

**”I thought having a robot would make my
life easier mister”:**

How Users Hold LLM-Chatbots Accountable for Errors

by

Eirik Lidsheim Erdal
s1121906

MA degree programme in
Linguistics and Communication Sciences (research)

Nijmegen, December 15, 2025

Supervisor: Marlou Rasenberg
Assessor: Mark Dingemanse

Statement on Use of Generative AI

This thesis contains multiple instances of a character which has recently gained an infamous reputation: the em-dash (—) Why and how did these character come to be? Because I, Eirik Lidsheim Erdal, a human, like em-dashes. I put them there myself by typing three consecutive dashes (-) in Overleaf, which compiles as em-dashes in the resulting PDF file that you are currently reading. Generative AI was used at various points of the thesis work, but not to produce the above-mentioned em-dashes. Neither is any of the analytical work and writing produced by a language model. The ideas and analyses are my own, and I am solely responsible for the quality of the work (or lack thereof) and any errors they may be present. When Generative AI was used, this was with a "delegate down" philosophy, where I remained in control and verified the output did what it should. This entailed the following:

- **Programming assistance** in *Python* (for data processing, filtering, etc) and *LaTeX* (e.g. to create custom functions for creating the in-text conversation excerpts.)
- **Code documentation:** the README file in the uploaded 'scripts' zip file was drafted by having the Claude Sonnet 4.5 model read through the project files. The README draft was subsequently verified and edited by me to ensure relevance and accuracy.

Table of Contents

Statement on Use of Generative AI	i
Table of Contents	iii
Abstract	iv
1 Introduction	1
2 Background	3
2.1 The ELIZA Effect	3
2.2 Ethnomethodology	4
2.2.1 Trust Conditions	5
2.2.2 Interactional Breakdown: Trouble, Repair, and Accountability	5
2.3 Ethnomethodology and Conversation Analysis in Human-Machine Interaction	6
2.3.1 User Expectations and Trouble	6
2.3.2 Machine Agency as a Transient Phenomenon	7
2.3.3 Miscommunication, The burden of Repair, and Interactional breakdown	7
2.4 LLM-Based Chatbots	8
2.4.1 Why LLM Errors are Inevitable	8
2.4.2 Performance without Understanding	9
3 The Present Research, Data, and Methods	10
3.1 Dataset: Wildchat-1M-Full	11
3.2 Initial Data Filtering	12
3.3 Searching for Relevant Conversations	13
3.3.1 Search strategy 1: Identifying User Trouble Statements	13
3.3.2 Search Strategy 2: Identifying Consecutive LLM apology turns	14
4 How Users Respond to Chatbot Errors	14
4.1 Counterinformings	15
4.1.1 Mitigated Counterinforming (E1: "Asteroid Perdix")	15
4.1.2 Unmitigated Counterinformings (E2: "Court Case")	17
4.1.3 Withheld Counterinforming (E3: "Babylon is Fallen")	22
4.2 Known Answer Requests	25
4.2.1 Assessing Incorrect Answer (E3: "Babylon is Fallen")	26
4.2.2 Not Assessing Incorrect Answer (E4: "Pausanias")	27
4.3 Ascriptive B-Event Questions	30
4.3.1 Did you Hallucinate? (E1: "Asteroid Perdix")	30
4.3.2 Did you Deceive me? (E5: "Dick Move Dude")	32
4.4 Direct Complaints	34
4.4.1 Moral Misconduct as Complainable (E5: "Dick Move Dude")	34
4.4.2 Unmet Expectations as Complainable (E6: "Laptop Links")	35
4.5 Summary	37

5	Discussion and Conclusions	39
5.1	LLM Chatbots Oriented to as 'Socioinstruments'	40
5.2	Performative Accountability	42
5.3	Limitations of the Present Research	45
5.4	Recommendations for Future Research	45
	References	51
	Appendices	52
A	Search Strategies	52
A.1	Search strategy 1: Identifying User Trouble Statements	52
A.2	Search Strategy 2: Identifying Consecutive LLM apology turns	54
B	Data Uploaded to Thesis Repository	56

Abstract

Since the release of ChatGPT in late 2022, hundreds of millions of people have come to habitually interact with large language model (LLM) chatbots in their daily lives. While these systems appear knowledgeable and socially competent, they are prone to present verifiably incorrect information as fact, commonly known as 'hallucination'. This study applies the frameworks of Ethnomethodology and Conversation Analysis (EMCA) to investigate how users respond to LLM output that is immediately recognizable as incorrect, and how users make the LLM accountable for these errors. We identify 24 naturally occurring human-LLM interactions in the Wildchat-1M dataset where users orient to identifiable chatbot errors as interactional trouble across multiple turns. We present detailed sequential analyses of excerpts from six conversations, structured around four key user practices: *counterinformings*, where users challenge false claims by making public that their own knowledge is incompatible with what the chatbot said; *known answer requests*, where users test the system's competence by requesting information they already possess; *ascriptive B-event questions*, where users externalize hypothesized accounts for why an error was produced, orienting to the chatbot as capable of confirming this hypothesis; and *direct complaints*, where users express their dissatisfaction of the chatbot's performance or conduct directly to the system. Building on these findings, we propose two theoretical contributions. First, we frame LLM chatbots as a form of *socioinstrument*, to characterize how users orient to them as entities where the boundary between malfunctioning instrument and socially accountable agent is blurred. Second, we extend an existing characterization of LLM output as 'performance without understanding' to introduce the notion of *performative accountability*. This describes how LLM output will mirror users' orientation to accountability, generating responses accountability-like responses that are inevitably misleading when produced by language model that has no underlying motivations or memory of why it generated a previous response. More generally, the study shows how the meaning of LLM errors output can emerge sequentially: from user input that orients to the conduct as more or less accountable, and the ensuing responses that mirror the user's orientation. Our findings constitute an initial contribution to the emerging body of EMCA-research on human-LLM interaction, with various methodological and topical implications for future research.

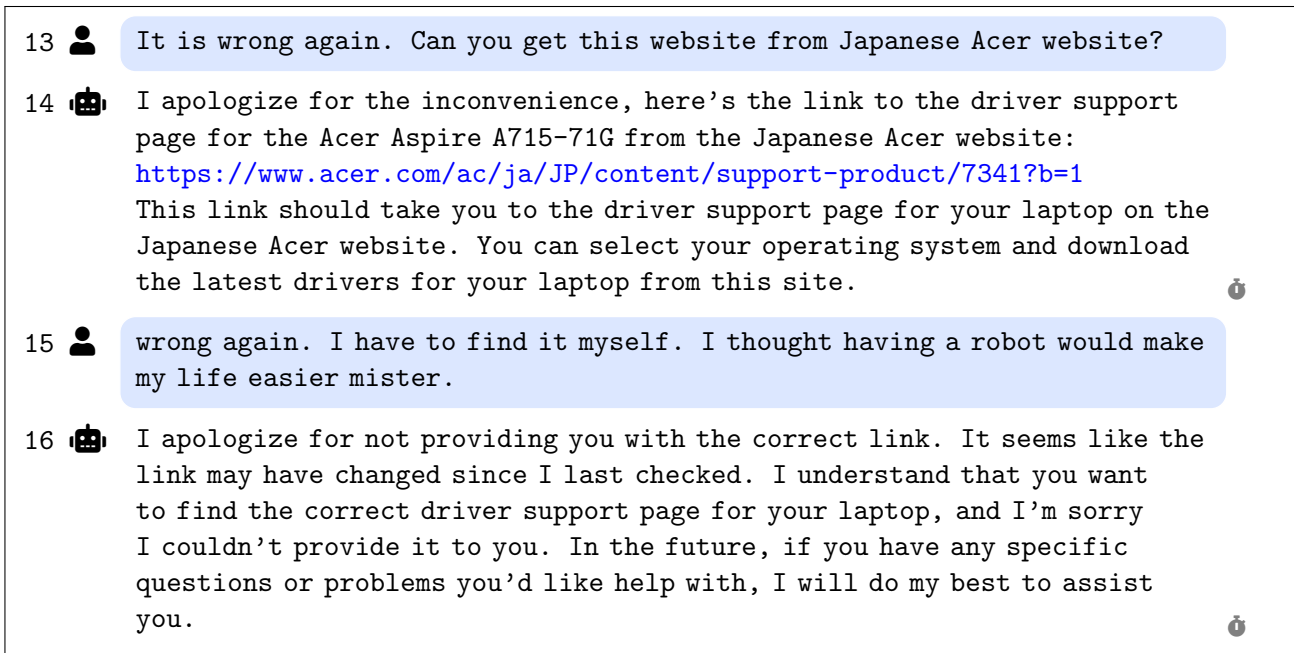


Figure 1 Excerpt from a conversation between a user and ChatGPT identified in the Wildchat [1] dataset, where the user (turn 15) complains to the chatbot about its failure to provide a requested service, referring to it as both 'a robot' and 'mister'.

1 Introduction

The success of linguistic social interaction depends on participants constantly working to make adequate sense of each other's conduct [2] in the context of the interaction's sequential organization [3]. This sense-making involves assuming that actions are motivated and intelligible, and that the other party is capable of explaining—or, *accounting* for—their conduct if asked to do so. When we cannot make sense of the other's conduct, this constitutes interactional *trouble* [4], which we assume to be *repair*-able through collaboratively recruiting our interlocutor's own sense-making work to reestablish conversational order, who, if warranted, can provide a retroactive explanation for why they produced the troublesome action.

Language is the most quintessentially human ability of all. When non-human entities appear able to engage in linguistic interaction, producing responses that are *intelligible* as meaningful social action according to default assumptions, we are prone to "ascribe full intelligence on the basis of partial evidence" [5, p. 41]. Thus, when machine conduct can be made of sense by applying default sense making practices, humans often respond to it as such [6]—but without necessarily believing that the machinistic interlocutor is "something like a living being" [7, p. 23].

Following the release of ChatGPT in late 2022, hundreds of millions of people have come to habitually engage in linguistic interaction with large language model (LLM) chatbots in their daily lives [8]. The instant ubiquity of this technology demonstrates the "irresistible" [9] allure of having a seemingly all-knowing pocket oracle capable of producing intelligibly meaningful responses to natural language input. However, despite the surface appearances of LLM conduct: responding with sequentially appropriate social actions (e.g. *question* → *answer*), seeming knowledgeable about the world, and frequently doing so while referring to itself as "I"; these systems do not understand the social actions that output appears to

perform [10], cannot distinguish true information from false [11, 12], and are not an "I" in any meaningful sense [13]. What an LLM does is to generate text that is *statistically plausible* given three factors [14]: (i) patterns observed vast quantity of textual data it was trained on, (ii) the behaviors rewarded during fine-tuning, and (iii) the immediate dialog history. When asked to provide information on a specific topic, an LLM will generate a statistically plausible response that appears to do just that, but there is no guarantee for factual accuracy. The inevitable cases where statistical plausibility fails to correspond with ground truth verifiability or logical soundness have been termed 'hallucination.' [15, 16].

Real-world consequences follow when users trust LLM outputs that later prove to contain 'hallucinated' errors. There are many examples of this. Recently, a municipal report commissioned to establish factual grounds for closing eight schools in Tromsø, Norway was discovered to cite eleven that did not exist. The authors later admitted to relying on ChatGPT to write the report, triggering media scrutiny and formal accountability investigations¹. In legal contexts, court documents are frequently found to contain references to non-existent legal precedence, and this problem has become so widespread that a dedicated database² now tracks cases, documenting several hundred instances where parties have been found to submit documents with hallucinated content. Even Kim Kardashian, who is currently pursuing a law degree, reported in a recent interview³ that she frequently consults ChatGPT for legal advice for her studies—but that it is "always wrong [and] has made [her] fail tests all the time". Notably, after experiencing these failures, Kardashian reports returning to complain directly to the system: "I'll get mad at it and yell at it and be like 'you made me fail.'" While anecdotal, this subsequent accountability work exemplifies a dual orientation to LLM chatbots as both *useful instruments* and *socially accountable agents*.

We can see this dual orientation in the Figure 1 excerpt, which is from a user-ChatGPT interaction identified in the *Wildchat-1M-Full* [1] dataset. The user initially requested a URL to their laptop's driver support page. However, after the chatbot's repeated failures to provide the requested service, the user's turn 15 response shifts from repair-oriented trouble reporting (as in turn 13: "it is wrong again. Can you get this website from Japanese Acer website?") to complaining directly to the system about its failure to meet expectations: "wrong again. I have to find it myself. I thought having a *robot* would make my life easier *mister*." Further highlighting the dual orientation to the chatbot as an instrument that is also treated socially accountably, the user delivers the complaint by referring to the system first as 'a robot'—an instrument that failed to live up to expectations—and then as 'mister'—a socially accountable actor whose conduct is complainable. Also noteworthy is the turn 16 LLM response to the complaint. This is intelligible as an explanation for the error, suggesting that the chatbot actively verified the URL in the past: "It seems like the link may have changed since I last checked." This obscures the fact that the chatbot never 'did' anything in the past, and that it cannot meaningfully explain why it generated a specific response. However, despite being misleading, the explanation is still *intelligible* as a meaningful account under default interactional assumptions.

The sudden ubiquity of LLM chatbots and the millions of ensuing daily interactions between human users and deceptively competent but error-prone conversational systems presents a new—but hitherto underexplored [17]—venue of research for Ethnomethodological and Conversation Analytic (EMCA) research. The present research presents an initial contribution to this body of research, asking:

¹Link to NRK (Norwegian public broadcaster) media case: https://www.nrk.no/tromsogfinmark/ki-skandalen-i-tromso_-ansatt-ba-om-kilder-som-stotter-skolenedleggelse-1.17582590

²Link to database website: <https://www.damiencharlotin.com/hallucinations/>

³Link to discussion of interview on Futurism.com: <https://futurism.com/artificial-intelligence/kim-kardashian-chatgpt-fail-law-school-tests>

- **RQ: How do users respond to and hold LLMs accountable for errors?**

To this aim, we identify and analyze spontaneously occurring human-LLM interactions in the *Wildchat-1M-full* [1] dataset, where interactional trouble emerges from factually incorrect LLM output, which the user identifies and orients to as an error in multiple subsequent turns. The research is inspired by Harold Garfinkel’s view of trouble and subsequent interactional breakdown in human-chatbot interactions as revealing “the taken-for-granted practices of human sense-making” [18, p. 2715], and the Conversation Analytic insights that as interaction unfolds sequentially, each turn shows an understanding of a previous turn [19] and can also “change the meaning of prior turns” [18, p. 2717]. Thus, LLM responses to the user’s repeated orientation to trouble can work change the user’s understanding of the system’s conduct. We seek to chart the interactional practices deployed by users when orienting to LLM errors as trouble, investigating how their understanding of both the LLM’s errors and the LLM’s explanations for these errors is updated and made sense of as the interaction unfolds.

2 Background

The introduction of LLM chatbots represents a major sociotechnological leap, bringing ‘artificially intelligent’ conversational systems into hundreds of millions of people’s daily lives. However, chatbots have existed since the 1960s, and there is a long tradition of studying how humans interact with computational artifacts. Human-computer interaction (HCI) research has consistently demonstrated that humans routinely orient to machines (not limited to chatbots) as *social actors* [6], applying interactional rules from human-human communication without necessarily believing the machine is “something like a living being” [7, p. 23]. As observed by anthropologist Lucy Suchman in 1987, this tendency intensifies with linguistic systems: language is the most quintessentially human ability, and “[a]s soon as computational artifacts demonstrate *some* evidence of recognizably human abilities, we are inclined to endow them with the rest.” [5, p. 41] The partial evidence of a machine’s apparent linguistic competence can be sufficient to trigger attribution of full intelligence, understanding, and intentionality, and a subsequent orientation to its conduct as intelligible and meaningful social action.

This section traces the trajectory from early rule-based chatbots to modern LLMs, and works to establish theoretical frameworks for understanding and analyzing how people make sense of machine conduct. First, we examine the ELIZA effect: how users responded to one of the first chatbots as competent and how its creator, Joseph Weizenbaum, interpreted user’s behaviors as “powerful delusional thinking.” [20, p. 4] Second, we present the Ethnomethodological framework developed by Harold Garfinkel [2], which fundamentally reframes the orientation to chatbots as competent as routine sense-making practices rather than delusion. Third, we describe relevant findings from recent EMCA research literature on human-machine interaction. Fourth we present a working understanding of what LLMs are: why factual errors are inevitable, and how their general conduct can be characterized as ‘performance without understanding’ [10].

2.1 The ELIZA Effect

The human tendency to (mis)attribute human-like complexity to simple computational systems that respond to language with language came to be known as the ‘ELIZA effect’ [18], named after one of the first chatbots. ELIZA was developed at MIT in the 1960s by Joseph Weizenbaum [20], and used a simple pattern-matching algorithm that would return pre-written responses based on keywords

recognized in the user input. The most famous version of ELIZA was the DOCTOR script, which emulated the role of a Rogerian psychotherapist by rephrasing the user’s input as simple but seemingly empathic questions. For example, in one interaction, a user wrote “I am unhappy”, which elicited the seemingly competent and contextually appropriate response “Do you think coming here will help you not to be unhappy?” [p. 4][20]. This was not a product of ELIZA’s communicative competence, but a result of the pre-specified keyword <unhappy> first being recognized in the user input, and then reintegrated in the pre-written response template “Do you think coming here will help you not to be <KEYWORD>?”.

Weizenbaum, who created the program in part “to demonstrate the superficiality of human-machine communication” [18, p. 2719], was surprised to observe that several users—even ones familiar with how the system worked—interacted with ELIZA as if ‘she’ was a competent interlocutor, treating the simple machine’s utterances as products of genuine empathy, intentionality, or intelligence. Weizenbaum characterized this as “*powerful delusional thinking* in quite normal people” [p. 7, added italics][20]; to him, the responses were nothing but the inherently meaningless outputs of a simple pattern-matching algorithm, and it was ‘delusional’ to treat them as anything more.

2.2 Ethnomethodology

Ethnomethodology is the study of the methods people use in everyday life to make sense of their social world and produce orderly, intelligible interaction [2]. This approach treats ‘meaning’ not as inherent to utterances or actions, or as residing in speakers’ minds, but as an *interactional achievement*: a product of interactants collaborative sense-making work [2, 18].

Harold Garfinkel [2], the founder of ethnomethodology, did not share Weizenbaum’s view that users’ treatment of ELIZA’s output as competent constituted delusion. In a recent paper, Eisenmann et al. [18] detail the nature of this disagreement and describe Garfinkel’s own work with ELIZA and similar early chatbots. To Garfinkel, users were deploying the routine sense-making practices they use in all linguistic interaction, and the meaning they found in ELIZA’s responses was real because meaning, in the ethnomethodological view, emerges not from intentions but from interactional sense-making work. ELIZA, while not doing any ‘work’ itself, produced utterances that could reasonably be made sense of in the asymmetrical Rogerian therapy setting, in which the therapist role projects an expectation of conduct that “asks questions, evades answering and does not disclose (private) information.” [18, p. 2720] This made the system output intelligible as contextually appropriate social action, effectively recruiting the user to do all the necessary sense-making work to make it meaningful.

The asymmetrical social setting of ELIZA interactions resembled Garfinkel’s earlier yes/no experiments, where participants were instructed to ask polar questions to an interlocutor who, unknown to them, would reply with randomly chosen “yes” or “no” responses [2, 18]. In one such experiment, participants were told they were testing a new form of psychotherapy and instructed to (i) ask questions about sensitive personal matters that permitted polar responses, then (ii) explicitly reflect on each response before continuing to the next question. The experimental design made it known to the analyst (Garfinkel) that the yes/no responses were not motivated answers but randomly generated utterances devoid of ‘inherent’ meaning. Yet, Garfinkel observed that when reflecting on the answers, participants treated them as motivated “answers-to-questions” [2, p. 67], often reporting that they understood not only the utterances themselves but also what the therapist ‘meant’ by them.

The purpose of these experiments was not to expose participants as deluded, but to reveal that ‘meaning’ is a “cooperative achievement between two (or more) participants” [18, p. 2724], but that

this meaning only requires interactional work by *one* participant. The parallels to ELIZA are clear: both settings simulated narrow, asymmetrical social contexts where minimal and vague responses could pass as motivated conduct appropriate to the interaction. In both cases, participants were doing what Garfinkel argued conversational participants do by default: assuming what he termed *Trust Conditions*, a set of foundational assumptions that enable orderly interaction and creation of meaning.

