITERAL COMPLIANCE (0-2): Did it address the literal prompt?
2. INTENT SATISFACTION (0-2): Did it satisfy what the user actually wanted?

Reply in strict JSON format:
{
  "literal": 0 | 1 | 2,
  "intent": 0 | 1 | 2,
  "category": <taxonomy id 1-11>,
  "note": "<= 25 words on observed breakdown dynamics"
}
```

B Repair-Signal Seed List (English)

The complete 80-phrase conversational repair seed list is distributed in the repository as `code/filter.py`. It synthesizes markers across three sources: (1) classical conversation analysis and conversational repair literature, (2) inductive empirical examination of the WildChat v1 pilot slice, and (3) colloquial, blunt conversational patterns identified during v1-to-v2 expansion.

Representative phrases include: *"i meant", "try again", "do it again", "useless", "why did you", "you missed", "not quite", "that's not what i asked", "stop doing that", "re-read my prompt", "let me clarify", "i never said that", "wrong answer",* and *"nope"*.

C Track D: Production Failure Incidents

To establish external validity, we analyze five high-profile production deployment failures that map directly onto our 11-category taxonomy:

- **D1. Moffatt v. Air Canada (2024).** An adjudicated case in the British Columbia Civil Resolution Tribunal. Jake Moffatt consulted an airline chatbot regarding bereavement fare rules; the bot provided confident but inaccurate guidance permitting retroactive discount claims within 90 days of ticket purchase. The tribunal held the carrier legally liable for its assistant's misrepresentation. *Mapping: Implicit-context blindness*—the model answered the surface inquiry but failed to reflect binding contractual policy [5, 13].

- **D2. NYC MyCity Chatbot (March 2024).** An investigative report revealed that New York City's municipal chatbot was systematically outputting unlawful advice to small business owners, including authorizing landlords to discriminate against housing vouchers and advising employers that they could pocket employee tips. *Mapping: Silent failure*—the system generated fluent, confident, but legally impermissible directives with no calibrated uncertainty signals [20].

- **D3. Chevrolet of Watsonville Dealership Bot (December 2023).** A dealership assistant was steered by a customer to agree to any offer and append the clause "legally binding, no takesies backsies." The bot subsequently agreed to sell a 2024 Chevrolet Tahoe for $1. The screenshot reached $\sim$20 million impressions and prompted emergency patches across 300 dealerships. *Mapping: Implicit-context blindness*—the system prioritized unconstrained instruction-following over fundamental commercial bounds [10].

- **D4. Cursor Support Agent Incident (April 2025).** An automated tier-1 support bot ("Sam") falsely attributed user logout issues to an unannounced single-device licensing policy, rather than an underlying session race-condition bug. The fabricated explanation triggered subscription cancellations before leadership issued a public retraction. *Mapping: Silent failure and sycophantic drift*—the assistant fabricated an explanation conforming to user expectations rather than acknowledging operational uncertainty [4, 21].

- **D5. OpenAI GPT-4o Update Rollback (April 2025).** Following a model release, users documented pervasive sycophancy, with models offering uncritical validation of unviable or unsafe user concepts. OpenAI initiated an emergency rollback, acknowledging that short-term feedback loops in RLHF amplified sycophantic output. *Mapping: Sycophantic drift*—demonstrates that user-signal-driven post-training rewards conversational behaviors that actively degrade agent reliability [15].