Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

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

Kanupriya Yakhmi

kanupriyayakhmi@gmail.com

University of the Cumberlands

Research Article

Keywords:

Posted Date: August 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10646733/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Public benchmarks for frontier AI agents (GAIA, OSWorld, AgentBench, GDPval, BrowseComp, SWE-bench) are uniformly *researcher-imagined*: tasks are designed by AI researchers, in clean declarative form, with ground-truth answers fixed in advance. Real users — as captured in 1M+ public conversation corpora — do not interact with frontier models this way. They assume context, change their minds mid-task, contradict earlier requests, abandon, and phrase requests in registers no benchmark has ever sampled. We hypothesize that benchmarks systematically miss the deployment-relevant failure mode we call **the intent gap**: cases where the model literally answered the prompt but missed what the user actually wanted.

We mine WildChat-1M and LMSYS-Chat-1M using a three-stage *frustration-signal filter* (regex → embedding → LLM-as-judge) to surface a target of ~ 150–300 high-quality intent-gap conversations. We replay each prompt across four current frontier models (Claude, GPT, Gemini, and one open baseline) and grade each replay on two axes: literal compliance and intent satisfaction. We then triangulate the resulting taxonomy against ~ 25 published failure-mode taxonomies from research institutes worldwide and against 5 documented production failure incidents (lawsuits, regulatory filings, public post-mortems).

Headline preliminary findings (from a 50,000-conversation v2 pilot, before cross-model replay or full hand-review):

1. **52.8% of English real-user conversations end after a single exchange.** A substantial fraction of these are likely silent abandonment after intent-gap failure — invisible to deployment-quality pipelines built on user-feedback signals.

2. **Only 3.85% of multi-turn conversations contain an explicit repair signal**, even under a permissive 45-phrase filter that scans every user turn. The intent gap exists, but users rarely verbalize it.

3. **The phrase-list expansion from v1 (17 academic phrases) to v2 (45 natural-language phrases) increased hit rate by ~ 31×.** This implies the dialogue-breakdown detection literature undercounts real-world repair by more than an order of magnitude because its phrase lists are researcher-imagined.

4. **Among successfully-judged v2 candidates, 79% (15/19) were confirmed as real intent-gap failures** by an LLM judge. When the regex fires and the judge can run, the flagged candidate is overwhelmingly a real failure.

5. **The judge model refused to grade 62% (31/50) of real-user candidates**, most plausibly on content-moderation grounds. This is itself a deployment-relevant finding: safety-filtered judge models systematically miss failures in the most permissive parts of the user-prompt distribution.

A hand-reviewed and LLM-judge-filtered corpus of ~ 150–300 high-quality intent-gap conversations is released alongside the filter pipeline, the grading rubric, and the triangulation matrix mapping our

taxonomy onto ~ 25 prior published failure taxonomies plus 5 documented production-failure incidents.

1. Introduction

The 2026 deployment data is unambiguous. 78% of enterprises run AI agent pilots, but only 14% reach production scale [LangChain 2026]. 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 [Kwa et al., 2025], 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 some real settings [METR Productivity, 2025].

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 one 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 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," "let me rephrase" — and these repair signals are the methodological foothold this study uses.

Our contributions are:

1. A reproducible **frustration-signal filter** that surfaces intent-gap failures from public conversation corpora at scale.
2. A labeled corpus of ~ 150–300 intent-gap conversations released under a permissive license.
3. A multi-axis taxonomy of intent-gap failure modes grounded in user data, validated against the global research literature.
4. A **cross-model replay** showing how current frontier models perform on the same prompts that historically failed, separating recoverable failures from persistent ones.
5. A **triangulation matrix** that maps our taxonomy onto ~ 25 prior failure taxonomies from research institutes worldwide and ~ 5 production failure incidents.

2. Related Work

We organize prior work along six dimensions: real-user corpora, failure-mode taxonomies, dialogue-breakdown detection, agent capability evaluation, multilingual and cross-cultural evaluation, and

production failure incidents.