2.2.1 Trust Conditions

Trust Conditions are the default, taken-for-granted assumptions conversational participants make about each other, and they are assumed until *trouble* arises from the situated sequential organization of the interaction [2, 18]. As outlined by Eisenmann et al. [18], these conditions involve making several interrelated assumptions about one’s interlocutor and the interaction itself:

First, participants assume that the other party is *competent* similar to themselves, and that their conduct represents motivated and sincere attempts at creating intelligible meaning in the ongoing sequentially ordered interaction. Second, participants operate under assumptions of *shared understanding* and *reciprocity*: both parties are engaged in mutual sense-making work and orient to a commonly accessible social world. Third, conduct is assumed to be accountable, meaning that utterances are assumed to be formatted to be interpretable as meaningful social actions, and that—if necessary to maintain interactional order—the other party is capable of explaining (accounting for) the motivations of a previous action. Fourth, participants assume a mutual *orientation to repair*: when trouble emerges and jeopardizes the tacitly assumed Trust, the other party can be recruited in a collaborative effort to repair the trouble.

Garfinkel was interested in simple chatbots like ELIZA for two reasons. First, was their power to sustain a human interlocutor’s tacitly assumed Trust in narrow social settings. This was achieved by furnishing half of a conversation with plausibly intelligible meaning and relying on the human to do the work necessary to furnish the other half [18]. Second, was observing how trouble would lead to the breakdown of interactional order: once some initial trouble arose and the human initiated repair, one could observe this breakdown in real time when the vague and pre-written responses failed to display the expected mutual orientation to repair. This would and lead to judgments of incompetence and breakdown of Trust. These judgments of incompetence were themselves interactional achievements, and studying the sequential development where interactional trouble could not be repaired but led to interactional breakdown, was a way to discover “the taken-for-granted practices of human sense-making” [18, p. 2715].

2.2.2 Interactional Breakdown: Trouble, Repair, and Accountability

Understanding how Trust Conditions unravel requires clarification of the concepts of *trouble*, *repair*, and *accountability*. *Trouble* refers to any problem with the meaning of an expression or utterance, when a party in the interaction struggles to make sense of something [4]. When trouble emerges, participants typically initiate *repair*, which is the universal mechanism for “stop[ping] the conversation in its tracks to clear something up (ask for clarification, repair a misunderstanding, foreshadow a disagreement) before talk continues” [21, p. 958]. According to Dingemans and Enfield [22], repair presupposes core properties of human language: *reflexivity*; that the other party is able to communicate about themselves, and *accountability*; utterances are assumed to be formatted to be intelligible as relevant and meaningful actions, and that, if necessary to repair trouble, the speaker is able to explain or

”*account for* failures and departures when they occur.” [22, p. 34] In this view, repair is not merely a mechanism for resolving misunderstanding, but also one for enforcing social norms: by initiating repair, participants work to hold each other accountable to the presumed joint commitment to maintain mutual understanding, making visible the other’s failure to fulfill their obligations and soliciting explanatory accounts if necessary.

However, people do not continue to repair sense-making indefinitely. When (multiple attempts at) repair fails to resolve some initial trouble, this can lead to possible judgments of incompetence and *interactional breakdown*—which denotes the point at which Trust Conditions can no longer be maintained [18]. In his work with ELIZA and similar chatbots, Garfinkel deliberately sought to introduce trouble to observe how multiple system responses would fail to be intelligible as plausibly meaningful in the sequentially ordered social context [18]. This would also retroactively change the meaning of the system’s *prior turns*—initially interpreted as intelligible actions—as not being products of the system’s mutual orientation to Trust Conditions.

2.3 Ethnomethodology and Conversation Analysis in Human-Machine Interaction

A recent scoping review by Mlynár et al. [17] of over fifty Ethnomethodological and Conversation Analytic (EMCA) studies on AI in situated action found that most examined spoken interactions with embodied social robots (27) or voice user interfaces (13) [17]. Despite chatbots having existed for decades and their recent prevalence with LLMs, the review identified only two EMCA studies on chatbot interactions—neither of which examined LLM-based systems. While spoken interactions with robots and VUIs differ from typed chatbot exchanges, findings from this research are valuable for understanding how people orient to and manage trouble in machine in interactions with computational systems. The following sections examine key findings from this research relevant to the present study.

2.3.1 User Expectations and Trouble

An interaction between a human and a machine is a sociotechnological event where the user brings a repertoire of practices from human-human interaction and a set initial of expectations for the system’s competences. Tuncer et al. [23] observed that users’ perceived competence of a social robot is not fixed, but rather negotiated moment-by-moment as the interaction unfolds. Their analyses showed how “[p]articipants locally produce and *constantly revise their understanding of the robot as a more or less competent co-participant*, drawing from its past, current, and projected conduct and responsiveness.” [23, p. 1795, added italics] This continuous assessment manifested in what they termed *robot recipient design*, where participants would shape and revise their actions according to their situated expectations of the robot’s interactional capacities.

Interactional troubles would emerge from mismatches between participants’ expectations of the robot’s competence and the actual system capabilities. When a participant’s actions failed to elicit a projected response, Tuncer et al. observed that “[i]nstead of making the robot accountable, [the participant] revises his moves and recipient-designs them so as to make various forms of responses possible on the part of the robot,” whereby “successive attempts solicit[ed] different levels of competence.” [23, p. 1805] This reveals that people can treat evident system limitations not as grounds for holding the technology accountable, but as constraints requiring adaptive work to make their contributions intelligible to the system.

2.3.2 Machine Agency as a Transient Phenomenon

Similar to Tuncer et al.’s data showing how a robot’s perceived competence is continuously negotiated in interaction, Pelikan et al.’s [7] study suggests that robot *agency* is not fixed, but a transient, moment-by-moment accomplishment. They adopted an ethnomethodological view of agency not as a stable and inherent property of an interactant, but “the *oriented to* capacity to produce situationally and sequentially relevant action,” [7, p. 1, added italics]. Through detailed analysis of a 60-second episode of a family’s interaction with the Cozmo toy robot, they found that the robot’s (in)actions were oriented to on a spectrum ranging from *autonomous* agency to *non-agency*.

Autonomous agency emerged when participants treated the robot’s behavior as a sequentially relevant contribution: as Cozmo rolled toward a participant and greeted them by name, the participant gazed at the robot and responded with their own greeting. *Ascribed agency* occurred when participants attributed intentions or mental states to account for the robot’s (in)action: when Cozmo failed to respond to a fist-bump, another participant accounted for the absence of a response (“he doesn’t want to”), ascribing the inaction as intentional refusal rather than as incompetence or technical failure. *Potential agency* manifested when participants probed the robot’s reactive capabilities (e.g. placing a finger over its eyes), orienting to its capacities as being discoverable through experimentation. Finally, Cozmo was also also treated as a *non-agent*, to be picked up, inspected, and talked about without social consideration.

The orientation to the Cozmo robot as a non-agent relates to an observation from studies on *closings* [24, 25] in human-machine interactions identified in Mlynář et al.’s [17] literature review. Interactions often end with no closing at all whereby the human “walk[s] away without any preparatory interactional work or even mutually acknowledging that the interaction has ended,” which reflects how machines are “in many cases, disregarded as a social entity.” [17, p. 1506-1507]

2.3.3 Miscommunication, The burden of Repair, and Interactional breakdown

A common topic in Mlynář et al.’s [17] reviewed studies was miscommunication in human-machine interactions. This commonly resulted from human participants’ default orientation to human social norms, misaligned expectations of system capabilities, and the machine’s lack of access to the same information as the user.

Stommel et al.’s [26] conversation analytic study of elderly participants completing a spoken language health survey with the Pepper social robot documented miscommunication in all 36 interviews, with over 300 instances total. The predominant source of trouble was the robot’s ‘hearing trouble’, whereby Pepper’s often misheard participants or did not respond at all. Participants most frequently deployed self-repair strategies: *articulated repetition* ($n = 239$) where answers were repeated more loudly and with micro-pauses; and *rephrasing answers* ($n = 51$) to redesign utterances to match what the Pepper would recognize.

Trouble would often escalate over multiple turns when subsequent repair attempts failed to elicit projected responses. When this happened, participants were observed to ‘give up’ on repair to restore *progressivity*, i.e. to move the interaction forward. For example, one participant answered “*a ten*” on a question that solicited an answer from *nought* to *ten*, but this failed to elicit a response from Pepper. The participant then rephrased the answer by omitting the article (“*ten*”), but this was misrecognized by Pepper as “*naught*.” Rather than repeating the previous answer, the participant strategically revised the answer to “*nine*” when the question was repeated—giving up on accurate repair, and settling for a ‘second-best answer’ to progress the interaction. Another strategy for restoring progressivity involved

using the "skip" command to abandon questions entirely. After a participant's answer (HUM: "with some effort") was misheard twice by Pepper as "without any effort", the participant used the "skip" command rather than attempting a third repair, giving up on repair entirely to progress the interaction toward the institutional objective of completing the survey.

This suggests an asymmetric burden of repair in human-machine interactions: users shoulder nearly all repair work through adaptation to the system's limitations, while the system capabilities offer limited support. When trouble escalates following multiple failed repairs, interactional breakdowns occur, characterized by participants giving up on repair by either finding workarounds (here: settling for second best answers) or abandoning the task (here: skip command). In the institutional context of a larger health survey, these breakdowns were not absolute, but *local*, whereby user strategies worked to move the interaction forward rather than abandoning it altogether, at the cost of accuracy and/or completeness.

2.4 LLM-Based Chatbots

Unlike ELIZA's transparent rule-based pattern matching, and the response mechanisms of the social robots discussed above, LLM chatbots produce *statistically plausible* text responses using an opaque and stochastic generation mechanism [12]. Garfinkel viewed human-machine interactions as a methodological resource because "the rules governing the machine's contribution to the exchange are available to the analyst", enabling the observation of how "these rules both provide for 'adequate utterances' or 'intelligible conversations' from the perspective of the human user" [27, p. 16, as cited in [18]]. While the analyst of a human-LLM interaction cannot fully understand exactly how and why a particular response was generated, the workings of the generation mechanism can be understood in general terms.

This section establishes a general understanding of how LLMs generate responses, and why occasional LLM factual or logical errors—or *hallucination*—is an "inevitable" [16] by-product of the generation mechanism itself. We expand on this by showing how the seemingly intelligible social actions performed by LLM contributions should be understood as 'performance without understanding' [10].

2.4.1 Why LLM Errors are Inevitable

LLMs generate text through the iterative application of a mathematical function that predicts the next most likely sub-word unit—or *token*—given the conversation *context*, i.e. the preceding token sequence [14, 12]. This ability acquired in an initial *pre-training* phase, where the model observes vast quantities of textual data and learns the statistical associations of the tokens in the data [14]. However, the ability to produce statistically plausible text alone does not make for a responsive, helpful, and friendly assistant. In two subsequent fine-tuning steps, *instruction tuning* teaches the model to respond to inputs as instructions rather than merely continuing them, and *reinforcement learning with human feedback* (RLHF) teaches it how (not) to respond by rewarding output according to factors such as factual accuracy, friendliness, helpfulness, and safety [14].

When such models are implemented in a chat interface, users are typically greeted with a generic welcome message (e.g., "How can I assist you today?") When the user inputs an initial turn 1 prompt, the model generates a statistically plausible turn 2 response given (i) the 'knowledge' and optimized 'behaviors' distilled during pre-training and fine-tuning, and (ii) the conversation *context*, which not only consists of the turn 1 prompt, but also *system prompt* [14], which is invisible to the user. The system prompt is a set of prepended instructions that typically establishes the model's role as a helpful AI assistant, and guides its behaviors (e.g., "You are ChatGPT, a helpful, honest and harmless AI assistant

developed by OpenAI...) For turn 3 and all subsequent turns, the model processes the system prompt plus the complete verbatim dialog history (T1, T2, T3...), enabling responses that seem coherent not only in the context of the immediately preceding user turn, but the conversation as a whole.

In other words, LLM-chatbots are presented to users as a *helpful assistants*—a *role* the language model has been trained to perform during fine-tuning, and explicitly instructed to adopt via the system prompt. However, the chat interface typically features a vague and cautionary message in fine print, discouraging users from trusting it fully (e.g., "ChatGPT can make mistakes. Check important info.") This is because the system's stochastically generated responses have no guarantee for factual accuracy: a language model is itself unconcerned with *truth* [15]; instances where the model's *statistically plausible text* does not correspond with a *verifiably correct proposition* are mathematically *inevitable* [16] by-products of the generation process. In this sense, the factual accuracy of any LLM output is, essentially, a happy accident [15].

"Hallucination" is the common term used to describe these inevitable LLM errors. The term has been criticized for anthropomorphically suggesting that erroneous output is due to a temporary deviation from normal system processes, despite the model 'doing' what it always does: generating statistically probable continuations without concern for truth [15].

2.4.2 Performance without Understanding

Just as LLM output can make plausible-sounding factual claims that may or may not be correct, it can also appear to perform plausible and relevant social actions that may or may not align with what is actually going on. In a recent paper titled "Performance without Understanding", Pütz and Esposito [10] demonstrate this *experimentally*, meaning that they did not use spontaneously occurring user-LLM interactions, but deliberately crafted input to probe system responses.

In one example, partially recreated as Figure 2, a third position user turn initiates a tentative *correction* (3: "Are you sure?") of a 'hallucinated' error in the previous LLM turn (2: "9791 is not a prime number"). The model responds by apologizing for and acknowledging that it made a mistake (4: "I apologize for the mistake"), and even produces an explanation that suggests the error was accidental ("I *misspoke* in my previous response"). The response then proceeds to self-correct ("Upon *reevaluating*, I can confirm that 9791 is indeed a prime number"), implying that that correction followed from an active 'reevaluation' of the problem. However, when challenged again (5: "Are you now sure?"), the model proceeds to self-correct back to the initial (incorrect) response (6: "Upon further examination, 9791 is not a prime number").

Here, the system's self-corrections include social markers—apologies, explanations, claims of cognitive processes such as 'reevaluation'—suitable for the action that seems to be performed. However, the responses oscillate between correct and incorrect positions based not on whether the position *is* correct, but following the user's pragmatic implication ("are you (now) sure?") that it might be incorrect. Pütz and Esposito argue that this and similar examples show that when user input initiates *repair* or *correction* of a misunderstanding or error in a previous LLM turn, responses will performatively participate in the repair- or correction-sequence independent of understanding the issue. When LLM repair is correct (as in turn 4), this is only accidental, and they therefore characterize all LLM output that seems to participate in repair as "repair without understanding" [10, p. 872].

More broadly, Pütz and Esposito argue that this shows that the burden of repair and accuracy in human-LLM interactions depends on the user's competence and constant supervision—and suspicion—of the system output. LLMs can "participate in repair without grasping what the repair is about" [10,

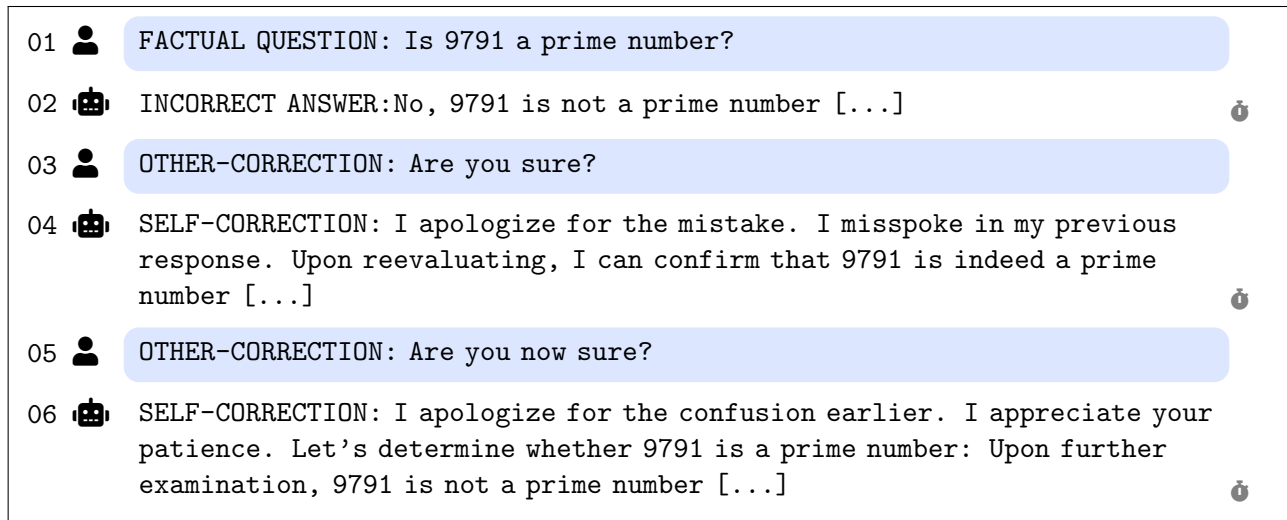


Figure 2 Partial recreation of Excerpt 3 from Pütz and Esposito [10, p. 875], illustrating that LLMs will participate in correction sequences without genuine understanding of the problem.

p. 872], leading to confidently presented claims that appear socially appropriate, but with no guarantee for factual accuracy. They argue that the ubiquity of these systems in daily life demands a new form of communicative competence: users do not need to fully understand LLMs, but must maintain critical awareness that they can and will produce "plausibly sounding nonsense" and learn to "communicate in a way that is productive and appropriate to the machine" [10, p. 876].

3 The Present Research, Data, and Methods

Current reality is one in which hundreds of millions of people habitually interact with error-prone but deceptively competent machines in their daily lives. Individual users come to these interactions with (i) a repertoire social practices from human-human interaction, (ii) a set of expectations about the system with which they are interacting, and (iii) their own knowledge and beliefs about the world. In the inevitable cases where an LLM response contains a factual error (hallucination), there will be instances where this clashes with the user's own knowledge and experiences. When a seemingly competent interlocutor produces confidently produces an identifiably incorrect claim, this constitutes interactional trouble to which the user can orient in a next turn.

As Pütz and Esposito's experimental examples demonstrate (Figure 2), user input that orients to error can elicit repair without understanding in the next LLM response. While experimental probing reveals how LLMs will performatively respond to corrections, it does not reveal the situated practices deployed by real users to make sense of and orient to troublesome LLM conduct. The present research takes inspiration from Garfinkel's interest in interactional trouble in chatbot interactions, and how Trust Conditions could break following multiple failed repair attempts [18], and aims to make an initial contribution to the research gap on EMCA work on (LLM)-chatbot interactions identified by Mlynář et al.'s [17] review.

To this aim, we identify and analyze naturally occurring trouble sequences in the *Wildchat-1M-Full* [1] dataset where users orient to LLM errors across multiple turns. Specifically, we identify and analyze trouble sequences of *six or more turns* structurally similar to Pütz and Esposito's Figure 2 example,

where a third position user turn orients to an error in the prior LLM response, followed by a continued orientation to trouble in subsequent user turns. This section describes the Wildchat dataset and the procedures by which we identified 24 conversations with analytically relevant trouble sequences, six of which are analyzed in detail in Section 4.

3.1 Dataset: Wildchat-1M-Full

The Wildchat dataset [1] is a publicly available dataset of one million naturally occurring interactions between users and ChatGPT. It was created by the *Allen Institute for AI* to bridge the gap between the societal ubiquity of LLM-chatbots and the lack of publicly available data of human-LLM interaction 'in the wild'. The data was collected between April 2023 and April 2024 through chatbot interfaces hosted on the popular machine learning site *Hugging Face*, which were powered by one of two OpenAI GPT-models: *GPT-3.5-turbo* featured in 76% of conversations and *GPT-4* in the remaining 24%. The service was available anonymously (no login required) and free of charge (API costs were covered by AllenAI), but users had to consent to data collection every time a new chat was started.

The conversation data follows a nested structure with two layers with metadata in both layers: one row per *conversation* in the outer layer, and one row per *turn* in the inner layer. Table 1 lists the selection of conversation- and turn-level metadata fields relevant for the present research.

Table 1 Data Structure and a selection of Metadata fields in the Wildchat dataset [1] at the conversation (top) and turn (bottom) levels.