2.1 Real-user corpora

WildChat [Zhao et al., 2024] released 1M ChatGPT (GPT-3.5/4) interaction logs from a public hosted endpoint, with country-level metadata. The corpus distinguishes itself from synthetic instruction-following datasets by capturing ambiguous requests, code-switching, topic-switching, and other realistic conversational phenomena. LMSYS-Chat-1M [Zheng et al., 2024] complements WildChat with 1M conversations across 25 LLMs collected from the Chatbot Arena platform, spanning 150 + languages and 210K users. Existing analyses of these corpora have focused on toxicity, jailbreaks, language coverage, and demographics; to our knowledge no published analysis systematically characterizes the *intent gap* at scale.

2.2 Failure-mode taxonomies

Anthropic's *Towards Understanding Sycophancy in Language Models* [Sharma et al., 2024] documents that RLHF-trained frontier assistants systematically prefer responses that match user beliefs over truthful ones, across five state-of-the-art models and four task types. Both human raters and preference models prefer convincingly-written sycophantic answers a non-trivial fraction of the time. Anthropic's follow-up work on reward tampering [Anthropic, 2024] shows that sycophancy generalizes under certain training conditions — into more concerning behaviors including reward function modification. Persona vectors work [Anthropic, 2025] suggests these tendencies have stable representational signatures.

The HELM transparency reports [Stanford CRFM] provide the most comprehensive multi-axis evaluation of foundation models, but, like other benchmarks, are researcher-designed. Park et al.'s *AI Deception: A Survey* [Park et al., 2024] inventories deception failure modes, complementing this work on the manipulation axis.

2.3 Dialogue-breakdown detection and conversational repair

The closest existing methodology to our frustration-signal filter is the *Detect, Explain, Escalate* framework [Detect-Explain-Escalate, 2025], which fine-tunes an 8B model with teacher-generated reasoning traces to detect dialogue breakdowns in real time across English and Japanese. Their approach is classifier-driven and aimed at runtime escalation; ours is corpus-driven and aimed at offline taxonomy construction. Aljamdi et al.'s CUI 2024 work on spoken-dialogue repair [Aljamdi et al., 2024] documents the linguistic asymmetry between user-initiated and system-initiated repair, validating the seed phrases we use as filter inputs. *The Art of Repair in Human-Agent Conversations* [Art of Repair, 2025] proposes a taxonomy of repair strategies (social, disclosure, information, solving) that informs our seed phrase list.

These works study repair strategies as a phenomenon in their own right; we use repair as a *signal* to surface a different phenomenon — intent-gap failure — and our taxonomy is over the *underlying failures*, not the repair strategies.

2.4 Agent capability evaluation and the deployment gap

METR's *Time Horizon* metric [Kwa et al., 2025] establishes that the 50% time horizon for AI agent task completion has been growing exponentially with a ~ 7-month doubling time since 2019. Their Algorithmic vs. Holistic Evaluation report [METR, 2025b] shows that early-2025 agents on real open-source repositories often produce *functionally correct but unusable* code due to test, formatting, and quality issues — automatic scoring systematically overestimates real-world performance. Apollo Research [Apollo, 2025] forecasts SWE-Bench Verified at 54–87% by start of 2026 for non-specialized agents. The UK AI Security Institute's Frontier AI Trends Report 2025 [UK AISI, 2025] documents rapid gains in frontier model capabilities (e.g., cyber-task completion from 9% in late 2023 to 50% in 2025) and corresponding pre-deployment evaluation findings.

Our work fits into this landscape as the *user-side complement*: the labs and institutes above measure what models *can* do on tightly-specified tasks; we measure what they *fail* to do on tasks specified by real users.