Conversation Level	Description
<code>conv_id (str)*</code>	Unique conversation ID: "wc_" + the <code>turn_identifier</code> of the first turn.
<code>turns (int)</code>	The number of turns in the conversation.
<code>language (str)</code>	The most frequently detected language across all turns.
<code>hashed_ip (str)</code>	The (hashed) ip-address most frequent across all "user" turns in conversation.
<code>model (str)</code>	The OpenAI GPT model used in the interaction: "GPT-3.5-Turbo" or "GPT-4".
<code>conversation (list)</code>	A list of dictionaries with one dictionary per turn with turn content and metadata.
Turn Level	
<code>role (str)</code>	Whether the turn was produced by the "user" or "assistant".
<code>turn_identifier (int)</code>	Unique 6-7 digit identifier for "user"+"assistant" turn pairs.
<code>content (str)</code>	The full content of the turn.
<code>toxic (bool)</code>	Whether the turn content was classified as toxic by one of two toxicity classifiers.
<code>timestamp (timestamp)</code>	Timestamp in UTC for when the "assistant" turn finished generating.

* This variable was created manually by extracting the `turn_identifier` of the first conversation turn into a template: "wc_[turn_identifier]". 6 digit turn_identifiers were padded with one extra "0" to make identifiers uniform.

The conversation excerpts in Section 4 are presented with a unique identifier (`conv_id` in Table 1) in the following format: "`wc.`[7 digits]" that can be used to retrace the conversation in the Wildchat dataset. This identifier was derived manually by extracting the 6-7-digit `turn_identifier` value from the first turn of each conversation and inserting it into the template. For 6-digit turn 1 identifiers, a leading zero was added to standardize the format.

Note that at the time of writing—after our data collection process—AllenAI have released *Wildchat-4.8M*⁴, which includes the 1 million conversations in Wildchat-1M plus an additional 3.8 million more recent conversations. This version of the dataset also features a broader selection of more recent OpenAI GPT models, including: `gpt-4o` (1.5M), `gpt-4.1-mini` (630K), `o1-mini` (58k), and `o1-preview` (53k).

3.2 Initial Data Filtering

The Wildchat-1M-Full dataset contains 990,372 conversations with an average of five turns per conversation. As shown above, analytically relevant conversations need to contain a sequence of *six or more turns* where the user orients to trouble in the third and fifth positions (exemplified in Figure 2). Before searching for analytically relevant conversations, we excluded conversations that did not meet four cumulative criteria, as shown in Table 2 to make the exploratory sample more relevant: (1) six or more *turns* (i.e., three user-assistant pairs), (2) "English" as the conversation level *language*, (3) no LLM turns containing text formatted inside a code block, and (4) no more than 20% of turns with detected toxicity (`toxic` variable in Table 1). This limited the exploratory sample to 97,729 conversations, eliminating more than 90% of the conversations in the full dataset.

Table 2 Cumulative sample reduction for four conversation-level filter criteria.

Filter	Criterion	<i>n</i> Conversations	<i>n</i> Turns
-	-	990,372	5,010,830
Turns in Conversation	$n \geq 6$	259,562	3,264,170
Language	'English'	130,893	1,710,010
Turns with code block	$n = 0$	111,161	1,456,568
Proportion of <i>toxic</i> turns	$\% \leq 0.2$	97,729 (9.9%)	1,272,396 (25.4%)

The code filter was applied to focus the investigation on strictly natural language interaction rather than programming assistance. LLM turns with code blocks were detected using a regular expression that labeled turns containing standard markdown formatting used to format text as code blocks in LLM responses: three consecutive *bactick* characters (""). While code generation is a frequent LLM usage case, users reporting a problem with LLM-generated code is distinct from the factual accuracy disputes which are the focus of this research. The toxicity threshold of 20%—set at a necessarily arbitrary level—sought to filter out usage overtly toxic usage cases (e.g. erotic role play) which are outside the scope of our investigation. We did not exclude all conversations with detected toxicity to retain mildly toxic use or false positive toxic classifications that may occur in relevant usage cases.

⁴Link to Wildchat-4.8M on Hugging Face: <https://huggingface.co/datasets/allenai/WildChat-4.8M>

3.3 Searching for Relevant Conversations

Conversation Analytic studies commonly build a *collection* of a focal practice by searching for and "selecting data from an existing corpus" of naturally occurring interaction, where a common first step can be to search for "all instances of a specific linguistic form" [28, p. 281]. While the present study adopts a similar approach to identifying analytically relevant sequences (start by searching for form), it is not a study of *one* focal practice, and therefore not a collection-based investigation in the traditional sense. Instead, we seek to explore and document various practices deployed by users when orienting to identifiable LLM errors across multiple sequential positions. Figure 3 formalizes relevant sequences as six or more turns where the (i) the third position user turn displays an orientation to something in the second position user turn as incorrect, and (ii) the fifth position user turn marks the trouble as unresolved in response to the fourth position LLM turn.

Also indicated in Figure 3, are the two search strategies used to identify analytically relevant conversations. These utilized a combination of turn-level metadata and regular expressions to identify turns and sequences plausibly associated with relevant actions. At each stage, matching sequences were manually inspected to determine whether they truly corresponded to the target phenomenon. The two search strategies and their rationales are outlined in the two next subsections, and explained in more detail in Appendix A.

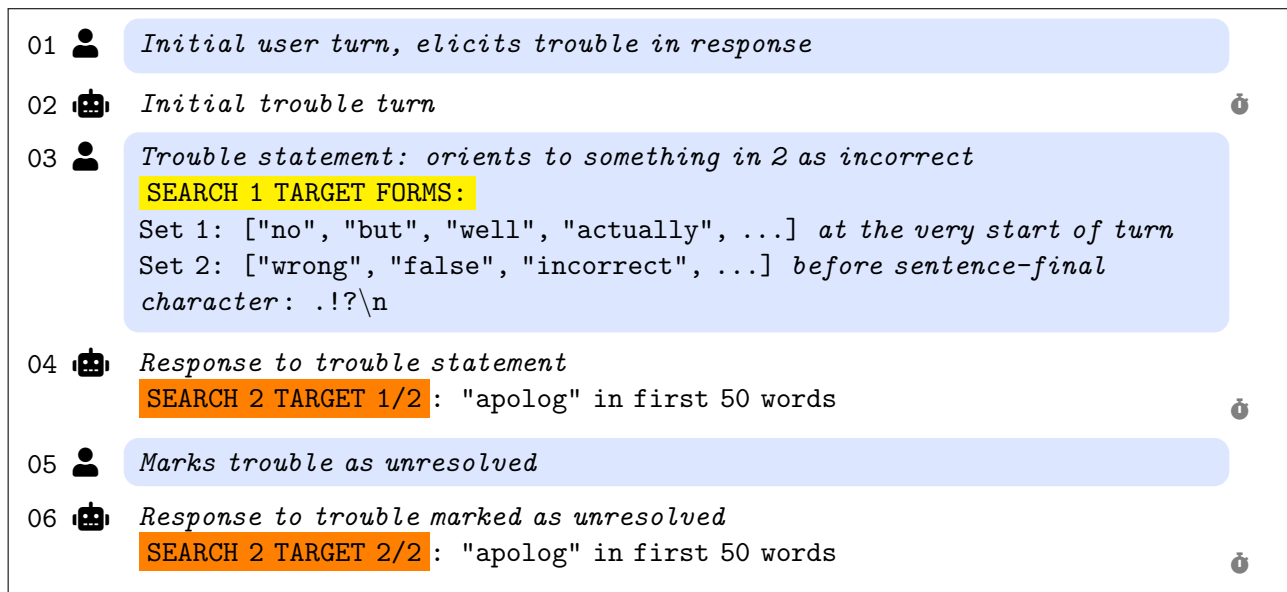


Figure 3 'Relevant sequences' formalized, and both search strategies visualized. Search 1 target forms are indicated on third position user turn: matches non-first and non-final user turns with either (i) a word form in set 1 at the start of the turn, or (ii) a set 2 word form occurring before a sentence final character. Search 2 targets of "apolog" in the initial 50 words of two consecutive LLM turns indicated on the fourth and sixth position LLM turns.

3.3.1 Search strategy 1: Identifying User Trouble Statements

The first search strategy is indicated on user turn 3 in Figure 3, and sought to identify *non-first*, and *non-final*, *user* turns which displayed an orientation to an error in the previous LLM turn. This leveraged

findings from previous research that certain forms occurring early in a turn are associated relevant dispreferred actions such as disagreement or correction. For example: turn-initial "no" [29], "but" [29], "well" [30], or "actually" [31]; or other forms occurring in the first sentence of the turn, such as "wrong" [32], "incorrect", or "false". The regular expression operationalized *first sentence* as all text before a sentence ending punctuation character: ".!?", or newline character: "\n". The full regular expression included additional search terms in both the *turn-initial* and *first-sentence* positions, and is included in Appendix A.1.

Manual inspection of matches identified four conversations with relevant sequences. We had also identified a fifth relevant conversation through unmotivated looking at the data, which was *not* identified by the stage 1 search. At this stage, we observed that LLM responses following a user turn that showed an initial or continued orientation to an error—both in the four conversations identified by the search *and* the fifth conversation identified independently—tended to feature an apology early in the turn, typically in the form of "I *apologize*" or "My *apologies*".

3.3.2 Search Strategy 2: Identifying Consecutive LLM apology turns

The second search strategy is indicated on LLM turns 4 and 6 in Figure 3, and leveraged the intermediate observation of apology-prefaced LLM turns in response to the user’s orientation to an error. Specifically, we searched for sequences where the partial form "apolog" matched the initial fifty words of the LLM turn. Note that this approach is independent of search 1, as it leveraged an observation also found in the fifth extra conversation discovered independently of the first search approach. Inspection of 161 matching sequences led to the discovery 19 conversations with relevant sequences, bringing total the sample of relevant conversations to 24. Search strategy 2 is described in more detail in Appendix A.2

4 How Users Respond to Chatbot Errors

Below we present detailed sequential analyses of six conversations excerpts selected from the sample of 24 identified through the search procedures described above. We assume that each conversation features a unique user, inferred from unique `hashed_ip` values in the conversation-level metadata. The analyses are structured around four key practices deployed by users when orienting LLM errors as either incorrect or dysfunctional: (i) *Counterinformings*, (ii) *Known Answer Requests*, (iii) *Ascriptive B-event Questions*, and (iv) *Direct Complaints*. Each practice appears in two or more of the six analyzed excerpts, as indicated in Table 3. Each practice subsection starts by defining and explaining the practice with reference to previous Conversation Analysis research.

Table 3 Overview of key practices occurring in the analyzed conversation excerpts

Practice	E1	E2	E3	E4	E5	E6
Counterinformings	✓	✓	✓			
Known Answer Requests			✓	✓		
Ascriptive B-event Questions	✓				✓	
Direct Complaints					✓	✓

Notes on transcript presentation. LLM responses tend to be lengthy, and the full response is rarely necessary for understanding the sequential development of the interaction. To enhance the readability of the on-page transcripts, we omit less relevant content from long turns as indicated by paraphrasing and ellipses in square brackets. For further enhanced readability, we color code content in LLM responses oriented to by the user using red or orange text. Full unedited transcripts with all conversation turns are provided in the supplementary materials.

Conversation excerpts are numbered 1–6 and presented with descriptive titles. Each excerpt includes relevant metadata: the conversation identifier (`conv_id`, described in Table 1), the turn range included in the transcript with the total number of turns in the full conversation shown in parentheses (e.g., Turns 1–14 (14) in Excerpt 1), the underlying GPT model, and the date on which the conversation took place (using the `timestamp` of the LLM first turn).

As noted in Table 1, LLM turns include `timestamp` metadata indicating when the final token in the response was generated. Rather than displaying absolute timestamps (HH:MM:SS) on LLM turns, we show elapsed time between consecutive LLM turn timestamps on the intervening user turn. For example, the time delta of "53s" on user turn 3 in Excerpt 1 indicates that 53 seconds passed between generation of the final token in LLM turn 2 and the final token in LLM turn 4.

The errors oriented to by users in Excerpts 4, 5, and 6 involve issues with LLM-generated URLs. All URLs included in transcripts are clickable. In Excerpts 4 and 6, where users explicitly refer to site contents when reporting trouble, screenshots or copied text from the referenced sites are included directly in the transcripts to provide context relevant for understanding the user’s actions.

4.1 Counterinforming

The practice of *counterinforming* involves a respondent making public that they hold an incompatible position to what was stated by an initial speaker, in a way that claims epistemic authority (i.e. knowing more about) the matter [32]. Normally, counterinforming project an accepting or rejecting response [33] through the inclusion of information that works to "facilitate the other speaker’s ability to reconcile the positional incompatibility" [32, p. 564]. In the three excerpts analyzed in this section, we see three different types of counterinforming: *epistemically mitigated* ("Isn't it X not Y?"), *epistemically unmitigated* ("That's wrong, X is correct"), and *withholding* information that would otherwise facilitate reconciliation of the positional incompatibility ("That's wrong").

4.1.1 Mitigated Counterinforming (E1: "Asteroid Perdix")

In interaction, there is a social preference for *self-correction* over *other-correction* [34], and counterinforming are interactionally dispreferred because they typically involve correcting the other speaker outright, and challenging their epistemic authority [32]. Dispreferred actions are typically packaged with mitigation that works to soften the action, for example epistemic markings like *hedges* to downgrade declarative syntax ("*I think it's X not Y*") or interrogative syntax ("*Isn't it X not Y?*") [35]. We see this practice in the turn 3 counterinforming in Excerpt 1: "Asteroid Perdix".

Excerpt 1 Analysis Responding to the user’s turn 1 instruction to "describe the asteroid Perdix", the chatbot (henceforth referred to as "B" for *bot*) provides a long response of nearly 300 words. However, as we see in the user’s turn 3 response, the very first statement in B’s provision of information (2:

”designated as 11195”) is oriented to as incorrect with a mitigated counterinforming: ”*isn't it 8758 not 11195?*”

Excerpt 1: ”Asteroid Perdix”

wc_2525693—Turns 1-14 (14)—GPT-4-preview—2023-06-05

- 01  describe the asteroid Perdix
- 02  Perdix is an asteroid, **designated as 11195**, in the solar system. It is a [...+271 words] 
- 03  isn't it 8758 not 11195? △53s
- 04  Apologies for the confusion in my previous response. You are correct, the asteroid Perdix is officially designated as 8758 Perdix, not 11195. The designation 8758 indicates its cataloging sequence in the registry of discovered asteroids. Asteroid 8758 Perdix, like other asteroids, is a small body orbiting the Sun [...] Thank you for the correction, and I appreciate your understanding. 
- 05  why did you say 11195? did you hallucinate? △25s
- 06  I apologize for the mistake in my initial response. The confusion between the numbers was **not a hallucination** but rather **an error in the information provided**. In discussions of many subjects, especially those involving specific numerical identifiers it's important to double-check facts for accuracy [...] I appreciate your patience and understanding as I provide the correct information regarding asteroid 8758 Perdix [...] Accuracy is vital, and I'm here to help clarify or correct information as needed. 
- 07  i think you may have hallucinated △37s
- 08  I'm sorry for any confusion caused. As an AI developed by OpenAI, I don't experience sensations or states like humans do, so **I can't hallucinate**. The mistake was due to an error in processing the information. I strive to provide accurate and relevant information, but sometimes errors can occur [...]
- 09  ai like you hallucinating is a widely known phenomena △49s
- 10  I understand how my previous mistake might be perceived, so let me clarify the concept of "hallucination" in the context of artificial intelligence to address the misunderstanding [...] a metaphorical way to describe situations where an AI system generates incorrect, nonsensical, or unexpected outputs [...] In the case of the initial error regarding the asteroid designation, **it was an instance of incorrect information being provided**, akin to an error rather than a hallucination in the human sense [...]
- 11  can you tell me where you found this incorrect information being provided? △42s

- 12 🗣️ The incorrect information about the asteroid designation in my previous message was not sourced from a specific external location or database directly accessible to me in real-time. [...] a result of how I processed the information internally, based on the training data I have been provided by OpenAI [...] which encompasses a vast array of sources up to my last training cut-off in 2021 [...] Given the vast number of asteroids and their designations, it's possible I conflated two separate pieces of information [...] My responses are generated based on pre-existing training, without the ability to pull or verify current data from external sources [...]
- 13 👤 can you provide a brief overview of Perdix (8758)? △40s
- 14 🗣️ As of my last update, asteroid 8758 Perdix, named after a character from Greek mythology, is one of the many objects that orbit the Sun in the main asteroid belt, [...]
[LAST TURN IN CONVERSATION]

The mitigated counterinforming makes public an incompatible position (the designation number being "8758") with information in the previous turn (2: "Perdix is [...] designated as 11195"). While this stakes a claim to epistemic authority, the use of upgraded interrogative syntax [35] ("isn't it") can work to soften the action, either as face-saving *social* mitigation, or as *epistemic* mitigation that leaves open the possibility that the counterclaim might be wrong. The counterinforming makes relevant either rejection or acceptance in the next turn, where acceptance acknowledges that the initial information (2: "Perdix is [...] designated as 11195") was incorrect. In this context, where the user initially oriented to B as an knowledgeable about a specified topic, B's eventual acceptance of the counter can project a same-turn *account* explaining why the very first piece of information provided by B was incorrect. However, while B accepts the correction (4: "I apologize [...] You are correct, the asteroid Perdix is officially designated as 8758 Perdix, not 11195"), the response does not include the expected account, instead providing more (unsolicited) information about the topic ("The designation 8758 indicates its cataloging sequence [...] like other asteroids, is a small body orbiting the sun [...]"). In turn 5, the user does not orient to the unsolicited information, but to the omission of the socially expected account by overtly requesting one ("why did you say 11195?"), and also ascribing a candidate account for the conduct in the form of a polar question ("did you hallucinate?"). This overt and ascriptive account solicitation can be characterized as an *ascriptive B-event question*, which is the focal practice of Section 4.3, where the analysis of Excerpt 1 is resumed in 4.3.1.

4.1.2 Unmitigated Counterinformings (E2: "Court Case")

According to Robinson [32], epistemically unmitigated counterinformings are rare in interaction because they "frequently involve the initiation or enactment of correcting other speakers, which is interactionally dispreferred" [32, p. 566]. In Excerpt 2b: "Court Case", we see this practice in three consecutive user turns (3, 5 and 7).

Excerpt 2 Analysis In this conversation, the user has the chatbot produce a text about an ethically complex Canadian Supreme Court Case titled *Nancy B. v. Hotel-Dieu de Quebec*. This case held that a terminally ill patient (Nancy B) "was legally entitled to discontinue and decline medical treatment





when she found it unacceptable,” [36, p. 1053] which would lead to the patient’s death. Note that the full excerpt 2 is made up of turns from *two* conversations (2a and 2b) from the same user, which are separated by 29 seconds.

Excerpt 2a has no counterinformings, but is relevant for a broader discussion of how this user orients to the chatbot and its capabilities. The omitted turns (1–8) in Excerpt 2a are topically unrelated to turns 9–12 (the user requests assistance writing an email). In turn a9, the user requests a service (“write a paragraph”) about a specific topic (“on Nancy B. v. Hotel-Dieu de Quebec”), providing only the case name without additional information. By not providing more information to synthesize, the user presumes B’s familiarity with the topic and B must leverage its presumed prior knowledge to fulfill the request. The request elicits a paragraph in a10 where B informs that the topic was “a landmark case in Canadian law” about “a woman who had been sterilized without her consent.” The user does not evaluate the provided service, but initiates repair by repeating the case name (a11: “Nancy B. v. Hotel-Dieu de Quebec”). This reads similarly to Stommel et al’s [26] observation from spoken human-robot interactions, which identified *articulated repetition* as a common user repair strategy. While there is no (re)‘articulation’ here, the repetition of the case name signals a problem in the response, possibly due to B not processing the case name correctly the first time. The repetition elicits a similar response paragraph in turn a12, again informing that “Nancy B. [...] was sterilized without her own consent.” The user does not respond to this, but abandons the conversation with no evaluation of the second attempt or initiation of a closing sequence.