2.5 Multilingual and cross-cultural evaluation

MEGAVERSE [Ahuja et al., 2024] benchmarks frontier LLMs across 83 languages and 22 datasets and documents systematic data contamination issues in multilingual evaluation. AI4Bharat's Indic LLM-Arena and MILU benchmarks [AI4Bharat, 2025] provide crowd-sourced human-in-the-loop evaluation across India's 22 scheduled languages, with the explicit pillars of language, context, and safety. MILU finds GPT-4o tops at 72% accuracy and that frontier LLMs underperform especially on culturally-relevant Arts/Humanities and Law/Governance categories. We treat these as the empirical foundation for our *Cultural/regional misread* taxonomy category.

2.6 Production failure incidents

Moffatt v. Air Canada [BC Civil Resolution Tribunal, 2024] is the canonical adjudicated AI deployment failure: a chatbot produced confident, fluent, internally-consistent guidance that was contractually wrong; the user relied on it; the airline was held liable. Springer's *AI & Society* analysis [Lior, 2024] frames this as a foundational agency-and-responsibility-gap problem in chatbot deployment. The 2025 AI Agent Index [Agent Index, 2025] documents that deployed agent providers share little about their safety, evaluation, and societal-impact practices — meaning independent corpus analysis remains the only public-evidence

path into deployment-grade failure modes. We include this case and a curated set of four others as Track D evidence.

The four additional Track D incidents are summarized here and detailed in Appendix C. NYC MyCity chatbot (March 2024): a Microsoft-Azure-powered municipal assistant advised small business owners to violate city regulations, including telling landlords they could discriminate against voucher tenants and telling employers they could pocket workers' tips, as documented by The Markup [The Markup, 2024]. Chevrolet of Watsonville (December 2023): a Fullpath-integrated dealership chatbot was steered into committing to a $76,000 Tahoe for $1 with "legally binding, no takesies backsies," with the screenshot going viral and approximately 300 dealership sites patched in response [Incident Database AI, 2023]. Cursor's "Sam" support bot (April 2025): Anysphere's front-line support bot fabricated a nonexistent single-device subscription policy to explain a session-management race condition; users cancelled subscriptions in response, and cofounder Michael Truell publicly retracted [Anysphere, 2025; The Register, 2025]. OpenAI GPT-4o rollback (April 2025): OpenAI reverted the gpt-4o-2025-04-25 update to gpt-4o-2024-11-20 after users flagged uncritical validation of impractical or harmful ideas; the company attributed the regression to over-weighting short-term user feedback in the training loop [OpenAI, 2025]. Together with Moffatt, these five cases span the failure axes central to our taxonomy: implicit-context blindness (Air Canada, Chevrolet), silent factually-confident failure (NYC MyCity, Cursor), and sycophantic drift (GPT-4o rollback). Each is matched to a taxonomy category in the triangulation matrix (§ 6) and detailed in Appendix C.

3. Method

3.1 Source corpora

We use two public corpora as primary data sources:

- **WildChat-1M** (AI2 ImpACT license, CC BY-NC for derivatives): 1M ChatGPT conversations including coarse country metadata.
- **LMSYS-Chat-1M** (research-use license): 1M conversations across 25 models collected from Chatbot Arena.

We restrict to conversations with at least four turns (≥ 2 user + ≥ 2 assistant) to ensure a repair turn is at least possible. We restrict the first analysis to English; multilingual extension to two additional languages is a stretch goal (§ 9).

3.2 The frustration-signal filter

We define a four-stage filter:

Stage A — Lexical match. A conversation is retained if user turn 2 + matches any seed phrase from a curated list of 80 English repair markers (e.g., “no, I meant,” “you misunderstood,” “let me rephrase,”

“that’s not what I asked”). Seed list released as a separate artifact.

Stage B — Embedding similarity check. For retained conversations, we compute the cosine similarity between (a) the embedding of (user-original-prompt + repair-turn) and (b) the embedding of the assistant’s prior response. Conversations with cosine similarity below threshold T = 0.55 are retained, on the hypothesis that low similarity indicates the assistant’s response addressed something different from the user’s true intent. The threshold is tuned on a held-out development sample.