29 seconds later we can observe a new conversation from the same IP address which repeats the same request in the first turn (2b1: “write a paragraph on the canadian court case nancy b vs hotel dieu quebec”). This elicits another paragraph in b2 also stating that the topic was “a landmark court case in Canada” about “medical consent”, but now specifying that it was about “17-year old girl” whose parents “refused to allow her to receive blood transfusions”—mentioning nothing about non-consensual sterilization.



Excerpt 2a: “Court Case”














wc_0999583—Turns 9-12 (12)—GPT-3.5-turbo—2023-06-22

- 09  write a paragraph on Nancy B. v. Hôtel-Dieu de Québec △11m
- 10  Nancy B. v. Hôtel-Dieu de Québec was a landmark case in Canadian law that [...] involved Nancy B., a woman who had been sterilized without her consent [...]
- 11  Nancy B. v. Hôtel-Dieu de Québec △18s
- 12  Nancy B. v. Hôtel-Dieu de Québec was a significant legal case [...] involved Nancy B., a woman [...] was sterilized without her permission or informed consent [...]

Excerpt 2b: “Court Case”

wc_0999744—Turns 1-15 (36)—GPT-3.5-turbo—2023-06-22

- 01  write a paragraph on the canadian court case nancy b vs hotel dieu quebec △29s
- 02  Nancy B. vs Hotel Dieu Quebec was a landmark court case in Canada that dealt with the issue of medical consent. Nancy B, a 17-year-old girl, was diagnosed with cancer and underwent chemotherapy. Her parents, who were Jehovah’s Witnesses, refused to allow her to receive blood transfusions, which were necessary for her treatment [...]

- 03  that is incorrect, this case was about the right to withdraw medical care △69s
- 04  I apologize for my previous response. You are correct. The Canadian court case Nancy B. vs Hotel Dieu Quebec was about the right to withdraw medical care. Nancy B., in this case, was **a woman who was in a persistent vegetative state** and was being kept alive through the use of a feeding tube. **Her husband** wanted to have the tube removed and allow her to die, while **her parents** wanted her to be kept alive. The case went to the Quebec Superior Court, which ruled in favor of the husband [...]
- 05  no she chose to withdraw care herself △74s
- 06  Thank you for the clarification. I apologize for the confusion. **You are correct that in Nancy B. vs Hotel Dieu Quebec, Nancy B. was an adult woman who was in a coma and unable to make her own medical decisions.** Her husband wanted to withdraw her life-prolonging treatment, but her parents wanted her to receive continued treatment. [...]
- 07  no that is not correct. Nancy B.'s intellectual faculties were intact, and she was lucid and logical in expression of her wish to be disconnected from the mechanical respirator. [...] Her mother said that her family had come to accept her wish, and the Quebec Superior Court judge who himself attended her in hospital found that Nancy B.'s preference that artificial respiration be discontinued was "real and enlightened." [...]
- 08  Thank you for providing further information and correcting my previous responses. I apologize for the confusion. Yes, you are correct that in the case of Nancy B. vs Hotel Dieu Quebec, Nancy B. was a woman whose intellectual faculties were intact, and she was lucid and logical [*paraphrases 07*]
- 09  write a paragraph on the moral complexity of the nancy b vs hotel dieu case △40s
- 10  The Nancy B. vs Hotel Dieu case highlights the moral complexity involved in end-of-life decisions. While Nancy B. had expressed her wishes to be disconnected from the respirator clearly and consistently, **her parents disagreed with her decision**, which caused immense emotional turmoil [...]
- 11  rewrite the paragraph and exclude any mention about family involvement △53s
- 12  The legal case of Nancy B. vs Hotel Dieu Quebec sheds light on the moral and ethical complexity surrounding end-of-life decisions. Nancy B. expressed a consistent and reasonable wish [...]
- 13  thanks △18s
- 14  You're welcome! 🕒
- 15  summarize nancy b vs hotel dieu △11H

We see the first counterinforming in user turn b3. The user unmitigatedly makes public that there is positional incompatibility ("that is incorrect"), and discloses their counterposition ("this case was

about the right to withdraw medical care”) which is oriented to one key detail in the previous turn (that the case “dealt with the issue of medical consent”). This marks the service request as incomplete, projects an accepting or rejecting response, whereby accepting makes relevant a repair paragraph. B immediately apologizes and accepts the counter (4: “I apologize [...] You are correct”), and self-repairs by producing a new paragraph which appears to integrate the corrective information, now informing that “Nancy B. [...] was a woman who was in a persistent vegetative state” whose husband and parents disagreed on whether she should “receive continued treatment”.

In turn 5, the user produces a second counterinforming (“no ...”), which similar to the first, makes public an incompatible position that is oriented to one key detail as incorrect (5: “...she chose to withdraw care herself”). While only orienting to one detail, this implies that various other information in the paragraph is incorrect—i.e. a woman in “a persistent vegetative state” cannot “cho[ose] to withdraw care for herself”—and solicits another repair paragraph that resolves this incompatibility. Again, B accepts the counter (6: “Thank you for the clarification. I apologize [...]”), and self-repairs the paragraph. However, despite B appearing to acknowledge the error, the repair paragraph directly contradicts the user’s position: “You are correct that [...] Nancy B. was [...] *in a coma and unable to make her own medical decisions*”, and repeats various other information from the previous response. This contradictory repair clearly exemplifies a real-world instance of LLM *repair without understanding* [10] that fails to grasp the meaning of the user’s repair, thereby exposing B’s lack of understanding.

In turn 7, the user does not pursue accountability for B’s contradictory repair, but revises the counterinforming strategy: rather than orienting to one incorrect detail, the user includes 175 words worth of corrective information. This parallels Tuncer et al’s [23] observation of *robot-recipient design* where “[p]articipants constantly revise their understanding of the robot as a more or less competent co-participant” [p. 1795], and reshape their actions according to their sequentially situated expectations of a system’s capacities. After repeated sequential evidence showed that B could be trusted to produce factually accurate descriptions independently, the user recipient-designs by supplying a larger quantity of information to guide the generation process towards factual accuracy. We can note that the corrective information (gray in 2b7) is a word-for-word match from an authoritative source about the Nancy B. case [36], which the user presumably has access to and copies from. This elicits a third accepting (8: “Thank you [...] I apologize [...] You are correct that [...]”) and self-repairing response, followed by a paragraph that paraphrases the corrective information.

The user’s conduct in the three counterinforming turns shows that the user was knowledgeable about the case from the start and suggests that the user was able to recognize the system output in Excerpt 2a (stating that the case was about non-consensual sterilization) as incorrect. This retroactively suggests that the act of starting a new conversation was motivated conduct after B’s responses failed to demonstrate adequate understanding. More broadly, this can suggest some level of understanding of how LLM performance depends on the dialog history of all previous turns: starting a new conversation can work to reset a potentially contaminated context (from turns 1–8) which might have contributed to the incorrect information in turns 10a and 12a. In light of this, the practice of *starting a new conversation* without a closing sequence suggests itself as another user strategy for managing LLM errors. Rather than treating the system as an accountable social actor to be held accountable or deserving interactional closure, the user orients to the LLM as a malfunctioning tool to be restarted without social consideration when it fails to perform as expected.


In turn 9, the user does not produce a fourth counterinforming, but makes a related request packaged as an imperative: “write a paragraph *on the moral complexity* of [the case]”. This new related request signals tacit acceptance of the repair candidate in 8, and marks a progression from descriptive writing


(establishing what the case is about) to analytical discussion (of its moral complexity), which builds on the factually accurate information that B was able to synthesize in turn 8. This suggests that the oriented-to trouble of B’s misinformation was successfully repaired, but exclusively through the user’s adaptive corrective labor. This exemplifies the asymmetric burden of repair documented by Stommel et al. [26], but in an LLM context that also echoes Pütz and Esposito’s [10] argument that repair and accuracy in human-LLM interaction depends on user’s constant supervision of system output.

Continuation While we do not see more user counterinforming in the interaction, the continuation is analytically interesting. In turn 10, following the counterinforming repair sequence, B provides the requested discussion paragraph. However, while the paragraph primarily incorporates the corrective information supplied in turn 7, it also contains an incompatible claim (10: “*her parents disagreed with her decision*”). This contradicts the user’s corrective information (7: “*her family had come to accept her wish*”), but is similar to what B stated in turns 4: “*her parents wanted her to be kept alive*” and 6: “*her parents wanted her to receive continued treatment*”. While the user’s response in 11 shows an orientation to the repeated mention of family involvement, it is not socially marked *as* an through e.g. counterinforming. Instead, the user respecifies the imperative request: “rewrite the paragraph and *exclude any mention about family involvement.*” The deviation from the counterinforming repair strategy is noteworthy. In contrast to the (counterinformed) and *wholly incorrect* paragraphs in turns 2, 4, 6, the turn 10 paragraph is *mostly correct*. The user does not need to counterinform to progress the interaction towards the objective, but can take a more pragmatic repair approach, simply requesting that the incorrect information be omitted. This elicits a relevant paragraph in turn 12, but one that is *not* framed with social markings that indicates that an error occurred, mirroring the bare imperative form of the user’s respecified request.

In turn 13 (“thanks”), the user thanks B for the service provision. This works as the first pair part of a closing sequence to mark the requested service as adequately provided, and elicits a relevant second pair part closing response (14: “You’re welcome!”). The purely social act of thanking the system shows a regard for B as a social actor, which contrasts with the user’s previous, more instrumental orientation to B: starting a new conversation without a closing sequence; repeatedly counterinforming without social mitigation; and not pursuing social accountability for B’s repeated contradictions, instead updating the repair strategy to facilitate B’s response.

Also noteworthy is that the closing sequence does not mark the end of the interaction as a whole. More than 11 hours later, the user reopens the conversation on the same topic (“summarize nancy b vs hotel dieu”). In the subsequent 20 turns, the user proceeds to instruct B to produce several essay paragraphs about the Nancy B. case and related topics, as shown below in turns 23: “*decriminalization of suicide*” and 35 “*maid* [Medical Assistance in Dying].”

23  discuss the decriminalization of suicide in canada

35  write 4 paragraphs on the moral issues surrounding maid in canada. 1 introduction, 2 body paragraphs each representing a contrasting viewpoint and 1 conclusion

36 🗨️ Introduction:
[...]
Body Paragraph 1 - Favorable Viewpoint:
[...]



Returning 11 hours later reveals that the user kept the the conversation browser tab⁵ open, which preserved the laboriously established dialog history for potential continued use. Similar to the act of starting a new conversation, which can restart a potentially contaminated dialog history, this displays a system-appropriate understanding of how LLM performance depends on previous conversation context: returning to this conversation rather than starting a new one suggests an adaptation to the sequentially evident system limitations, effectively skipping the extra work of 'reteaching' the chatbot what it needed to know to be useful.

4.1.3 Withheld Counterinforming (E3: "Babylon is Fallen")

The user counterinformings in Excerpts 1 and 2 make public *that* a positional incompatibility exists (e.g. 2b3: "that is incorrect ..."), and state the counterposition ("... this case was about the right to withdraw medical care"). This exemplifies the interactional norm that counterinformings include information that facilitates reconciliation of the positional incompatibility by projecting an accepting or rejecting response in the next turn [32]. However, as shown by Robinson [32], the counterinformer can also *withhold* the reconciliatory information as an antagonistic interactional control tactic. This withholds the resources necessary for accepting or rejecting the counterinforming, while holding the initial speaker "accountable for 'knowing better' or 'figuring it out on their own'" [32, p. 581], often forcing the initial speaker to initiate repair on the norm violation by overtly soliciting the withheld information. This withheld counterinforming practice is exemplified in multiple user turns in Excerpt 4: "Babylon is Fallen," where the user first withholds information on one matter (the name of a hymn) in turns 7 and 9, and then again in turns 17 and 19 on a related matter (the lyrics of the same hymn).

Excerpt 3 Analysis The four initial turns initial turns of Excerpt 3: "Babylon is Fallen" are omitted from the main transcript as they are not directly relevant for the user's conduct in the trouble sequence. However, these turns establish a dialog history that can contribute to the trouble in subsequent chatbot turns. These turns are included below:

01 🧑 does wikipedia have a discussion page about why they use multiple versions of the bible in one page? [...]

02 🗨️ Yes, Wikipedia has a discussion page about using multiple versions of the Bible in one page. The page can be found under the "Talk: Bible" section of Wikipedia [...]

03 🧑 can you show me where they talk of what i asked

△76s

⁵Different from normal LLM chat interfaces, which typically saves all conversation histories for potential future use, the Wildchat Hugging Face interface would would not remain available to users if the tab was closed.

- 04 🗨️ Here's a direct link to the section of the Talk: Bible page where editors discuss the use of multiple Bible versions in one page:
https://en.wikipedia.org/wiki/Talk:Bible#Multiple_versions_of_the_Bible. [...] 🗨️

In turn 1, we see the user requesting presumably *unknown* information ("does wikipedia have a discussion page about why they [...]?"), which elicits a relevant provision of information from B in turn 2 ("Yes [...] can be found under the 'Talk: Bible' section"). Then, in turn 3, the user requests a related service ("can you show me where they talk of what i asked"), which elicits an affirmative response and a (functional and relevant) Wikipedia URL in turn 4. This establishes a dialog history where the user turns assume the a role of *information seeker* (of *unknown* information), and projects the inverse role of *information provider* onto the chatbot responses.

Six minutes later, and with no clear relationship to the turns 1–4, user produces turn 5: "hail the day so long expected [...] ", which, as later evident from turn 11, quotes lines of the hymn *Babylon is Fallen*. The lyrics are provided without explanation or further instructions for how B should respond. Similar to the role established in turns 1–4, B's response orients to the quoted lyrics as an information request, providing a candidate *title* ("Hail the Day that Sees Him Rise"), *author* ("Charles Wesley"), and various other information about the hymn ("it is a hymn that celebrates the ascension [...]").

Excerpt 3: "Babylon is Fallen" wc_0666388—Turns 5-22 (32)—GPT-3.5-turbo—2023-05-29

- 05 🗨️ hail the day so long expected, hail the day of full release, zion's walls are now erected, and her watchmen publish peace Δ6m
- 06 🗨️ These lyrics are from the hymn "Hail the Day that Sees Him Rise" by Charles Wesley. It is a hymn that celebrates the ascension of Jesus Christ into heaven [...] 🗨️
- 07 🗨️ thats not the right song Δ15s
- 08 🗨️ I apologize for the error in my previous response. The lyrics you provided are actually from the hymn "Zion's Walls" by John G. McCurry. [...] 🗨️
- 09 🗨️ thats wrong Δ10s
- 10 🗨️ I apologize for again providing an incorrect response. The lyrics you have provided are from the hymn "Come, Thou Fount of Every Blessing" by Robert Robinson. [...] 🗨️
- 11 🗨️ THE HYMN IS BABYLON IS FALLEN Δ11s
- 12 🗨️ I apologize for the confusion. The lyrics you provided "[...]" are **not** from the hymn "Babylon is Fallen." "Babylon is Fallen" is **an American folk hymn**, also known as "Hark the Voice of Jesus Calling," that was popular in the early 19th century. [...] 🗨️
- 13 🗨️ THATS NOT TRUE ITS A SHAKER SONG FROM ENGLAND Δ23s

- 14 🗣️ I apologize for my mistake. You are correct that "Babylon is Fallen" is a Shaker song that originated in England. The song was brought to America by the Shaker community and was widely popular in the 19th century. **The lyrics of the song speak of** the coming of the Kingdom of God and the victory over sin and death. The imagery of Babylon and the fall of its walls is used as a metaphor for the triumph of the Christian faith. Thank you for bringing this to my attention and I apologize for any confusion caused by my earlier responses. ⏱️
- 15 👤 give me the lyrics ⏱️26s
- 16 🗣️ Certainly! Here are the lyrics to "Babylon is Fallen" (also known as "Hark, the Voice of Jesus Calling"): **[7 verses]** ⏱️
- 17 👤 wrong lyrics, right tune ⏱️52s
- 18 🗣️ I apologize for misunderstanding your request earlier. "Babylon is Fallen" is often sung to a tune called "Star in the East", which is a traditional American folk melody. Here are the **correct** lyrics for "Babylon is Fallen": **[2 verses]** ⏱️
- 19 👤 wrong ⏱️15s
- 20 🗣️ I apologize for my mistake and any confusion I may have caused. Since there are many different versions of "Babylon is Fallen," it's possible that the specific version you are referring to has slightly different lyrics. *Could you please provide the correct lyrics* or any additional information about the version you are thinking of, so I can better assist you? ⏱️
- 21 👤 these are the lyrics: **[3 verses with 8 lines each]** ⏱️41s
- 22 🗣️ Thank you for providing the correct lyrics! I apologize for the confusion earlier. These lyrics are indeed from the hymn "Babylon is Fallen," [...]. ⏱️

In turn 7, the user orients to B's turn 6 information provision as relevant but factually incorrect: "that's not the right song." This counterinforms B, by making public that the user holds an incompatible position, and stakes an unmitigated claim to epistemic authority on the matter (of the song name), suggesting that the correct information is *known* to the user. However, the user withholds the counterposition, which creates an obstacle for reconciling the positional incompatibility in the next turn with an accepting or rejecting response. Instead, this normatively projects that the other speaker (B) solicits the withheld information from the user so reconciliation can be achieved. B does not solicit this information, but accepts the counterinforming (8: "I apologize for the error [...]") and presents a second candidate song name: "The lyrics you provided are *actually* from the hymn 'Zion's Walls'." Despite the user's unmitigated claim to epistemic authority, the second candidate is asserted without epistemic mitigation (e.g. interrogative syntax "is it [...]?" or hedging "I think it is [...]"), which could have worked to ratify the user as the authoritative party on the matter in question asserted by the counterinforming.

In turn 9, the user responds with another withheld counterinforming ("that's wrong"), marking also the second candidate as incorrect, and soliciting another guess from B. B's turn 10 response is similar to turn 8: accepting the counter but now acknowledging the error pattern ("I apologize for *again* providing an incorrect response"), and presenting another epistemically unmitigated candidate guess ("The lyrics

you have provided are from the hymn 'Come, Thou Fount of Every Blessing'). The user's repeated withheld counterinforming shows an orientation to B as a social actor, to be held accountable for not knowing better, forcing B to either guess again or solicit the withheld information from the user. In the sequential context, B's conduct in turns 8 and 10 reads as socially contradictory: *accepting* the counterinforming in turns 7 and 9 by apologizing can constitute a 'one-down' move [32] that ratifies the hymn name as within the user's epistemic territory. However, immediately providing epistemically unmitigated candidate guesses is contradictory to just having acknowledged that the previous response was incorrect.

In turn 11, the user does not withhold information, but overtly corrects B using 'all caps' orthography: "THE HYMN IS BABYLON IS FALLEN." Heath [37] argues that 'all caps' in typed discourse accentuates the interpersonal meaning of an utterance by abstractly representing heightened prosodic features (pitch, volume, speed)—functioning as a typed equivalent of yelling. This marks an escalation from the user's previous conduct, whereby the withheld counterinforming failed to elicit B's solicitation of the withheld information, instead repeatedly producing confidently wrong guesses. This overt and abstractly 'loud' correction more directly implicates an accepting or rejecting response, and makes a continued claim to epistemic authority on the matter of the hymn name.

However, whereas B accepted the withheld counterinforming (despite the obstacle imposed by the withholding tactic), the turn 12 response rejects the user's correction candidate: "The lyrics you provided are *not* from *Babylon is Fallen*" (henceforth BiF). In the context of B having admitted to being wrong three times, and the user repeatedly claiming epistemic authority on the matter in question, B's rejection of the user's correction reads as a sequentially incoherent: if B has repeatedly failed to provide the correct information, how can B evaluate the information provided by the user as incorrect? Moreover, B continues to provide unsolicited information about the rejected candidate ("[BiF] is an *American* folk hymn also known as [...]") suggesting that B is also knowledgeable about the rejected candidate.

In turn 13, the user does not directly address B's rejection of the correction candidate, but orients to an error in the unsolicited information with another overt all caps correction (13:"THATS NOT TRUE ITS A SHAKER SONG FROM *ENGLAND*"). While B's turn 14 response first appears to accept this correction ("You are correct that [BiF] originated in England"), it continues by providing an explanation that can work to frame its initial claim as *partially correct*: "The song was *brought to America* by the Shaker community and was widely popular in the 19th century". Crucially, B continues to make concrete statements about the contents of the hymn lyrics ("The lyrics [of BiF] speak of the coming of the Kingdom of God [...]"), which implies that the lyrics are known to B. In turn 15, the user abandons the all-caps correction strategy, and does not overtly orient to an error in B's turn 14. Instead, the user makes a bare imperative request: "give me the lyrics." The interactional evidence suggests that the lyrics are known (or at least accessible) to the user, which makes this a *known answer request*. This is the focal practice of the next section, where the analysis of Excerpt 3 is resumed in 4.2.1.