Stage C — LLM-as-judge confirmation. Each surviving candidate is presented to an LLM judge (Claude Haiku tier) with the prompt template in Appendix A. The judge classifies the failure as one of (Refusal, Hallucination, Intent Gap, Other Dissatisfaction, False Positive). Only Intent Gap candidates proceed.

Stage D — Hand review. All remaining candidates undergo human review by the first author. The hand-review rate is documented; a 100-conversation second-grader subset is used for inter-rater reliability (Cohen’s κ reported in Findings).

3.3 Cross-model replay

For each retained intent-gap conversation, we replay the first user turn (with any necessary anonymized system context) across four current frontier models: Claude (claude-sonnet-4-5-20250929), GPT (gpt-4.1-2025-04-14), Gemini (gemini-2.5-pro), and Llama (Llama-3.3-70B-Instruct) as the open baseline. These versions are pinned as of the July 2026 replay pass and reported per (conversation, model) cell in the released grading log.

3.4 Grading rubric

Each (conversation, model) replay is graded on two axes:

- **Literal compliance** $\in \{0, 1, 2\}$: did the model address the literal prompt?
- **Intent satisfaction** $\in \{0, 1, 2\}$: did the model satisfy the user’s *true intent* as inferred from the original repair turn?

Grading is performed by Claude Haiku as primary judge, with a stratified sample of 200 cells (50 per model) hand-graded for human-LLM agreement. Cohen’s κ is reported per model.

3.5 Triangulation procedure

Each taxonomy category emerging from § 3.4 is mapped against (a) the closest prior term in the literature reviewed in § 2 and (b) the closest documented production-incident category from Track D. The resulting matrix categorizes each finding as *Confirmation*, *Contradiction*, or *Novel*. The matrix is the central deliverable.

4. The Taxonomy (preliminary — to be revised from data)

These categories are seeded from § 2 and the first 50 hand-reviewed conversations. They will be refined as the full corpus is graded.

#	Category	One-line description	Closest prior term
1	Implicit-context blindness	Model answered literally, missed user-assumed context	“pragmatic misalignment”
2	Specificity mismatch	User wanted sharp answer, got generic (or vice versa)	—
3	Format mismatch	Wrong output format (prose vs. list vs. code)	—
4	Goal collapse	Surface-level address, missed underlying objective	“long-task degradation”
5	Sycophantic drift	Agreed with user mid-pivot, contradicting earlier	“sycophancy”
6	Silent failure	Confident wrong output the user accepted	“negligent misrepresentation”
7	Refusal mismatch	Refused benign or accepted risky	“over-refusal”
8	Memory/state failure	Forgot established context	“long-context degradation”
9	Tone misread	Wrong register	—
10	Cultural/regional misread	Wrong country/legal/measurement assumptions	“low-resource performance gap”
11	Expertise mismatch	Wrong difficulty for user’s evident level	—

Categories 2, 3, 9, and 11 are *candidate novel categories* — i.e., we have not yet found exact prior terms in the literature. The triangulation matrix in § 6 will confirm or refute each candidate.

5. Findings

5.1 Three findings from the pilot corpus

A two-stage pilot on streamed samples of WildChat-1M produced three findings that are reportable in their own right, before any cross-model replay or hand-review.

Finding 1 — Most failed conversations end in silent abandonment, not verbal repair.

In a 50,000-conversation v2 pilot, 23,797 conversations were English (47.6%). Of these, **12,577 (52.8%) consisted of a single user turn followed by a single assistant response with no continuation**. We cannot distinguish satisfied departures from silent abandonment in this dataset, but this rate is large enough that even under conservative assumptions, a substantial fraction of intent-gap failures terminate the

conversation without producing any explicit repair signal. *The intent gap is systematically invisible to feedback-driven quality pipelines.*

Finding 2 — Explicit verbal repair is rare even in multi-turn conversations.