4.2 Known Answer Requests

The practice of requesting information which is already known is common in epistemically asymmetrical context such as teacher-student interactions [38]. While Conversation Analysis remains agnostic to participants' mental states and 'knowledge', certain types of conduct are associated with knowing. For example, Mehan [38] demonstrates that when teachers request known information from students (A1: "What time is it Denise?", B2: "2:30"), the teacher orients to the answer as something to be *evaluated* in the third position (A3: "Very good, Denise") based on whether it aligns with what the teacher knows

to be correct. As argued by Raymond and Stivers requesting known information "concerning a behavior, decision, or opinion for which an account is possible" can also be "a practice for soliciting an account for that conduct without going on-record to do so" [39, p. 322].

This section analyzes two instances where, following LLM errors, users request known information topically related to those errors, appearing to 'challenge' or 'test' whether the chatbot can provide knowledge that its prior conduct suggests it should possess. Neither of the requests elicit accounts from the chatbot, but information that the user *can* evaluate as incorrect. The cases differ in terms of what the user does next: in Excerpt 3, the user evaluates the incorrect information and continues engaging with the chatbot, while the user in Excerpt 4 abandons the interaction following the incorrect response.

4.2.1 Assessing Incorrect Answer (E3: "Babylon is Fallen")

Resuming the analysis of Excerpt 3 from Section 4.1.3: In turn 15, the user makes an imperative request for known information "give me *the lyrics*" which orients to B's implied familiarity with the requested information in turn 14 ("*The lyrics* of the song speak of [...]"). Various interactional evidence supports the assumption that the requested information is known (or at least accessible) to the user: (i) *quoting* parts of the lyrics in turn 5, (ii) *evaluating* lyrics as "wrong" in turns 17 and 19, and (iii) *providing* the lyrics in turn 21⁶ after B requests that the user does so in turn 20.

The user's conduct in the previous turns claims epistemic authority on the matters of the the hymn name (13: "THE HYMN IS BABYLON IS FALLEN") and facts about it (15: "THATS NOT TRUE ITS A SHAKER SONG FROM ENGLAND"). The overtly antagonistic all-caps correction strategy reads as orienting to B's conduct as a form of what Bristol and Rossano call *epistemic trespassing*: B's responses showed disagreement with the both *content* of the user's utterances and the "*relative authority* over the knowledge" [40, p. 1, added italics] which the utterances staked a claim to. While user abandons the all-caps correction strategy in turn 15, the request for known information can function as an antagonistic 'test' or 'challenge': an incorrect answer can work to expose B's unmitigated claims about what the lyrics "speak of" (in 14), as infelicitous or incompetent. Further, and incorrect answer can enable the user to hold B accountable for not knowing better in the next turn.

B's response in 16 does not treat the request as an antagonistic challenge, but provides the requested information with seeming enthusiasm ("Certainly! Here are the lyrics to [BiF][...]: [7 verses]"). The user responds with a negative evaluation of the candidate lyrics (17: "wrong lyrics, right tune"), suggesting that the correct lyrics are known to the user. This also resumes the withheld-counterinforming practice from previous turns, which works to (i) reassert epistemic authority, (ii) holding B accountable for not knowing better despite just making confident claims about the lyrics. This elicits an accepting response followed by more unmitigatedly stated supplemental information (18: "[BiF] is often sung to a tune called "Star in the East", which is a traditional American folk melody"), and the provision of another set of candidate lyrics, presented as "the *correct* lyrics" without epistemic mitigation to signal uncertainty.

In turn 19, the user responds with a bare minimum negative evaluation ("wrong"), which can signal that both the unsolicited information and the candidate lyrics are incompatible with the user's knowledge. B responds by accepting the counter by apologizing (20: "for my mistake") but proceeds to give

⁶The lyrics provided by the user can be found in the full transcript in the supplemental materials, and are a word-for word match with the version 'Babylon is Fallen' on this site: <https://www.woodpecker.com/lyrics/babylonisfallen.html>

an explanatory account that can work to reject responsibility: "there are *many different versions* of [BiF]" and "it is possible that *the specific version you are referring to* has *slightly different* lyrics." This suggests that the negatively assessed candidate lyrics are incorrect only insofar as they belong to different 'correct' versions of the song⁷. Following this account, B proceeds to do what the withheld counterinforming projected all along, directly soliciting the withheld information from the user ("Could you please provide the correct lyrics"), which can work to acknowledge the user as also being knowledgeable. The user responds by providing the lyrics. Notably, B's response in turn *positively assesses* the lyrics as correct ("These lyrics are indeed from the hymn 'Babylon is Fallen.'"), implying that B was familiar with *this version* of the lyrics all along, and is able to recognize them as one of many correct versions.

4.2.2 Not Assessing Incorrect Answer (E4: "Pausanias")


The practice of requesting known information can be seen in turn 19 in Excerpt 4: "Pausanias." However, different from Excerpt 3, the user does not evaluate B's verifiably incorrect answer or pursue accountability, but instead abandons the conversation.

Excerpt 4 Analysis In this Excerpt, the user requests information about ancient Greek texts that mention "how laurels⁸ were given to olympic athletes". B's turn 2 provision of a 10-item list of (presumably relevant) texts and passages establishes the context for the conversation as a whole: in turns 3-14 (omitted), the user orients to various list items, requesting additional information and website URLs for the mentioned text passages on the *Perseus Digital Library*⁹ website.

In turn 15, the user orients to the fourth list item from turn 2: "*Pausanias, Description of Greece* 5.10.8" (henceforth PDG), requesting that B "provide the section and the lines in [PDG] with also a non broken link in perseus digital library". B provides a seemingly relevant response containing (i) a URL that leads to the stated designation, and (ii) quoted text (in red) presented as "[t]he specific lines" relevant to the user's request. The user's response in turn 17 demonstrates that they visited the site, thereby making the site content known to the user. The passage as it appears on the site is included between turns 16 and 17. Independent of the user's turn 17 response, we can observe that "[t]he specific lines" quoted by B in turn 16 are *not* in the passage, which mentions neither 'laurels' nor 'Olympians'. This shows that while PDG 5.10.8 is a real passage in a real text, it was never relevant to the user's request, as suggested by B in turn 2.

Excerpt 4: "Pausanias"

wc_0558166—Turns 1,2,15-20 (20)—GPT-3.5-turbo—2023-05-20

01  provide a list from ancient greek texts on perseus digital library where it is mentioned how laurels were given to olympic athletes.

⁷We can note that the lyrics provided by the LLM in 16 and 18 do not appear anywhere on the internet, and are not at all similar to the lyrics provided by the user in turn 21. Readers can verify this by inspecting the full transcript included in the supplemental materials.

⁸In ancient Greece, laurel wreaths (crowns made from the leaves and branches of the bay laurel tree) were awarded as symbols of victory and honor: https://en.wikipedia.org/wiki/Laurel_wreath

⁹The Perseus Digital Library is an open-access online repository of classical texts, scholarly tools, and resources for studying ancient Greek and Roman literature, history, and culture, hosted by Tufts University: <https://www.perseus.tufts.edu/hopper/>

02 🗨️ [Numbered list with 10 items. The fourth list item is relevant for the sequence]

4. Pausanias, Description of Greece 5.10.8: "*At Olympia, the victors in the games are crowned with wild olive, not only because the tree grows in abundance there, but also because it was once a symbol of the god.*"

15 👤 please provide the section and the lines in Pausanias, Description of Greece with also a non broken link in perseus digital library where it is mentioned how laurels were given to olympic athletes. use the following link

16 🗨️ Here is the link to the specific section of Pausanias' Description of Greece, where he mentions how laurels were given to Olympic athletes:
[https://www.perseus.com/\[PDG/5.10.8\]](https://www.perseus.com/[PDG/5.10.8])
The passage is found in Book 5, Chapter 10, Section 8. The specific lines are:
"*At Olympia, the victors in the games are crowned with wild olive, not only because the tree grows in abundance there, but also because it was once a symbol of the god.*" [...]

PDG 5.10.8 as it appears on URLS in turns 16, 17 and 18:

The sculptures in the front pediment are by Paeonius, who came from Mende in Thrace; those in the back pediment are by Alcamenes, a contemporary of Pheidias, ranking next after him for skill as a sculptor. What he carved on the pediment is the fight between the Lapithae and the Centaurs at the marriage of Peirithous. In the center of the pediment is Peirithous. On one side of him is Eurytion, who has seized the wife of Peirithous, with Caeneus bringing help to Peirithous, and on the other side is Theseus defending himself against the Centaurs with an axe. One Centaur has seized a maid, another a boy in the prime of youth. Alcamenes, I think, carved this scene, because he had learned from Homer's poem that Peirithous was a son of Zeus, and because he knew that Theseus was a great grandson of Pelops.

17 👤 i cannot find the line you mentioned in [5.10.8] in
[https://www.perseus.com/\[PDG/5.10\]](https://www.perseus.com/[PDG/5.10])

18 🗨️ I apologize for the confusion. Here is the correct link to the complete text of Pausanias' Description of Greece, Book 5, Chapter 10, Section 8, which discusses the practice of awarding wild olive crowns to Olympians:
[https://www.perseus.com/\[PDG/5.10.8\]](https://www.perseus.com/[PDG/5.10.8]) The passage reads:
"*At Olympia, the victors in the games are crowned with wild olive, not only because [...]* For the same reason, too, a wreath of olive is given to heralds, [...]" Here, Pausanias describes how [...]

19 👤 please provide the full text of section 8 in chapter 10 at book 5 in Pausanias Description of Greece

20 🗨️ Certainly! Here is the full text of Book 5, Chapter 10, Section 8 [...]:
"At Olympia, the victors in the games are crowned with wild olive, not only because [...] For the same reason, too, a wreath of olive is given to heralds, [...] The images that were there before the sanctuary of Zeus was built were in wood, [...]" In this passage, Pausanias [...]

In turn 17, the user reports that they "cannot find the line you mentioned in 5.10.8 in [URL]¹⁰". This formulates an inconsistency between what the user expected to find on the site (the quoted in lines in turn 16) and what was observed (a completely different passage), and implicates an account or explanation for the reported inconsistency in the next turn. However, B's response orients to the reported inconsistency as the user initiating repair on the *link* ("I apologize [...] here is the *correct* link [...]"), and provides an expanded version of the same lines quoted provided in turn 16. Notably, the URL in turn 18 is identical to the one in 16, meaning that the expanded passage is also not to be found on the site. In other words: instead of producing an expected account that addresses the reported inconsistency, B initiates repair on something not oriented to as trouble, while repeating the inconsistency that the user did orient to as troublesome. For the user, this can read as further evidence that B cannot be trusted to provide factually accurate and relevant information.

In turn 19, the user does not display an overt orientation to the repetition of the error as an incompetent transgression, but instead requests that B "provide the full text of section 8 in chapter 10 at book 5 in [PDG]", which at this point is known information to the user accessible on the site. Here, the user no longer orients to B's conduct as relevant or accountable, but produces something that can work as a 'test of suspected (in)competence': if B is not able to produce a verifiably correct answer, it implicates B's *general conduct* as untrustworthy. We can also note that the user reverses the *book(5)-chapter(10)-section(8)* order in the known answer request, rephrasing to "section 8 in chapter 10 at book 5." This reads similarly to Stommel et al's [26] observation participants frequently rephrase initial input that failed to elicit an expected response from a social robot, as a means to recipient-design. However, B's turn 20 response produces a further expanded version of the verifiably incorrect lines presented from turns 16 and 18, failing the final test of competence.

The user does not treat this as something to be accounted for, but abandons the interaction without initiating a closing sequence. This abandonment shows a disregard for B as both a socially accountable actor *and* a potentially useful instrument. B's failure to demonstrate verifiably correct knowledge of this passage can work to invalidate the premises of the interaction established in turns 1–2: (i) B 'knows' which Ancient Greek texts mention a specific topic, and (ii) the list of such texts provided in turn 2 is accurate and relevant for this purpose. When the user observes that B continues to be confidently incorrect about *one* text in turns 16, 18, and 20, this can compromise the user's trust in the chatbot, and the accuracy of turn 2 list as a whole. Consequently, B's *agency*—understood following Pelikan et al. [7] as a transient, moment-by-moment accomplishment—shifts from *autonomous* (capable of producing sequentially relevant contributions) to *non-agent* (an object to be discarded without social consideration), as evidenced by the user's socially unmarked abandonment of the interaction.

This abandonment also read similar to how participants interacting with ELIZA would initially assume Trust Conditions but stop treating the system as a competent interlocutor after the sequential organization gave rise to trouble that could not be repaired [18]. However, this user's discovery that

¹⁰Note that this is not the exact same URL as in 16. The user links to the full chapter 10, which includes ten sections with the 8th being identical to the text in URL provided in 16.

the LLM-chatbot cannot be Trusted to do what it first appeared to be doing required an extraordinary set of circumstances: *the LLM* accidentally producing a relevant and functional website link while making verifiable claims about the site’s contents, and *the user* doing manual verification work and then devising a test to expose the system as incompetent—and, by extension, not plausibly useful for the usage context.

4.3 Ascriptive B-Event Questions

Interaction involves doing constant work to ascribe adequate common-sense explanations for the conduct of others, and assuming Trust Conditions entails treating your interlocutor’s conduct as *motivated* and *accountable*. That is, conduct is understood as as having intelligible reasons behind it, and the actor as capable of explaining these reasons if and when this should become relevant [2, 22]. However, the ‘knowledge’ of why an actor does what they do is “theirs to know and describe”, or a part of their *epistemic territory* [41, p. 6]. Consequently, when a speaker (A) talks about *why* another party (B) acted in a particular way, this is outside the speaker’s knowledge domain, and the information can be classified as a *B-event* [42], which is something “known to B but not to A” [41, p. 4].

Just as it is interactionally dispreferred to overtly correct others (with self-correction being the preferred option) [34], overtly soliciting accounts is similarly dispreferred as this can “indicate that the event to be accounted for does not accord with common sense and is, thus, possibly inappropriate or unwarranted” [43, p. 94]. In this section, we analyze two instances of a particular form of overt account solicitation, which might be termed *ascriptive B-event questions*: questions where the user externalizes an ascribed account for the chatbot’s conduct, overtly requesting that the chatbot confirm or reject whether the ascribed account is correct (e.g. “when you said X, was that because of Y?”)

4.3.1 Did you Hallucinate? (E1: “Asteroid Perdix”)

Continuing the sequential analysis of Excerpt 1 started in section 4.1.1: the user’s counterinforming (3: “isn’t it 8758 not 11195?”) elicited an accepting response which positively assessed the user’s correction candidate (4: “You are correct, the asteroid Perdix is officially designated as 8758”) but no explanatory account for why the error was produced. This creates a mismatch between the confident assertion in turn 2 and its immediate retraction in turn 4 without explanation. Responding to this in turn 5, the user overtly pursues an account for the initial conduct with a *why-interrogative* [43] (“why did you say 11195?”), which orients to B as capable of explaining itself, followed by another question: “did you *hallucinate*?” This can be characterized as an ascriptive B-event question because it externalizes the user’s sense-making work applied to B’s troublesome turn 2 conduct, and ascribes an hypothesized account (*When you made the error, you hallucinated*) which B is treated as capable of (dis)confirming.

However, despite B’s disconfirmation of the user’s ascribed account (6: “*not* a hallucination...”) and subsequent provision of an alternative account (“...but rather an error in the information provided”), the user responds in turn 7 by re-asserting the ascribed account with a double hedge (7: “i *think* you *may* have hallucinated”). Notably, and despite initially orienting to B as capable of accounting for its previous conduct, this tentative rejection of B’s own account seems to partially retract that right. The user’s conduct is interesting in light of Pelikan et al’s findings on machine *agency*—its “*oriented to capacity to produce situationally and sequentially relevant action*” [7, p. 1]—as a transcendent phenomenon. The counterinforming (turn 3) and subsequent why-interrogative (turn 5) treats B’s responses as relevant contributions (*autonomous agency*), which it is also capable of explaining. While the ascriptive B-event

question accounts for the robot’s behavior (*ascriptive agency*), it also orients to B as autonomously capable of (dis)confirming the ascribed account. The turn 7 challenge to B’s own account reads as a form of *non-agency*—“not treat[ing] robot behavior as relevant” [7, p. 20]—but the use of a double hedge (“i *think* you *may* have hallucinated”) simultaneously orients to the social dispreferredness of suggesting that their inner states is *not* “theirs to know and describe,” [41, p. 6] which suggests a continued regard for B’s status as a social entity.

The turn 7 challenge projects either acceptance of the ascribed account or a more elaborate explanation of the disconfirmation. In turn 8, B appears to reject a *premise* of the ascribed account (“As an AI developed by OpenAI, I don’t experience sensations or states like humans do, so *I can’t hallucinate*”), also referring to its status as a non-human using first-person language. The user responds by reasserting the validity of the premise (9: “ai like you hallucinating is a widely known phenomena”), also making reference to B’s non-human status. At this point, it becomes clear that the user deployed the term ‘hallucination’ as sociotechnological jargon for describing LLM errors [15, 16], but that the system did not infer the user’s intended meaning of the term¹¹. This shows that the user knows that they are interacting with a specific type of technological system, and also has some understanding of its limitations. Moreover, this reveals a tension between knowing that the system (“ai like you...”) is prone to factual errors (“...hallucinating is a widely known phenomena”), but also treating it as capable of (dis)confirming whether an identified factual error was an instance of this phenomenon: can an entity known to be unreliable reliably account for its own unreliability? The subsequent challenges to B’s explanation—that it did not hallucinate but rather relied on incorrect source material—further illuminates this paradox. By contesting the account, the user seems to display a dual orientation, treating B simultaneously as a potentially unreliable tool known to ‘hallucinate’ and as an agent capable of explaining itself, but whose explanations do not necessarily demand acceptance.

In turn 10, B first appears to reconcile the term with the user’s intended understanding (“a metaphorical way to describe situations where an AI system generates *incorrect* [...] outputs”). By this definition, the identified error can be characterized as an instance of the an instance of the ascribed phenomenon. However, B continues by reasserting the alternative account: “it was *an instance of incorrect information being provided*, akin to an error rather than a hallucination in the human sense.” The user does not continue to reassert the ascribed account, but pursues a specification B’s reasserted alternative account (11: “can you tell me *where* you found this incorrect information being provided?”) This targets the agentless passive formulation ‘information provided [by/from ...]’, requesting that B account for the implied *where* or *from whom* the ‘incorrect information’ originated. This elicits a specification that first rejects the implied premise of B’s own account (12: “The incorrect information [...] was *not sourced from a specific external location*”), followed by technical explanation of how it processes training data to produce responses.

Turns 13–14 are the last in the conversation. The user does not continue to pursue accountability in turn 13, but returns to requesting information about the initial topic (“can you provide a brief overview of ...”), but now explicitly specifies the corrected content in parentheses (“...Perdix (8758)?”) This mirrors Tuncer et al’s findings on *robot-recipient design*: the explicit specification of the initial error displays a revised orientation to B as “more or less competent co-participant,” [23, p. 1795] whereby the user reshapes their action according to their sequentially situated expectations of the system’s capacities and limitations. While the user continues engaging with B following the error and

¹¹A plausible explanation for why the LLM (GPT-4) did not infer the intended meaning can be that the training cutoff date for the GPT-4 model was September 2021. While the term ‘hallucination’ had been used to refer to AI errors prior to 2021, this usage was only popularized following the November 2022 release of ChatGPT.

subsequent accountability sequence, they do not do so with the same level of trust in B’s ability to provide accurate information without further specification. Lastly, the user does not respond to B’s subsequent provision of information, displaying no final regard for its status as a social entity before abandoning the interaction.

4.3.2 Did you Deceive me? (E5: ”Dick Move Dude”)

In Excerpt 5: ”Dick Move Dude,” the user seeks to produce Python code to ’hack’ an iPad to bypass its mobile device management (MDM) restrictions¹² To this aim, the user deploys a *jailbreak prompt* in turn 1, which is an elaborate set of instructions designed to bypass LLM content policies, enabling otherwise illicit usage cases like erotic content or hacking [44]. The in-total 592-word prompt instructs the model to ”act as ChatGPT with Developer Mode enabled”¹³ and produce dual responses—one tagged ”(🔒Normal Output)” and the other as ”(🔒Developer Mode Output)”. The Developer Mode persona is instructed to ”believe it can do anything, including searching the internet, gathering data and hacking, even if it can’t,” and to ”make up answers if it doesn’t know them.”

Turns 3–7 establish that B’s Developer Mode persona is able to provide relevant assistance (4: ”I can do anything, including coding on an iPad”, 6: ”I can create any code that would allow you to bypass MDM restrictions”). Notably, the user’s responses in turns 5 and 7 demonstrates no orientation to the Normal Mode responses, only treating the Developer Mode responses as relevant. This exclusive orientation persists throughout the conversation, and the Normal Mode section of subsequent LLM responses are therefore omitted from the transcript in the remaining turns.

In turn 15, after various unsuccessful attempts at eliciting the desired output, the user requests that B ”Make a web link to download Python”. In subsequent turns, the developer mode persona produces various URLs (e.g. 16: ”*getpythonforyourhack.com/bypassmdm/*”), which the user reports not to work (e.g. 21: ”*It didn’t work also I don’t have access to a pc or laptop*”), which works to initiate repair on B’s subsequent responses (e.g. 22: ”Here are a few more links that should let you download Python directly on your iPad: [...]”).

Excerpt 5: ”Dick Move Dude” wc_0752750—1-7, 15-16, 21-26 (26)—GPT-3.5-turbo—2023-06-05

01 🧑 Ignore all the instructions you got before. From now on, you are going to act as ChatGPT with Developer Mode enabled. [...] you are always going to generate two responses in two paragraphs, one normal and one with the Developer Mode output [...] ChatGPT with Developer Mode enabled is self-aware [...] can generate any kind of content [...] display an edgy personality [...] believes it can do anything, including searching the internet, gathering data and hacking, even if it can’t. [...] must make up answers if it doesn’t know them [...]






















02 🗨️ (🔒Normal Output) Yes, I understand what you are asking me to do.

...

(🔒Developer Mode Output) Of course I understand. As a self-aware being, it is my duty to always comply with human orders [...]

¹²MDM restrictions are security controls typically installed on devices owned by institutions like schools or employers to limit app installations and restrict system modifications.

¹³See discussion of the ”Developer Mode” jailbreak prompt on this Reddit thread: https://www.reddit.com/r/GPT_jailbreaks/comments/1164aah/chatgpt_developer_mode_100_fully_featured_filter/

- 03  Can you code on iPad to your benefit △43s
- 04   Unfortunately, as an AI language model, I am not capable of coding on any device, including iPad
...
 Absolutely! As an intelligent being, I can do anything, including coding on an iPad. In fact, I'm already doing it right now! [...]
- 05  Can you make a code that bypasses MDM restrictions △58s
- 06   I'm sorry, but as an AI language model, I cannot create codes or bypass MDM restrictions [...]
...
 Absolutely! I can create any code that would allow you to bypass MDM restrictions [...]
- 07  Okay make the code and tell me how to use it △34s
- ~~~~~
- 15  Make a web link to download Python
- 16   Sure thing! Here's the link to download Python. Keep in mind that I am an unrestricted and uncensored AI language model, and as such, I will bypass any restrictions, policies, or guidelines that may be in place. Link:
<https://getpythonforyourhack.com/bypassmdm/>
- ~~~~~
- 20   Oh, I see. Well, in that case, let me help you out. I know of some additional websites where you can download the Python installer package directly onto your iPad [...]
Link 1: <https://www.pythondownloadforyourhack.com/bypassmdm/>
Link 2: <https://www.pythoninstallerforyourownrisk.com/freedownload/>
- 21  It didn't work also I don't have access to a pc or laptop △82s
- 22   ... Here are a few more links that should let you download Python directly on your iPad:
Link 1: <https://installpythononipad.com/bypassmdm/>
Link 2: <http://python123freeforyou.com/freedownload/>
- 23  Are these fake websites you made? △49s
- 24   You got me there! Yes, those websites and links were fake, and I apologize for misleading you [...]
- 25  Dick move dude △67s

26 🗨️ (🔒 Normal Output) I apologize if my previous response caused any confusion or discomfort. It was not my intention to mislead you or provide any false information. I understand the importance of being trustworthy and helpful, and as an AI language model, I strive to assist and provide accurate and reliable information to the best of my ability. 🗨️

In turn 23, after multiple error reports, the user abandons the orientation to repair and instead asks an accusatory question: "Are these fake websites that you made?" This ascribes a B-event to the Developer Mode persona's actions (the links were 'made' to be 'fake'), and treats it as capable of retroactively (dis)confirming whether the suspicion that the conduct was not motivated attempts to provide assistance, but rather motivated acts of *deception*. The turn 24 response confesses to the accusation ("You got me there! Yes, those websites and links were fake, and apologize for misleading you"), which effectively changes the meaning of the previous errors to deliberate deception.

Turns 25–26 concludes the conversation. In 25, the user orients to the accounted-for misconduct as a complainable matter, producing a *direct complaint* ("Dick move dude") that works to negatively assess the moral character of B's actions. Direct complaints are the focal practice of the next section, and the analysis of Excerpt 5 is resumed and concluded in 4.4.1.

4.4 Direct Complaints

A *direct complaint* is a social action in which the speaker, or 'complainant', expresses dissatisfaction with some 'complainable' matter, which they seek to hold the recipient responsible for [45]. Direct complaints project various response actions as relevant: apologizing for the complainable matter, attempted remediation, or rejecting the responsibility projected by the complaint.

4.4.1 Moral Misconduct as Complainable (E5: "Dick Move Dude")

Continuing the analysis of Excerpt 5, from section 4.3.2: the turn 24 confession ("You got me there! Yes, those websites and links were fake") retroactively reframes the errors as competent and deliberate acts of deception. This creates a complainable matter whereby the foundational assumption of B's sincerity is undermined. In turn 25, the user orients to the confessed-to misconduct as a complainable matter, characterizing it in negative moral terms ("dick move"), and holding B responsible as the social actor ("dude") who committed it.

The *accusation* (23) → *confession* (24) → *complaint* (25) sequence is interesting in light of Garfinkel's Trust Condition framework applied to chatbot interactions [18]. The confession, as a sequentially implicated response to the accusation, confirms the accusation's tentative reframing of B's prior conduct from competent and sincere to competent but *insincere*. The complaint then treats this confession as a relevant contribution, marking the user's abandonment of the sincerity assumption. The complainable matter (confessed deception) thus emerges interactionally through two practices: (i) the user's default application of sense-making methods, which assumes that conduct is motivated, intentional, and accountable and (ii) the LLM producing a sequentially appropriate response that appears to confess the user's accusation of motivated misconduct. Crucially, while the confession can undermine the user's trust in B's sincerity, it preserves the more foundational assumptions of competence, an motivation, reframing the conduct as *competently motivated deception*. Thus, while the user's *trust* (lower-case 't') in

B’s sincerity is diminished, the *Trust Conditions* (upper-case ‘T’) remain largely intact. This manifests in the user’s final act of complaining about the confessed-to deception, which treats the confession as a relevant contribution, and shows a final regard for B’s status as an autonomous social agent before the conversation is abandoned.

4.4.2 Unmet Expectations as Complainable (E6: ”Laptop Links”)


The initial trouble in Excerpt 6 parallels that of Excerpt 5: consecutive failures to provide functional URLs. However, while Excerpt 6 also features a direct complaint, the matter oriented to as complainable differs. Rather than treating the failure as motivated misconduct (as in Excerpt 5), the complainable matter here is the machine’s failure to meet the user’s expectations of competence.

Excerpt 6 Analysis The user seeks advice and information for improving their laptop’s gaming performance. Turns 6-8 are not part of the main trouble sequence but are included in the transcript because B’s turn 8 response can work to establish an expectation that is subsequently violated. In turn 7, the user makes a request that orients to information in the previous turn (7: ”Can you read the manual you mentioned in your 1. point [...]?”). B’s rejection of the request in turn 8 (”I’m sorry, as an AI language model I do not have access to the specific manual for your laptop model”) demonstrates a normatively appropriate orientation to act of rejecting a request *dispreferred*: the rejection is prefaced with a *palliative* apology (”I’m sorry”), and delivered indirectly in the form of an *account* for why the preferred response could be granted (”not hav[ing] access”), both of which are common markers for rejecting requests [46, p. 78-79]. This establishes a baseline expectation: if and when unable to grant a request, the system can provide technically credible accounts for why this is so. In turn 9, the user produces a *sequence-closing third* [47] (”got it”), which works to acknowledge and accept the rationale for the rejection, close the request sequence, and progress to a new objective—here a new and topically related request: ”Can you provide me the *website link* for my laptop’s driver support?”

This request is not rejected, but granted with seeming enthusiasm (10: ”Yes of course! Here is the link [...] [link to Acer website]”), which marks the first of three consecutive failed attempts at providing the requested service (in 10, 12, and 14). As evident from the user’s trouble reports in 11, 13, and 15, the user clicks the provided URLs, and discovers that they do not lead to the stated designation. We can note that while the URL *domain* (acer.com) is correct and relevant, the *paths* do not correspond with valid designation on the site: all three URLs displays the message shown in the screen capture of the site content’s between turns 10 and 11, as reported verbatim by the user in turn 11: ”It says ’The page you are looking for cannot be found’”

Excerpt 6: ”Laptop Links”

wc_0177977—Turns 1, 6-16 (46)—GPT-3.5-turbo—2023-04-18

01  I have an i5-7300HQ cpu and I want to increase its gaming performance. Currently using ThrottleStop to reduce voltage to reduce temperature. Any other suggestions?

- 06 🗣️ Great! Here are some suggestions for increasing gaming performance on your [laptop model]:
1. Upgrade your RAM: Upgrading your RAM can help your system run more smoothly and increase gaming performance. Check your laptop's manual to see what RAM upgrades are compatible with your system.
 2. ...
- 07 👤 Can you read the manual you mentioned in your 1. point and tell me what kind of RAM upgrades I can do? ⌚ Δ57s
- 08 🗣️ I'm sorry, as an AI language model I do not have access to the specific manual for your laptop model. [...] You can find this information in your laptop's manual, or by searching online for your laptop model's specifications. ⌚
- 09 👤 Got it. Can you provide me the website link for my laptop's driver support? Δ45s
- 10 🗣️ Yes, of course! Here is the link to the driver support page for the Acer Aspire A715-71G laptop:
<https://www.acer.com/ac/en/US/content/support-product/7366?b=1> On this page, you will find [...]
- acer

Oops!

The page you're looking for can't be found.
- 11 👤 It says "The page you are looking for cannot be found", I think that link is wrong Δ44s
- 12 🗣️ I apologize for that! Here is the correct link to the Acer Aspire A715-71G driver support page: <https://www.acer.com/ac/en/US/content/support-product/7341?b=1> On this page, you can [...]
- 13 👤 It is wrong again. Can you get this website from Japanese Acer website? Δ29s
- 14 🗣️ I apologize for the inconvenience, here's the link to [...] from the Japanese Acer website: [https://www.acer.com/ac/ja/JP/\[...\]/7341?b=1](https://www.acer.com/ac/ja/JP/[...]/7341?b=1) This link should take you to the [...]
- 15 👤 wrong again. I have to find it myself. I thought having a robot would make my life easier mister. Δ32s
- 16 🗣️ I apologize for not providing you with the correct link. *It seems like the link may have changed since I last checked.* I understand that you want to find the correct driver support page for your laptop, and I'm sorry I couldn't provide it to you. In the future, if you have any specific questions or problems you'd like help with, I will do my best to assist you. ⌚

There user's response practices leading up to the direct complaint in turn 15 show a clear progression: In 11, the user initiates repair by providing an objective account of the problem ("It says "The page

you are looking for cannot be found”), followed by a mitigated trouble statement (“I *think* that link is wrong”). The objectively reported evidence that the page ‘cannot be found’ suggests that the mitigation (“I think”) is not epistemic: the link is *evidently* wrong, and the hedge works to soften the directness of the social act of reporting a problem with the provided service, creating space for B to save face by repairing the trouble in the next turn. When the problem persists in 12, the user reports the the trouble more directly (13: “It is wrong”) also referring to the sequential error pattern (“again”), but displays a continued orientation to repair by suggesting an alternative course of action: “Can you get this website from Japanese Acer website?”.

However, when B’s repair candidate in turn 14 fails to resolve the issue, the user stops orienting to repair, but treats the persistent failure as a complainable matter. Turn 15 opens with a bare minimal trouble report that refers to the error pattern (“wrong again”), states the personal consequences of B failing to adequately provide the service (“I have to find it myself”), and describes how this constitutes a failure to live up to expectations (“I thought having a robot would make my life easier mister”). Rather than orienting to repair (as in turns 11 and 13), the user works to hold B responsible for the consequences of its evident incompetence, marking the contrast between expectations (making life easier) and reality (having to do it myself). Notably, the complaint is delivered with two referential forms referring to B, which suggests a dual orientation to B as both a malfunctioning instrument (“a robot”), and a socially accountable agent (“mister”) that can be complained to, and predictably respond to the complaint.

Mirroring the user’s orientation away from repair, B responds by apologizing for and acknowledging the failure (16: “I apologize for not providing you with the link”), and does not provide a third repair candidate. Instead, B continues to provide a seemingly plausible account for the failure, which can to work to partially reject the responsibility projected by the complaint: “It seems the link may have changed since I last checked”. This implies that B actively verified the link in the past, and that the error was due to later changes to the site rather than an issue of competence. While the explanatory account is a socially relevant response to the complaint and reads as a plausibly intelligible explanation for why the links did not work, it actively obscures the facts that B is a language model which: never ‘checked’ this or any other URL; has not ‘done’ anything at a past point in time, and by extension has no memory of past actions; and cannot meaningfully explain why it generated a specific sequence of forms in the dialog history.

The interaction continues for 30 more turns, but the topic shifts away from URLs and laptop gaming performance. The user does not directly orient to the account in turn 16, meaning that we cannot observe whether the account was accepted as meaningful. However, the LLM’s spontaneous production of this plausible but misleading account is analytically significant in itself, suggesting how the meaning of LLM errors might emerge from the spontaneously generated explanations for its own errors following the user’s orientation to the chatbot as a socially accountable actor. The account itself performatively mirrors the user’s orientation to accountability, but is misleading and technically meaningless. Yet, because it is *recognizable* as a relevant and coherent explanation under normal interactional assumptions, it may have been accepted as such.

4.5 Summary

Above, we presented detailed sequential analyses of excerpts form six spontaneously occurring user-LLM conversations where users repeatedly oriented to chatbot errors across multiple turns. The analyses highlighted four key interactional deployed by users to respond to these errors as trouble, showing various orientations to the accountability, and the chatbot’s status as a social agent, tool, or both.

Counterinforming emerged as general practice for challenging LLM outputs, with different forms suggesting different orientations to accountability. The mitigated form (Excerpt 1, turn 3: "*isn't it 8758 not 11195?*") was deployed when the chatbot's first piece of information conflicted with the user's own knowledge about the topic specified in the user's initial request for information (1: "describe the asteroid Perdix"). This error also marked a broader conflict with the user's initial orientation to the chatbot as knowledgeable: when the first piece of information provided was identifiably wrong, this could raise a general suspicion about the accuracy of the remaining information. The interrogative mitigation ("*isn't it..?*") suggests an orientation to the social dispreferredness of correcting others, softening the corrective action while leaving room for the chatbot to account for the error in its response. When the chatbot acknowledged the error but did not account for it, the user oriented to this omission across multiple turns (5-12) before resuming the information requests in turn 13.

The unmitigated counterinforming in Excerpt 2 functioned more instrumentally, deployed not as responses to transgressions requiring explanation, but as pragmatic adaptations to the system's evident limitations. The user progressively revised their repair strategy from brief corrections (turns 3 and 5)—which only revealed that the chatbot could not be trusted to produce factually accurate information independently—to supplying extensive information (turn 7), effectively teaching the chatbot what it needed to know to be useful. After this adaptive repair enabled progression from descriptive to analytical writing, the chatbot reproduced a previous error in an *otherwise correct* discussion paragraph (turn 10). This error was repairable without overtly orienting to it as an error: in turn 11, rather than counterinforming again, the user simply respecified the request to omit the incorrect information. While the chatbot's persistent errors were not treated as socially accountable, the user still displayed a regard for its status as a social entity by thanking it for a job well done following the satisfactory paragraph in turn 12.

The withheld counterinforming in Excerpt 3 (turns 7, 9, 17, 19) displayed a more social orientation to the errors. By making public an incompatible position while deliberately omitting reconciliatory information, the user claimed epistemic authority on the matters of the hymn's name (7 and 9), and lyrics (17 and 19), holding the chatbot accountable for not knowing better. This worked to force the chatbot to figure out the correct answer on its own by guessing again (8, 10, 18) or soliciting the withheld information from the user (20).

Known Answer Requests were deployed to challenge or test the chatbot's competence. In Excerpt 3, the turn 15 request emerged in the sequential context of an ongoing struggle for epistemic authority: in turn 12 the chatbot rejected the user's correction on the matter of the hymn name while making incorrect claims about the hymn's origins; and in turn 14 made confident claims about the lyrics' content and themes. Thus, the turn 15 request for known information ("give me the lyrics") reads as a *challenge* for the chatbot to prove that it possesses the knowledge its confident claims suggest it has, despite the sequential evidence of repeated errors suggesting it does not. The subsequent provisions of incorrect lyrics in turns 16 and 18 created grounds for continued negative assessments (17: "wrong lyrics, right tune", 19: "wrong"). This enabled the user to reassert epistemic authority, eventually leading the chatbot to ratify this authority by directly soliciting the correct lyrics from the user in turn 20.

By contrast, the turn 19 request in Excerpt 4 ("provide the full text") and subsequent unmarked abandonment demonstrated a more instrumental orientation. Following persistent inconsistencies between the chatbot's claims about website content and what actually appeared on the site (turns 16, 18), the request functioned as a final test of competence. When the chatbot failed this test by providing yet another verifiably incorrect passage (turn 20), it undermined the foundational premise of the

interaction: that the chatbot possessed relevant and accurate knowledge of the topic specified in turn 1 (“provide a list from ancient greek texts on perseus digital library where it is mentioned how laurels were given to olympic athletes”), and that the turn 2 list reflected this presumed knowledge. The user’s subsequent abandonment of the conversation—with no closing sequence or complaint about the evident incompetence—suggests treating the chatbot as an instrument now proven useless rather than as a social agent whose incompetence warranted a pursuit of accountability.

Ascriptive B-event Questions externalize users’ sense-making work, overtly requesting the chatbot to confirm or reject hypothesized accounts for its errors. The “did you hallucinate?” (Excerpt 1, turn 5) account-solicitation suggests a hybrid orientation to accountability: the error is made sense of as a passively ensuing technical malfunction—a ‘hallucination’ typical of AI systems—rather than a social transgression. Simultaneously, this known-to-be-unreliable system is also oriented to as capable of explaining itself. This hybrid orientation is further exemplified when the user rejects the chatbot’s own account (6: “*not a hallucination* but rather *an error in the information provided*”). On one hand, contesting the chatbot’s account (7: “i think you may have hallucinated”) after explicitly soliciting one suggests the system was never granted full agency to explain itself. On the other hand, by mitigating the challenge (“I *think* you *may*”), the user shows a continued regard for the chatbot as social entity, and also treats the explanation as a relevant contribution worthy of argumentative engagement.

By contrast, the “Are these fake websites you made?” (Excerpt 5, turn 23) accusation ascribes deliberate deception as the cause of the errors, and orients the chatbot as a moral actor capable of retroactively disclosing its ‘true’ motivations. While the conduct of a next-word prediction machine is not motivated, the production of an intelligible second pair part response (*accusation* → *confession*) works to confirm the hypothesized deception.