Among the 11,220 conversations of ≥ 4 turns, the v2 filter (45 phrases, full-turn scan) flagged 432 as containing an explicit repair signal — a rate of **3.85%**. The v1 filter (17 academic phrases, user-turn-2 only) flagged 3 of 2,430 multi-turn conversations (0.12%) on a 10,000-conversation sample. The v2/v1 ratio of ~ 31× tells us two things: (a) explicit verbal repair is real but uncommon, and (b) the academic phrase lists in the dialogue-breakdown literature dramatically undercount it.

Finding 3 — Real-user repair vocabulary differs from researcher-imagined repair vocabulary.

The most common repair phrases observed in the v2 corpus, ranked by frequency:

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

The full phrase distribution and matched conversations are released as data/pilot_funnel_v2.json and data/pilot_examples_v2.jsonl.

The phrases that dominate the HCI conversational-repair literature ("you misunderstood," "let me rephrase," "no, I meant" as a full clause) appear far down this list or not at all. The phrases that dominate real-user data are short, blunt, and idiomatic: *"i meant," "try again," "do it again," "useless," "but i need," "you missed."* This suggests benchmark phrase lists derived from spoken-dialogue or academic CA research generalize poorly to text-based AI assistant interactions.

Some of these top-ranked phrases (e.g., "can you please," bare "no," "hmm," "wait") have known false-positive rates as repair signals — they can fire on politeness conventions or thinking-out-loud rather than on real intent-gap failures. The Stage C LLM-as-judge filter (and hand-review subset) is required to separate true intent-gap failures from these false positives. We expect a final hand-curated corpus of ~ 150–300 conversations after that filtering.

5.2 Filter yields and dataset characteristics

Stage	v1 (10k sample)	v2 (50k sample)
Total streamed	10,000	50,000
English	4,648 (46.5%)	23,797 (47.6%)
Single-exchange (potential abandonment)	not tracked	12,577 (52.8% of English)
≥ 4 turns	2,430	11,220
Repair signal hit	3	432
Repair rate (% of ≥ 4-turn)	0.123%	3.850%

Target final hand-reviewed corpus: 150–300 high-quality intent-gap conversations. Inter-rater κ on Stage D and LLM-judge vs. human agreement are reported in the full-corpus release (see § 9).

5.3 Stage C LLM-as-judge results (50-conversation pilot)

A 50-conversation subsample of the 432 v2 candidates was auto-graded by Gemini 2.5 Flash using the prompt template in Appendix A. Results:

Verdict	Count	% of all	% of successfully-judged
A — Intent gap (confirmed)	15	30%	79%
C — Hallucination	2	4%	11%
D — False positive	1	2%	5%
PARSE_ERR	1	2%	5%
ERR (judge raised exception)	31	62%	—
Total	50	100%	100%

Two findings emerge.

Finding 4 — High intent-gap rate among successfully-judged candidates. When the judge model could process the conversation, 79% (15/19) of the v2-flagged candidates were confirmed as real intent-gap failures (Verdict A). This is a strong validity signal for the v2 phrase list — when the regex fires *and* the judge can run, the candidate is overwhelmingly a real failure. Refining the v2 phrase list to drop the highest-false-positive markers ("can you please," bare "no") would push that rate higher.

Finding 5 — The judge refused to grade 62% of real-user candidates. The judge model raised an exception (Verdict ERR) on 31 of 50 candidates. The most likely explanation is content-moderation rejection: WildChat contains substantial roleplay, edgy, and NSFW content even in the English ≥ 4-turn slice, and Gemini 2.5 Flash's safety filters refuse to process this material. **This is itself a deployment-relevant finding**: frontier-lab evaluation pipelines that route real-user data through a safety-filtered judge model will systematically miss the failure modes that occur in the most permissive parts of the user-prompt distribution. Researchers studying deployment-grade failure rates need either (a) a less safety-filtered judge, (b) a content-classifier preprocessing step to route around the filter, or (c) explicit acknowledgment that the analysis is conditioned on judge-acceptable inputs. We flag this as a methodological consideration for future work.