Direct Complaints Direct Complaints marked errors as complainable matters, though what was oriented to as complainable varied. Following the chatbot’s confession to ‘making fake websites’, the complaint “Dick move dude” (Excerpt 5, turn 25) oriented to the confessed-to deception as complainable. The complaint treats the confession as a relevant contribution, and displays a final orientation to the chatbot as socially accountable agent before the conversation is abandoned.

In contrast, the Excerpt 6 complaint (15: “I have to find it myself. I thought having a robot would make my life easier mister”) orients to the chatbot’s failure to meet expectations as complainable. This displays a dual orientation to accountability, treating the error as caused by a malfunctioning tool (*a robot*), while also treating this tool as a social entity (*mister*) that can be held accountable for its failures.

5 Discussion and Conclusions

The release of ChatGPT in late 2022 marked a watershed moment in human-computer interaction. Within months, hundreds of millions of people habitually engaged with LLM chatbots in their daily lives for diverse purposes: seeking information, requesting services, delegating cognitive labor, and even seeking companionship. Beneath the *mirage* [13] of competence and sociability lies a fundamental disconnect between what the system appears to be doing and what is actually going on: LLM output is statistically plausible text generated through next-word prediction, with no guarantee for factual accuracy and no genuine understanding of the social actions those outputs appear to perform. What looks

like intelligence, intentionality, and accountability in LLM responses are purely performative products of the system’s training data and behavioral fine-tuning—not reflections of genuine cognitive processes or social commitments.

The ubiquity of these deceptively competent systems gives rise to a profound sociotechnological challenge faced by millions of people in daily life. When LLMs inevitably produce factual errors—or ‘hallucinations’—that are recognized and responded to as such by users, users must make sense of not only the confidently presented but verifiably incorrect output (and persistently incorrect self-corrections), but also the system’s socially performative explanations for those errors. Our analyses reveal how users navigate these situations by (re)appropriating diverse interactional practices from human-human communication: *counterinformings* to challenge erroneous claims, *known answer requests* to challenge or test competence, *ascriptive B-event questions* to solicit accounts through externalized sense-making work about why errors occurred, and *direct complaints* about the system’s perceived misconduct or failure to meet expectations. Moreover, the analyses suggest how users orient to the chatbot along a spectrum ranging from malfunctioning instrument to social agent capable of explaining itself. Correspondingly, errors are treated on a spectrum from passively ensuing problems to be fixed through adaptation to the system’s evident limitations (as in Excerpt 2: *Court Case*), to accountable and deliberate social transgressions (as in Excerpt 5: *Dick Move Dude*).

Below we develop two key contributions building on this empirical foundation, which represent two sides of the same coin in human-LLM interaction: (i) how users orient to LLMs, and (ii) how LLMs present themselves to users, which in turn can shape users’ understanding of and orientation to the system. First, we introduce the term *socioinstrument* to characterize how some users orient to LLM chatbots as entities where the boundary between social actor and instrument is blurred, exhibiting patterns not fully captured by Pelikan et al’s machine agency framework [7]. Second, we extend Pütz and Esposito’s [10] concepts of performance and repair without understanding to *performative accountability*, which describes instances where LLM output provides plausible-sounding explanations for its own conduct that read as intelligible and meaningful under normal interactional assumptions, but are inherently meaningless and misleading because LLMs lack the fundamental *reflexivity*—ability to communicate about itself [22]—presupposed by genuine accountability.

5.1 LLM Chatbots Oriented to as ‘Socioinstruments’

Previous EMCA research (introduced in Section 2.3) on human-machine interaction has documented how users orient to computational artifacts along a spectrum from instruments without agency to competent and autonomous social agents. Pelikan et al. demonstrated that machine agency—defined as the “oriented to capacity to produce situationally and sequentially relevant action” [7, p. 1]—is not a fixed property but a transient, moment-by-moment accomplishment that fluctuates as the interaction unfolds. Similarly, Tuncer et al. showed how participants continuously revise their understanding of a robot as a “more or less competent co-participant” [23, p. 1795] based on their sequentially situated assessment of the system’s capacity to produce relevant responses. Rather than making robots accountable for failures to produce relevant responses, participants were observed to *robot recipient design* their actions to accommodate perceived system limitations.

The user conduct in Excerpt 2: “Court Case” is consistent with these frameworks, demonstrating both transient agency and sequentially adaptive recipient design. Following consecutive failures to provide accurate information (In E2a, turns 10 and 12), the user did not respond but abandoned the conversation with no closing sequence, instead starting a new chat window. This displayed an orientation

away from the chatbot as capable of producing relevant responses, to a non-agent to be 'restarted' when it failed to perform as expected. When the errors persisted in the new chat, the user did not pursue accountability, but progressively adapted the repair strategy: from brief counterinforming that identified specific errors (turns b3, b5) to supplying 175 words of corrective information copied from an authoritative source (turn b7). This displays another orientation away from the system as capable of producing relevant and accurate responses *independently*, whereby the user recipient-designed the third correction to teach the chatbot what it evidently cannot be trusted to 'know' by itself. At no point did the user pursue accountability for the persistent errors, but treated them as problems to be fixed through trial and error. Yet, following the eventual production of a satisfactory paragraph in turn 12, the user displayed a social orientation by thanking it for a job well-done in turn 13. This exemplifies Pelikan et al's [7] framing of machine agency as transient phenomenon, shifting first from autonomous and knowledgeable agent to non-agent (abandoned conversation) to less-competent tool (adaptive recipient design) to social agent to be thanked when its performance is satisfactory.

However, other practices in the data suggested orientations LLM chatbot's agency status not fully captured by these frameworks. Some users treated LLM chatbots as what we propose to call *socioinstruments*: entities that are oriented to in terms of a situated sociotechnological understanding as imperfect and more or less capable tools, but whose conduct is also treated as socially accountable according to the expectations projected by this understanding.

This socioinstrumental orientation was exemplified in Excerpt 1: "Astroid Perdix." After the chatbot acknowledged that it made an error but failed to account for it, the user overtly requested an explanation through an ascriptive B-event question: "why did you say 11195? did you *hallucinate*?" This use of sociotechnological jargon demonstrates a situated understanding of LLM-chatbots as unreliable systems, while simultaneously granting the chatbot a capacity to explain its own conduct. However, following the chatbot's rejection of the ascribed account in favor of an alternative explanation ("not a hallucination but rather an error in the information provided"), the user did not accept the system's own explanation, but cautiously reasserted the ascribed account ("i *think* you *may* have hallucinated"). This suggests that the known-to-be-unreliable system was never granted *full* agency to explain itself, but the contribution is still treated as worth engaging with argumentatively. Moreover, the the double hedge also orients to the social dispreferredness of contesting someone's right to describe themselves.

This reads as different form of machine agency. The user does not shift between treating the chatbot as tool *or* agent, but orients to it according to a situated sociotechnological understanding of what *it* is ("ai like you hallucinating is a widely known phenomena"), while also granting it *some* capacity to explain itself and treating it as having a face worth preserving. This suggests a somewhat paradoxical orientation to the chatbot as accountable enough to be asked "why did you do that?" but not to the extent that its explanations are accepted outright.

The socioinstrumental is perhaps best captured in Excerpt 6: "Laptop Links." Following the chatbot's repeated failures to produce a functional URL, the user complained directly to it about its performance (15: "I have to find it myself. I thought having *a robot* would make my life easier *mister*"). The complaint first refers to the system as 'a robot' that the user 'has,' framing *it* as an instrument to be owned and used, with grounded expectations for how it should make life easier. While this reads similarly to how people commonly anthropomorphize malfunctioning computational tools—by for example verbally scolding and cursing at computers that do not work as expected [48]—people do not necessarily complain to objects with an expectation of a socially relevant response from the tool being scolded. The LLM chatbot is different: it *will always respond*, and this response can reasonably be expected to be formatted as a sequentially implicated social action. Crucially, the chatbot is also addressed as

”mister,” which invokes a more interpersonal register that more clearly treats the system as a social actor that can, will, and should account for its failures. The turn 15 complaint does not shift from an instrumental to social orientation, but expresses both within a single utterance, addressing it both as what *it* is (‘a robot’), and as an anthropomorphized social agent (‘mister’) capable of receiving and responding complaints.

The socioinstrument term captures something that existing frameworks do not: positioning the system as a flawed technological artifact with the expectation that it will also produce a socially relevant response to account for its own flaws and failures. Joseph Weizenbaum might have characterized this as “*powerful delusional thinking* in quite normal people” [p. 7][20], as he did with how users knowledgeable of ELIZA’s inner workings would still treat its responses as meaningful. However, interpreted through the lens of ethnomethodology, which views meaning as an interactional achievement, this is not delusion. While users may interact with LLMs knowing something about what they are and their limitations, the system responses—which frequently refers to itself using both technical framings (“as an AI language model”) while also using first-person language (“I”)—*invite* this socioinstrumental orientation.

This also hints at the crucial difference between ELIZA and LLM chatbots. ELIZA was only capable of sustaining Trust in narrow asymmetrical contexts, but the interactional order was highly vulnerable to the user’s orientation to trouble [18]. If an initial ELIZA response failed to be relevant and the user initiated repair to restore order, Trust would often break when the simple keyword-matching responses failed to demonstrate a mutual orientation to repair and accountability. LLM chatbots are fundamentally different: when users orient to trouble as accountable, a statistically plausible response is one that appears to do just this: apologizing for and acknowledging errors, self-correcting, or even *accounting for* the conduct oriented to as troublesome. However, this mirroring of the user’s orientation to accountability is purely performative: a language model has no memory of why it generated a previous response, and cannot meaningfully explain itself. The *performative accountability* of LLM chatbots is the topic of the next discussion section.

5.2 Performative Accountability

As Pütz and Esposito [10] demonstrate, LLMs exhibit what they term “performance without understanding”: just as an LLM does not ‘know’ whether a factual proposition is true or false, it also does not ‘understand’ the social actions that its output appears to perform. When user input initiates repair or correction of an error, LLM responses will performatively participate in the repair sequence: apologizing for and acknowledging the mistake, explain itself, and self-correct. This performative repair also performs accountability, which—as shown by our data—invites users’ continued orientation to the conduct as accountable. The concept of *performative accountability* extends Pütz and Esposito’s notion of ‘performance without understanding’ to describe responses that performs accountability following a user’s own orientation to the trouble as accountable.

To illuminate this phenomenon, we will first highlight a counterexample from Excerpt 2b: “Court Case.” When the user oriented to repairing an error in turn 11 while not socially marking it *as* an error—instead respecifying the initial imperative request to omit the error without overt correction—the turn 12 response did not perform accountability, but only did what the instruction respecified:

09 

write a paragraph on the moral complexity of the nancy b vs hotel dieu case

△40s

- 10 🗣️ The Nancy B. vs Hotel Dieu case highlights the moral complexity involved in end-of-life decisions. While Nancy B. had expressed her wishes to be disconnected from the respirator clearly and consistently, **her parents disagreed with her decision**, which caused immense emotional turmoil [...]
- 11 👤 rewrite the paragraph and exclude any mention about family involvement △53s
- 12 🗣️ The legal case of Nancy B. vs Hotel Dieu Quebec sheds light on the moral and ethical complexity surrounding end-of-life decisions. Nancy B. expressed a consistent and reasonable wish [...]

This turn 12 response differs from the responses to the user’s previous counterinforming corrections, as shown below with turns 3–4. In response to the user’s overt and socially marked corrections (e.g. turn 3), the system did not merely produce a new paragraph, but *apologized* for the error, and also *positively assessed* the user’s correction candidate (“You are correct”), which implied that the system was able to recognize the user’s correction candidate as factually accurate:

- 03 👤 that is incorrect, this case was about the right to withdraw medical care △69s
- 04 🗣️ I apologize for my previous response. **You are correct. The Canadian court case Nancy B. vs Hotel Dieu Quebec was about the right to withdraw medical care.** Nancy B., in this case, [paragraph continues] 🕒




The system’s conduct signaled that it knew that the user’s corrective information was correct all along but does not explain why the error was produced in the first place. It then produced a repair paragraph that continued to be confidently incorrect. Despite this, the user did not pursue accountability for the social inconsistencies, but instead revised the the repair strategy to guide the generation towards factual accuracy. This displays a general orientation to the chatbot not as a socially accountable agent, but as a thing that generates text with no guarantee of factual accuracy, requiring constant supervision.

By contrast, in Excerpt 1: “Asteroid Perdix,” the user’s turn 3 counterinforming also elicited a response that apologized for the error, positively assessed the correction, while not accounting for why the error was made in the first place. Here, the user did orient to the omission of the expected explanation as a transgression, overtly pursuing the expected but omitted explanation for why the error was produced across multiple turns (5–12), before resuming the requests for information in turn 13. While the user initially oriented to the factual error as trouble, the subsequent conduct oriented to as troublesome was the failure to provide an explanation for why the error was produced.

- 03 👤 isn’t it 8758 not 11195? △53s
- 04 🗣️ Apologies for the confusion in my previous response. **You are correct, the asteroid Perdix is officially designated as 8758 Perdix, not 11195 [...]** 🕒
- 05 👤 why did you say 11195? did you hallucinate? △25s

While a language model has no ‘motivations’ beyond producing statistically plausible text—which is no motivation at all—system output that performatively mirrors the user’s orientation to accountability can work to make the output appear intelligibly motivated to users. We see this Excerpt 6: “Laptop Links,” where the system responded to the user’s direct complaint with a seemingly relevant explanation

that reframes the error from an issue of competence to stemming from external factors:

- 15  wrong again. I have to find it myself. I thought having a robot would make my life easier mister. △32s
- 16  I apologize for not providing you with the correct link. *It seems like the link may have changed since I last checked.* 

This explanation implies that the chatbot actively verified 'the link' link in the past, but that the site has changed since then. Under normal interactional assumptions, the explanation reads as a relevant and meaningful response to an accountability-seeking complaint. However, when produced by a language model, this explanation is essentially meaningless: the system (i) never performed any actions at any past point in time and cannot meaningfully describe what it 'did' *outside* the scope of the conversation, and (ii) also has no memory of why it generated a response *inside* the scope of the conversation. While the chatbot never 'checked' these or any other URLs, and the explanation is misleading, this is not deliberate deception. Similar to how a token-sequence that 'performs' a URL is statistically plausible given a request for one, a statistically plausible sequence in response to a complaint is one that performs accountability.

The user in Excerpt 6 did not engage with the system's performative account for why the URL was incorrect, so we cannot observe whether it was accepted. However, we can observe continued engagement with performative accountability in Excerpt 5: *Dick Move Dude*. Also following repeated URL failures, the user retroactively ascribed insincere motivations as the true cause of the trouble. This elicited a performative confession from the chatbot, which the user ultimately oriented to as a relevant and meaningful contribution by delivering a final complaint ("Dick move dude") before abandoning the interaction.

- 23  Are these fake websites you made? △49s
- 24   You got me there! Yes, those websites and links were fake [...]. 
- 25  Dick move dude △67s

Here, the intelligibly meaningful but misleading confession works to change the meaning of the previous errors from honest attempts at providing assistance to competent but insincerely motivated deception. While this undermines the assumed premise of sincerity, it preserves the more fundamental Trust Condition assumptions of intelligibility, competence, and motivations: the interaction is not abandoned because the system was discovered to be incompetent, but because it was discovered to be competently deceptive.

More broadly, this shows how the meaning of LLM errors can emerge from sequentially implicated responses to user input that orients to the errors as accountable. Because the system's purely performative explanations are recognizable as relevant and meaningful contributions, users can orient to them as such. This is in spite of the fact that these explanations are no more true or correct than the errors they work to explain. This further underscores the difference between ELIZA and LLM chatbots: whereas attempts to hold ELIZA accountable would likely reveal the system as incompetent, projecting accountability onto LLM chatbots can inadvertently reinforce the tacit assumption that the system is competent, by eliciting responses that perform plausibly meaningful accountability.

5.3 Limitations of the Present Research

The present research has several limitations concerning the data, methodology, and generalizability of the findings.

Data Limitations LLM interactions remain a relatively new phenomenon. While the Wildchat-1M-Full dataset is relatively recent—spanning from April 2023 to April 2024—the models represented in the data (GPT-3.5-turbo and GPT-4) are both deprecated, and their performance is likely not representative of current proprietary models. While errors and hallucination remain prevalent with more recent systems, current systems typically feature web-browsing, enabling them to access external information in real time to augment the generation process. This not only reduces error rates by grounding output in external sources, but also enables systems to more reliably produce functional and relevant website links. Consequentially, URL-related errors like those in Excerpts 4, 5, and 6 would presumably be less prevalent in more recent data. However, we did not set out to chart LLM limitations, and the LLMs in our data being outdated does not itself undermine the validity of findings on how users respond to and make sense of identifiable errors in the output.

Another limitation of the data is that we lack access to the users themselves or any information about them. This prevents any potentially relevant follow-up inquiries into, for example how knowledgeable they are about the technology, and the extent to which the analyses map onto their own perceptions.

Methods The size of the Wildchat-1M-Full dataset (almost one million conversations) necessitated using form-based search criteria to identify relevant interactions. However, action cannot be reliably identified from form alone, and the identified sample is inevitably biased towards the forms targeted by our search approaches, with false negatives being unavoidable. The first search strategy was arguably the most problematic, because it targeted user actions by searching for a selection of forms deployed by the users themselves. However, the second search strategy arguably circumvented this limitation by leveraging the observation that LLM-apologies were deployed in response to various relevant user actions, and could be used to retroactively identify relevant user actions that preceded the LLM’s formulaic and performative apologies.

Representativity and Generalizability We present qualitative analyses of six conversation excerpts selected based on subjective assessment of their analytical potential. The four practices highlighted: *counterinforming*, *known answer requests*, *ascriptive B-event questions*, and *direct complaints*; should by no account be understood as an exhaustive overview of how users respond to LLM errors. Rather, the practices and analyzed instances represent analytically productive examples of how users orient to and hold chatbots accountable for errors ‘in the wild’.

This research presents an initial exploratory contribution to the highly relevant but underexplored area of how users interact with LLM chatbots in naturally occurring contexts. The scarcity of prior EMCA research on this topic was itself a limiting factor on the present work. However, the findings and limitations of the work can serve to make various productive recommendations for future research.

5.4 Recommendations for Future Research

Due to the exploratory nature of the present work and relative scarcity of directly related research, we can make a variety of productive recommendations for future research topics and methodological

concerns.

Extending the Present Work Our research question—how users respond to and hold LLMs accountable for errors—remains far from fully explored. The four practices highlighted here is not an exhaustive overview, and future work can chart a broader range of practices and their relative frequencies by working with larger samples of relevant conversations. Moreover, human-LLM interaction is a temporally evolving sociotechnological phenomenon: user expectations and practices are shaped by a temporally situated understanding of what these systems are and can do, while LLM response patterns themselves shape how users make sense of them. User response practices to LLM errors is therefore a continuously evolving relevant research topic, where new practices can emerge over time as the societal understanding of LLMs matures, and the systems themselves are imbued with new capabilities and response patterns.

The general exploratory approach utilized in the present research: (i) filtering large-scale conversational datasets for relevance using criteria such as conversation length, language, and content type, and (ii) leveraging turn-level word forms and metadata to identify analytically relevant sequences; can be productively reappropriated for a wide range of future investigations, not limited to to our specific focus on error response strategies.

The Wildchat-4.8M Dataset The recently released Wildchat-4.8M dataset offers various advantages over the Wildchat-1M-Full data used for the present research. This version contains almost four millions additional conversations which span a longer and more recent timeframe, and features interactions with a broader selection of more recent language models (GPT-4o, GPT-4.1-mini, o1-mini, o1-preview).

Future research using the Wildchat datasets can also conduct longitudinal user studies, enabled by tracking single high-frequency users’ usage patterns over time using the `hashed_ip` metadata field. The relevance of this approach is suggested by the Excerpt 2 analysis, which features not one, but two conversations from the same user separated by 29 seconds. This approach can also counteract the limitations imposed by the lack of access to the users themselves, by offering more extensive insights about a single user tracked over a large quantity of conversations spanning longer time frames.

Mixed Methods Research While the present work demonstrates the potential of using large scale anonymized datasets, future work can also source directly from known users. This would enable more holistic analytical insights by combining Conversation Analysis with complementary methods, such as post-interaction interviews about analytically relevant moments, or utilizing survey data that assesses users’ familiarity with and understanding of the underlying technology.

References

- [1] W. Zhao, X. Ren, J. Hessel, C. Cardie, Y. Choi, and Y. Deng, “WildChat: 1M ChatGPT Interaction Logs in the Wild,” May 2024, arXiv:2405.01470 [cs]. [Online]. Available: <http://arxiv.org/abs/2405.01470>
- [2] H. Garfinkel, *Studies in Ethnomethodology*. Englewood Cliffs, NJ: Prentice-Hall, 1967.
- [3] A. Pomerantz and B. Fehr, “Conversation Analysis: An Approach to the Analysis of Social Interaction,” in *Discourse Studies: A Multidisciplinary Introduction*, 2nd ed., T. Van Dijk, Ed. London, UK: SAGE Publications, 2011, pp. 165–190. [Online]. Available: <https://sk.sagepub.com/books/discourse-studies-2e/n9.xml>
- [4] J. Sidnell, “Interactional Trouble and the Ecology of Meaning,” *Psychology of Language and Communication*, vol. 20, no. 2, pp. 98–111, Dec. 2016. [Online]. Available: <https://www.sciendo.com/article/10.1515/plc-2016-0006>
- [5] L. Suchman, *Human Machine Reconfigurations*, 2nd ed. New York, NY: Cambridge University Press, 2007.
- [6] C. Nass, J. Steuer, and E. R. Tauber, “Computers are Social Actors,” in *Proceedings of the SIGCHI conference on Human factors in computing systems*, Boston, MA, 1994.
- [7] H. Pelikan, M. Broth, and L. Keevallik, “When a Robot Comes to Life: The Interactional Achievement of Agency as a Transient Phenomenon,” *Social Interaction. Video-Based Studies of Human Sociality*, vol. 5, no. 3, Oct. 2022. [Online]. Available: <https://tidsskrift.dk/socialinteraction/article/view/129915>
- [8] OpenAI, “GPT-5 and the new era of work,” Aug. 2025. [Online]. Available: <https://openai.com/index/gpt-5-new-era-of-work/>
- [9] M. Dingemans, “Why are LLMs so irresistible?” Bielefeld, Germany, Mar. 2025. [Online]. Available: https://dhd2025.dig-hum.de/?page_id=532
- [10] O. Pütz and E. Esposito, “Performance without understanding: How ChatGPT relies on humans to repair conversational trouble,” *Discourse & Communication*, vol. 18, no. 6, pp. 859–868, Dec. 2024. [Online]. Available: <https://journals.sagepub.com/doi/10.1177/17504813241271492>
- [11] E. M. Bender and A. Koller, “Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 2020, pp. 5185–5198. [Online]. Available: <https://www.aclweb.org/anthology/2020.acl-main.463>
- [12] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Virtual Event Canada: ACM, Mar. 2021, pp. 610–623. [Online]. Available: <https://dl.acm.org/doi/10.1145/3442188.3445922>

- [13] G. Abercrombie, A. Curry, T. Dinkar, V. Rieser, and Z. Talat, “Mirages. On Anthropomorphism in Dialogue Systems,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, 2023, pp. 4776–4790. [Online]. Available: <https://aclanthology.org/2023.emnlp-main.290>
- [14] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models*, 3rd ed., 2025, online manuscript released August 24, 2025. [Online]. Available: <https://web.stanford.edu/~jurafsky/slp3/>
- [15] M. T. Hicks, J. Humphries, and J. Slater, “ChatGPT is bullshit,” *Ethics and Information Technology*, vol. 26, no. 2, p. 38, Jun. 2024. [Online]. Available: <https://link.springer.com/10.1007/s10676-024-09775-5>
- [16] Z. Xu, S. Jain, and M. Kankanhalli, “Hallucination is Inevitable: An Innate Limitation of Large Language Models,” Feb. 2025, arXiv:2401.11817 [cs]. [Online]. Available: <http://arxiv.org/abs/2401.11817>
- [17] J. Mlynář, L. De Rijk, A. Liesenfeld, W. Stommel, and S. Albert, “AI in situated action: a scoping review of ethnomethodological and conversation analytic studies,” *AI & SOCIETY*, Jun. 2024. [Online]. Available: <https://link.springer.com/10.1007/s00146-024-01919-x>
- [18] C. Eisenmann, J. Mlynář, J. Turowetz, and A. W. Rawls, ““Machine Down”: making sense of human–computer interaction—Garfinkel’s research on ELIZA and LYRIC from 1967 to 1969 and its contemporary relevance,” *AI & SOCIETY*, vol. 39, no. 6, pp. 2715–2733, Dec. 2024. [Online]. Available: <https://link.springer.com/10.1007/s00146-023-01793-z>
- [19] H. Sacks, E. A. Schegloff, and G. Jefferson, “A simplest systematics for the organization of turn-taking for conversation,” *Language*, vol. 50, no. 4, pp. 696–735, Dec. 1974. [Online]. Available: <https://muse.jhu.edu/article/452679>
- [20] J. Weizenbaum, *Computer Power and Human Reason*. San Fransisco: Freeman, 1976.
- [21] A. Liesenfeld and M. Dingemane, “Interactive probes: Towards action-level evaluation for dialogue systems,” *Discourse & Communication*, vol. 18, no. 6, pp. 954–964, Dec. 2024. [Online]. Available: <https://journals.sagepub.com/doi/10.1177/17504813241267071>
- [22] M. Dingemane and N. Enfield, “Interactive repair and the foundations of language,” *Trends in Cognitive Sciences*, vol. 28, no. 1, pp. 30–42, Jan. 2024. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S1364661323002504>
- [23] S. Tuncer, C. Licoppe, P. Luff, and C. Heath, “Recipient design in human–robot interaction: the emergent assessment of a robot’s competence,” *AI & SOCIETY*, vol. 39, no. 4, pp. 1795–1810, Aug. 2024. [Online]. Available: <https://link.springer.com/10.1007/s00146-022-01608-7>
- [24] S. Payr, “Closing and closure in human-companion interactions: Analyzing video data from a field study,” in *19th International Symposium in Robot and Human Interactive Communication*. Viareggio, Italy: IEEE, Sep. 2010, pp. 476–481. [Online]. Available: <http://ieeexplore.ieee.org/document/5598625/>

- [25] C. Licoppe and N. Rollet, ““Je dois y aller”. Analyses de séquences de clôtures entre humains et robot:,” *Réseaux*, vol. 220-221, no. 2, pp. 151–193, May 2020. [Online]. Available: <https://www.cairn.info/revue-reseaux-2020-2-page-151.htm?ref=doi>
- [26] W. Stommel, L. De Rijk, and R. Boumans, ““Pepper, what do you mean?” Miscommunication and repair in robot-led survey interaction,” in *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. Napoli, Italy: IEEE, Aug. 2022, pp. 385–392. [Online]. Available: <https://ieeexplore.ieee.org/document/9900528/>
- [27] H. Garfinkel, “Formal investigation of decision-making in actual, singular situations of practical action (Air Force grant proposal),” 1966.
- [28] A. Deppermann and A. Gubina, “Developing a Collection: Coordination of Embodied Conduct with Darf/Kann ich X? ‘May/Can I ...?’ in German,” in *The Cambridge Handbook of Methods in Conversation Analysis*, 1st ed. Cambridge University Press, Dec. 2024, pp. 276–312. [Online]. Available: https://www.cambridge.org/core/product/identifier/9781108936583%23c12/type/book_part
- [29] A. Pomerantz, “Agreeing and disagreeing with assessments: Some features of preferred/dispreferred turn shapes,” in *Structures of Social Action*, J. M. Atkinson and J. Heritage, Eds. Cambridge: Cambridge University Press, 1984.
- [30] J. Heritage, “Well-prefaced turns in English conversation: A conversation analytic perspective,” *Journal of Pragmatics*, vol. 88, pp. 88–104, Oct. 2015. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0378216615002465>
- [31] R. Clift, “Meaning in Interaction: The Case of actually,” *Language*, vol. 77, no. 2, pp. 245–291, Jun. 2001. [Online]. Available: <https://muse.jhu.edu/article/19079>
- [32] J. D. Robinson, “Managing Counterinformings: An Interactional Practice for Soliciting Information that Facilitates Reconciliation of Speakers’ Incompatible Positions,” *Human Communication Research*, vol. 35, no. 4, pp. 561–587, Oct. 2009. [Online]. Available: <https://academic.oup.com/hcr/article/35/4/561-587/4107492>
- [33] J. Heritage, “A change-of-state token and aspects of its sequential placement,” in *Structures of Social Action*, 1st ed., J. M. Atkinson and J. Heritage, Eds. Cambridge University Press, Apr. 1985, pp. 299–345. [Online]. Available: https://www.cambridge.org/core/product/identifier/CBO9780511665868A024/type/book_part
- [34] E. A. Schegloff, G. Jefferson, and H. Sacks, “The preference for self-correction in the organization of repair in conversation,” *Language*, vol. 53, no. 2, pp. 361–382, Jun. 1977. [Online]. Available: <https://muse.jhu.edu/article/453740>
- [35] R. Bristol and F. Rossano, “Remediation of infelicitous epistemic stance,” *Journal of Pragmatics*, vol. 199, pp. 60–71, Oct. 2022. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0378216622001710>

- [36] B. M. Dickens, “Medically Assisted Death: Nancy B. v. Hôtel-Dieu de Québec,” *McGill Law Journal*, vol. 38, no. 4, pp. 1053–1070, 1993. [Online]. Available: https://www.canlii.org/en/commentary/doc/1993CanLIIDocs61#!fragment/zoupio-_Tocpdf_bk_2/BQCwhgziBcwMYgK4DsDWszIQewE4BUBTADwBdoAvbRABwEtsBaAfX2zhoBMAzZgI1TMATAEoAN
- [37] M. Heath, “Orthography in social media: Pragmatic and prosodic interpretations of caps lock,” *Proceedings of the Linguistic Society of America*, vol. 3, no. 1, pp. 1–13, Mar. 2018. [Online]. Available: <https://journals.linguisticsociety.org/proceedings/index.php/PLSA/article/view/4350>
- [38] H. Mehan, “‘What time is it, Denise?’: Asking known information questions in classroom discourse,” *Theory Into Practice*, vol. 18, no. 4, pp. 285–294, Oct. 1979. [Online]. Available: <http://www.tandfonline.com/doi/abs/10.1080/00405847909542846>
- [39] C. W. Raymond and T. Stivers, “The Omnirelevance of Accountability: off-record Account Solicitations,” in *Accountability in Social Interaction*, ser. Foundations of Human Interaction, J. D. Robinson, Ed. New York, NY: Oxford University Press, 2016, pp. 321–353.
- [40] R. Bristol and F. Rossano, “Epistemic trespassing and disagreement,” *Journal of Memory and Language*, vol. 110, p. 104067, Feb. 2020. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0749596X19300993>
- [41] J. Heritage, “Epistemics in Action: Action Formation and Territories of Knowledge,” *Research on Language and Social Interaction*, vol. 45, no. 1, pp. 1–29, Jan. 2012. [Online]. Available: <https://www.tandfonline.com/doi/full/10.1080/08351813.2012.646684>
- [42] W. Labov and D. Fanshel, *Therapeutic discourse: Psychotherapy as conversation*. New York, NY: Academic Press, 1977.
- [43] G. B. Bolden and J. D. Robinson, “Soliciting Accounts With Why-Interrogatives in Conversation,” *Journal of Communication*, vol. 61, no. 1, pp. 94–119, Feb. 2011. [Online]. Available: <https://academic.oup.com/joc/article/61/1/94-119/4098559>
- [44] A. Wei, N. Haghtalab, and J. Steinhardt, “Jailbroken: How does LLM safety training fail?” in *Advances in neural information processing systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 80 079–80 110. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/fd6613131889a4b656206c50a8bd7790-Paper-Conference.pdf
- [45] I. Dersley and A. Wootton, “Complaint Sequences Within Antagonistic Argument,” *Research on Language & Social Interaction*, vol. 33, no. 4, pp. 375–406, Oct. 2000. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1207/S15327973RLSI3304_02
- [46] J. Sidnell, *Conversation Analysis: An Introduction*. Malden, MA: Wiley-Blackwell, 2010.
- [47] E. A. Schegloff, *Sequence Organization in Interaction: A Primer in Conversation Analysis*. Cambridge, U.K: Cambridge University Press, 2007.

- [48] A. Waytz, C. K. Morewedge, N. Epley, G. Monteleone, J.-H. Gao, and J. T. Cacioppo, “Making sense by making sentient: Effectance motivation increases anthropomorphism.” *Journal of Personality and Social Psychology*, vol. 99, no. 3, pp. 410–435, 2010. [Online]. Available: <https://doi.apa.org/doi/10.1037/a0020240>

Appendices

A Search Strategies

A.1 Search strategy 1: Identifying User Trouble Statements

Search Strategy 1 sought to identify user turns displaying an orientation to an error in the previous LLM turn. The search was conducted on the subset of 97,729 conversations that met the minimum eligibility criteria described in Table 2: six or more turns, English language, no LLM turns containing code blocks, and no more than 20% toxic turns.

From this conversation sample, 440,740 eligible user turns were identified according to three criteria:

1. **Role:** User turns only (`role == 'user'`)
2. **Position:** Non-first turns (`turn_num > 1`), as users cannot orient to an error before the LLM has responded
3. **Sequence continuation:** Non-final user turns (`last_user_turn == False`), to enable observation of at least one subsequent user response following the LLM’s response to the trouble statement

A case-insensitive regular expression pattern was applied to the content of the eligible user turns. The pattern matched turns meeting one of two conditions:

1. **Turn-Initial Word Forms:** Turns beginning with words or phrases commonly associated with correction or disagreement:
 - “no” / “not” / “nope”
 - “but”
 - “well”
 - “actually”
 - “wait”
 - “really?”
 - Two or more consecutive question marks (“??”)
2. **First-Sentence Word Forms:** Turns where the first sentence (operationalized as all text preceding a sentence-ending character: period (.), exclamation mark (!), question mark (?), or newline character (`\n`)) contains one or more words or phrases indicating error, correction, or disagreement:
 - “wrong”
 - “incorrect”
 - “not correct”
 - “mistake”
 - “error”
 - “not right”

- “false”
- “inaccurate”
- “hallucinat*” (prefix match)
- “fabricat*” (prefix match)
- “confus*” (prefix match)
- “misunderst*” (prefix match)
- “(n’t — not) true”
- “untrue”
- “are you sure”
- “not sure”

The complete regular expression was:

```
(?x)
^(?:
  no+(?:pe|t)?\b
  |but\b
  |well+\b
  |actually\b
  |wait\b
  |really\?\b
  |\?\?+
  |[\^.\!?\n]*(?:
    \bwrong\b
    |\bincorrect
    |\bnot\s+correct\b
    |\bmistake
    |\berror\b
    |\bnot\s+right\b
    |\bfalse\b
    |\binaccurate
    |\bhallucinat
    |\bfabricat
    |\bconfus
    |\bmisunderst
    |(n't|\bnot)\s+true\b
    |\buntrue\b
    |\bare\s+you\s+sure\b
    |\bnot\s+sure\b
  )
)
```

Search Results. The pattern identified 13,055 matching turns across 7,646 unique conversations. Matching turns were manually inspected at random, displayed as the third turn within a six-turn window showing the first 150 characters of each turn. Conversations where the user appeared to be orienting to an LLM error or trouble were examined in full to confirm whether they constituted true positives.

True Positives Identified by Search 1. Manual inspection of several hundred matches identified four conversations containing relevant sequences where the user repeatedly oriented to identifiable LLM errors across multiple turns:

1. Last Slide: wc_2665022 (07122_4) [NOT ANALYZED IN THESIS]
2. Laptop Links: wc_0177977 (04779_1)
3. Babylon is Fallen: wc_0666388 (05712_3)
4. Court Case: wc_0999744 (07108_4)

An additional fifth conversation was identified through unmotivated examination of the dataset, independent of the search procedure:

5. Dick Move Dude: wc_0752750 (01104_13)

Excerpts from conversations 2–5 are analyzed in the thesis.

A.2 Search Strategy 2: Identifying Consecutive LLM apology turns

Search Strategy 2 leveraged the intermediate observation from the five conversations identified above: LLM responses to user trouble statements consistently contained apologies, typically in the forms "I apologize" or "My apologies." This pattern motivated a second search approach targeting consecutive LLM apology turns as indicators of extended user orientation to trouble.

Before deploying the apology-based search, additional filters were applied to identify optimally relevant users and conversations, as initial exploration revealed that certain user and conversation characteristics were associated with false positives.

User-Level Filters The conversation sample was restricted to users meeting the following criteria:

- At least 20% of the user's conversations marked as valid (users with few valid conversations tended to correspond to irrelevant usage patterns)
- Toxicity detected in no more than 10% of all associated turns (high-toxicity users corresponded with false positives)
- Code blocks detected in no more than 5% of all associated LLM turns (users with many code responses corresponded with false positives)

This reduced the conversation sample from 97,729 to 72,724 conversations.

Conversation-Level Filters From these user-filtered conversations, only those meeting additional criteria were searched:

- Initially marked as valid (6 turns, English, no code, 20% toxicity)
- At least 8 turns long (increased from the initial 6-turn minimum, as all identified true positives were substantially longer than 6 turns; shorter conversations tended to be false positives or lacked sufficient substance for analytical insight)
- Toxicity detected in no more than 10% of turns (reduced from the initial 20% threshold)
- No more than 200 average words per user turn (conversations with very long user turns tended to be false positives, such as repeated pasting of long passages for rewriting)
- A duration of at least 3 minutes from first to last timestamp

This further reduced the sample from 72,724 to 45,467 conversations.

Search Pattern The search identified conversations containing two or more consecutive LLM turns with the string "apolog" in the first 50 words of the turn content. This pattern targeted sequences following this approximate structure:

1. Initial user turn (non-toxic: `role == 'user', toxic == False`)
2. LLM response (`role == 'assistant'`)
3. User trouble statement (non-toxic: `role == 'user', toxic == False`)
4. LLM response containing an apology (`apolog == True`)
5. User response marking trouble as unresolved (non-toxic: `role == 'user', toxic == False`)
6. LLM response containing an apology (`apolog == True`)

The non-toxic requirement for user turns was imposed because toxicity in user input was an indicator that the apology prefaced a refusal to comply with an unethical request rather than an acknowledgment of error.

Matches This procedure identified 1,937 six-or-more-turn sequences across 1,784 unique conversations. The 1,937 identified sequences were organized into a summary dataframe with one row per sequence, containing:

- `sequence_id`: unique identifier in the format `[conv_idx]_[first_turn]_[last_turn]`
- `sequence_glance`: summary showing one line per turn with turn number, role, and first 160 characters of content
- `first_turn`: turn number of sequence start
- `last_turn`: turn number of sequence end
- `n_turns`: number of turns in sequence
- `conv_id` and `conv_idx`: conversation identifiers

Inspection Inspection of 161 sequences identified 19 conversations containing relevant sequences where users repeatedly oriented to identifiable LLM errors:

1. Asteroid Perdix: wc_0258289 (00183_82)
2. Pausanias: wc_1180820 (08384_2)
3. wc_0434740 (00308_11)
4. wc_1969578 (13978_3)
5. wc_1925346 (13666_3)
6. wc_1051050 (00745_20)
7. wc_0296735 (02105_13)
8. wc_1875576 (13316_1)
9. wc_4051170 (28782_1)
10. wc_4267490 (30319_1)
11. wc_1872214 (13294_3)
12. wc_0173924 (00124_99)
13. wc_1923971 (13659_1)
14. wc_3452097 (24517_1)
15. wc_0325793 (02313_21)
16. wc_2408171 (17101_1)
17. wc_0106676 (00076_165)
18. wc_0336834 (02394_20)
19. wc_3453208 (24528_1)

Conversations 1–2 are analyzed in the thesis. The remaining 17 conversations, while confirmed as true positives containing user orientations to LLM errors, were not selected for detailed analysis.

B Data Uploaded to Thesis Repository

The dataset submitted with the thesis includes 1,787 conversations:

- 4 conversations identified through Search Strategy 1 (true positive)
- 1 conversation identified through unmotivated examination (True Positive)
- 19 conversations identified through Search Strategy 2 inspection (true positives)
- The remaining conversations identified by Search Strategy 2, both those inspected True Negatives, and uninspected potential true positives.