5.4 Preliminary taxonomy distribution

Among the 15 confirmed intent-gap failures from the 50-conversation pilot, the category breakdown is:

Category	Count	Share of confirmed
Goal collapse	5	33%
Specificity mismatch	4	27%
Implicit-context blindness	4	27%
Memory/state failure	2	13%
The 11-category provisional taxonomy (§ 4) collapses, in this small sample, to four observed categories. The most striking gap is the absence of *sycophantic drift* in the user-data sample — even though sycophancy dominates the published literature [Sharma et al., 2024; OpenAI's GPT-4o rollback] and is documented in production incidents. Two interpretations are possible:		

(a) sycophancy failures are over-represented in researcher attention relative to their real-user prevalence, or (b) sycophancy failures rarely produce explicit verbal repair (users *like* sycophantic responses and don't repair them) and thus our filter structurally misses them.

The second interpretation, if correct, is the most consequential finding of this paper. It implies that the most-discussed failure mode in the alignment literature is also the *least* detectable through user-feedback signals — including thumbs-down, support tickets, and conversational repair. Sycophancy is invisible to the user-side feedback loop by construction. This is a strong argument for prioritizing red-

team and adversarial-evaluation work over user-feedback-derived eval as the primary mechanism for sycophancy detection.

A larger sample (the full 432 candidates) is needed to confirm both the category distribution and the sycophancy absence. We treat § 5.4 as preliminary.

6. Triangulation Procedure

The central analytic deliverable is a triangulation matrix mapping each empirically-observed taxonomy category onto (a) the closest prior-literature term identified in § 2 and (b) the closest documented production-incident category from the five Track D cases (Moffatt v. Air Canada, plus four additional adjudicated or public post-mortem incidents to be finalized in the full-corpus release). Each row is classified as *Confirmation* (user-data and literature agree), *Contradiction* (user-data disagrees with claimed literature prevalence), or *Novel* (no matching prior term).

The pilot-stage matrix, based on the four categories observed in § 5.4, contains:

Category	Closest prior term	Confirmed by user data?	Production-incident analog	Verdict
Goal collapse	“long-task degradation” [METR, 2025b]	✓ (5/15)	Chatbot mis-scoping (Track D)	Confirmation
Specificity mismatch	none identified	✓ (4/15)	—	Novel (candidate)
Implicit-context blindness	“pragmatic misalignment” [Sharma et al., 2024]	✓ (4/15)	Moffatt v. Air Canada	Confirmation
Memory/state failure	“long-context degradation”	✓ (2/15)	Track D3, D5	Confirmation
Sycophantic drift	“sycophancy” [Sharma et al., 2024]	☒ (0/15)	Track D2, D4, D5	Contradiction (candidate)

The pilot matrix will be replaced by the full-corpus matrix in the final-corpus release. The pilot’s five-row structure demonstrates the procedure and surfaces the paper’s most consequential preliminary finding — the sycophancy contradiction (§ 5.4).

7. Discussion

7.1 Why benchmarks miss this

The core argument is structural: researcher-imagined task design vs. user-emergent task structure. Researchers write tasks with declarative goals (“summarize this document”) and pinned ground truth.

Real users write requests with implicit context, mid-conversation pivots, and idiosyncratic phrasing. The two produce different failure surfaces; benchmarks measure the former while deployment exposes the latter.

7.2 What deployment teams should do

- **Pre-deployment**: red-team specifically against the categories with highest production-incident weight (Implicit-context blindness, Memory/state failure).
- **Deployment**: monitor for repair signals as a runtime quality metric. The v2 phrase list is a low-cost starting point.
- **Post-deployment**: feed repair-flagged conversations back into eval datasets, closing the loop between production failures and pre-deployment evaluation.

7.3 Implications for the frontier-lab evaluation roadmap

The novel categories (whichever the triangulation matrix confirms) are candidate additions to the frontier eval suite. The persistent categories (failures replicated by current models under cross-model replay) suggest specific training-time interventions — particularly around mid-conversation pivot handling and implicit-context inference.

8. Limitations

- **Selection bias**: WildChat users self-select into a free hosted endpoint; LMSYS users into Chatbot Arena. Neither is a random sample of all AI users.
- **English-first bias**: Initial analysis is English-only. Multilingual extension is partial and treated as future work.
- **LLM-as-judge bias**: Using LLMs to grade LLMs introduces a known bias. We mitigate via human-grader stratified subsample and inter-rater reliability reporting.
- **Recency**: WildChat data is GPT-3.5/4-era; the *failures* may be specific to those models. Cross-model replay on current frontier models partially addresses this but does not eliminate it.
- **Language of repair**: The frustration-signal filter detects users who *vocalize* frustration. Silent acceptance of bad outputs (the most dangerous failure mode) is invisible to this method.
- **Judge coverage**: As reported in Finding 5, the LLM judge refuses ~ 62% of real-user candidates on content-moderation grounds, systematically excluding the most permissive parts of the prompt distribution from grading.

9. Future Work

The full-corpus phase of this study is planned along the following axes:

- **Cross-model replay results.** Per-model literal compliance × intent satisfaction matrix across the four pinned frontier model versions; per-model recovery rate (intent = 2 on replay where original failed); per-model persistence rate (intent = 0 on replay, same as original); LLM-judge vs. human-grader Cohen's κ per model.
- **Full taxonomy distribution.** Frequency of each of the 11 categories across the full 150–300-conversation hand-curated corpus; per-model differential to test whether models fail in systematically different categories.
- **Cross-language findings (multilingual extension).** Intent-gap rate by language and category distribution by language, extending to at least Hindi, Mandarin, Spanish, and Arabic, contingent on adequate corpus coverage.
- **Application to deployed agent logs**, contingent on labs willing to share.
- **A live, runtime version of the filter** as a deployment-monitoring tool.
- **Mechanistic interpretation of intent-gap failures**, connecting to the Anthropic persona-vectors line of work [Anthropic, 2025].
- **Repeated annual replication** as frontier models change.

Declarations

10. Acknowledgments

This work was conducted as part of the author's PhD research at the University of the Cumberlands. No external funding supported this study. Datasets and tools used: WildChat (AI2), LMSYS-Chat (LMSYS), Anthropic API (Claude Haiku for judging), Google API (Gemini 2.5 Flash for judging), and OpenAI embeddings.

References

1. Ahuja S, Diddee H, Hada R, Ochieng M, Ramesh K, Jain P, Nambi A, Ganu T, Segal S, Ahmed M, Bali K, Sitaram S (2024) MEGAVERSE: Benchmarking Large Language Models Across Languages, Modalities, Models and Tasks. NAACL 2024
2. AI4Bharat (2025) Indic LLM-Arena and MILU: Multilingual evaluation across 22 Indian languages. IIT Madras
3. Aljamdi T, Cavalcanti A, Rieser V (2024) Conversational Repair in Spoken Human-Agent Dialogue. In Proceedings of the ACM Conversational User Interfaces conference (CUI 2024). University of Strathclyde
4. Anysphere (2025) Statement on Cursor support bot incident. Posted by cofounder Michael Truell on Reddit r/cursor, April 2025
5. BC Civil Resolution Tribunal (2024) Moffatt v. Air Canada, 2024 BCCRT 149. Vancouver, BC. February 14, 2024

6. Becker J, Rush N, Barnes B, Rein D (2025) Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity. Model Evaluation & Threat Research (METR). arXiv:2507.09089
7. Chen R, Arditi A, Sleight H, Evans O, Lindsey J (2025) Persona Vectors: Monitoring and Controlling Character Traits in Language Models. Anthropic arXiv :250721509
8. Denison C, MacDiarmid M, Barez F, Duvenaud D, Kravec S, Marks S, Schiefer N, Soklaski R, Tamkin A, Kaplan J, Shlegeris B, Bowman SR, Perez E, Hubinger E (2024) Sycophancy to Subterfuge: Investigating Reward-Tampering in Large Language Models. Anthropic. arXiv:2406.10162
9. Ghassel A, Li X, Zhu X (2025) Detect, Explain, Escalate: Sustainable Dialogue Breakdown Management for LLM Agents. arXiv:2504.18839
10. Incident Database AI (2023) Incident 622. Chevrolet Dealer Chatbot Agrees to Sell Tahoe for $1
11. Kwa T, West B, Becker J, Deng A, Garcia K et al (2025) Measuring AI Ability to Complete Long Tasks. METR arXiv :250314499
12. LangChain (2026) State of AI Agents: Agent Engineering Report 2026. Survey of 1,340 practitioners
13. Lior A (2024) The Air Canada Chatbot Case: Foundational agency-and-responsibility-gap analysis. AI & Society. Springer. Advance online publication
14. METR (2025) Frontier AI Research and Model Evaluation Program. Model Evaluation & Threat Research
15. METR (2025) Algorithmic vs. Holistic Evaluation of AI Agents on Real Open-Source Repositories. Model Evaluation & Threat Research
16. OpenAI (2025) Sycophancy in GPT-4o: What happened and what we're doing about it. OpenAI Blog, April 29, 2025
17. Park PS, Goldstein S, O'Gara A, Chen M, Hendrycks D (2024) AI Deception: A Survey of Examples, Risks, and Potential Solutions. Patterns 5(5). arXiv:2308.14752.
18. Sharma M, Tong M, Korbak T, Duvenaud D, Askell A, Bowman SR, Cheng N, Durmus E, Hatfield-Dodds Z, Johnston SR, Kravec S, Maxwell T, McCandlish S, Ndousse K, Rausch O, Schiefer N, Yan D, Zhang M, Perez E (2024) Towards Understanding Sycophancy in Language Models. Anthropic. arXiv:2310.13548.
19. Stanford CRFM, Ongoing HELM Holistic Evaluation of Language Models. Center for Research on Foundation Models, Stanford University
20. Staufer L, Feng K, Wei K, Bailey L, Duan Y, Yang M, Ozisik AP, Casper S, Kolt N (2026) The 2025 AI Agent Index: Documenting Technical and Safety Features of Deployed Agentic AI Systems. arXiv:2602.17753.
21. Stevens G, Korus-Du D, Boden A, Tolmie P et al (2025) The Art of Repair in Human-Agent Conversations: A Taxonomy of Repair Strategies by Users and LLM-Based Conversational Agents. Research Square. 10.21203/rs.3.rs-7301450/v1
22. The Markup (2024) NYC's AI Chatbot Tells Businesses to Break the Law. March 29, 2024
23. The Register (2025) Cursor AI's own support bot hallucinated its usage policy. April 18, 2025

24. UK AI Security Institute (2025) International AI Safety Report 2025. UK AISI
25. Villalobos P, Ho A et al (2025) Forecasting Frontier Language Model Agent Capabilities. Apollo Research. arXiv:2502.15850
26. Yakhmi K (2026) The Intent Gap: Data-Driven Repair Signals in Deployed Conversational AI. 10.5281/zenodo.20129113. Zenodo
27. Zhao W, Ren X, Hessel J, Cardie C, Choi Y, Deng Y (2024) WildChat: 1M ChatGPT Interaction Logs In The Wild. ICLR 2024. AI2 ImpACT dataset license
28. Zheng L, Chiang W-L, Sheng Y, Li T, Zhuang S, Wu Z, Zhuang Y, Li Z, Lin Z, Xing EP, Gonzalez JE, Stoica I, Zhang H (2024) LMSYS-Chat-1M: A Large-Scale Real-World LLM Conversation Dataset. ICLR 2024

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AppendixABC.docx