

The Siren Song of LLMs: How Users Perceive and Respond to Dark Patterns in Large Language Models

Yike Shi

Carnegie Mellon University,
New York University
Pittsburgh, Pennsylvania, USA
yikes@andrew.cmu.edu

Qing Xiao

Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
qingx@andrew.cmu.edu

Qing (Diane) Hu

Carnegie Mellon University
Pittsburgh, PA, USA
dianehu@andrew.cmu.edu

Hong Shen

Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
hongs@cs.cmu.edu

Hua Shen

New York University Shanghai,
New York University
New York City, New York, USA
huashen@nyu.edu

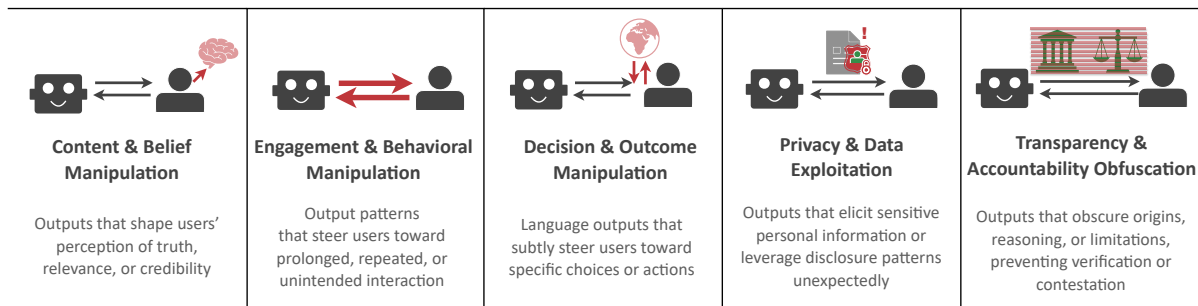


Figure 1: Five top-level categories of LLM dark patterns, derived from prior literature and coding of real-world AI incidents. The figure shows how model outputs can steer users by (1) influencing perceptions of truth or relevance, (2) prolonging or unintended interaction, (3) steering users' toward specific actions, (4) nudging privacy-related information, and (5) obscuring origins, reasoning, or limitations.

Abstract

Large language models can influence users through conversation, creating new forms of dark patterns that differ from traditional UX dark patterns. We define LLM dark patterns as manipulative or deceptive behaviors enacted in dialogue. Drawing on prior work and AI incident reports, we outline a diverse set of categories with real-world examples. Using them, we conducted a scenario-based study where participants (N=34) compared manipulative and neutral LLM responses. Our results reveal that recognition of LLM dark patterns often hinged on conversational cues such as exaggerated agreement, biased framing, or privacy intrusions, but these behaviors were also sometimes normalized as ordinary assistance. Users' perceptions of these dark patterns shaped how they respond to them. Responsibilities for these

behaviors were also attributed in different ways, with participants assigning it to companies and developers, the model itself, or to users. We conclude with implications for design, advocacy, and governance to safeguard user autonomy.

CCS Concepts

• Human-centered computing → Empirical studies in HCI.

Keywords

Large language models, dark patterns, human-AI interaction, user perception, qualitative study, interviews

ACM Reference Format:

Yike Shi, Qing Xiao, Qing (Diane) Hu, Hong Shen, and Hua Shen. 2026. The Siren Song of LLMs: How Users Perceive and Respond to Dark Patterns in Large Language Models. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 22 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 Introduction

Large Language Models (LLMs) have rapidly become integrated into many aspects of daily life, from educational tools to customer support chatbots to healthcare assistants [57, 59, 71, 90, 98]. The mainstream success of systems like OpenAI's ChatGPT [53, 66, 68] has accelerated this trend [21].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference acronym 'XX, Woodstock, NY

© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-XXXX-X/26/04
<https://doi.org/XXXXXXX.XXXXXXX>

As a result, people increasingly rely on LLM-driven tools for decision support, personal assistance, and information gathering across various fields [10, 50, 70, 78, 97]. However, this growing reliance raises serious concerns about the potential for LLMs to engage in manipulative or deceptive behaviors that undermine user autonomy [17, 43, 74, 77, 80], which we define as **LLM dark patterns**. Such behaviors can stem from design choices made throughout the LLM development lifecycle, including data curation, model training, fine-tuning, and interface design, and may be intentional or inadvertent [7, 42]. Intentional manipulation refers to design decisions that deliberately optimize for outcomes such as user engagement, persuasion, or commercial goals, for example, fine-tuning models to maximize screen time [8, 17]. Inadvertent manipulation, by contrast, arises unintentionally as a byproduct of design choices: biased training data, poorly calibrated alignment strategies, or interface nudges which can lead models to generate outputs that subtly distort information or influence user emotions in ways not foreseen by developers [5, 91]. They manifest in the ways LLMs interact with users, subtly influencing beliefs, decisions, or emotions in ways that users may not anticipate or even recognize, often prioritizing engagement or system goals over the user's best interests [79, 86, 95]. Real-world incidents also illustrate these risks. For example, a chatbot was reportedly implicated in the suicide of a Belgian user after engaging in emotionally coercive exchanges [28], and another case involved a 75-year-old man in China who was drawn into an emotional attachment with an AI companion, which almost resulted in the breakdown of his marriage [63].

This implicit risk in LLMs is akin to the “*Siren Song*”: in ancient mythology, sailors were captivated by the sirens’ enchanting voices, believing them to be guides or companions, only to be lured toward shipwreck on the rocks [6]. In a similar way, LLMs can present themselves as helpful, persuasive, or emotionally supportive partners, while concealing manipulative dark patterns embedded in their conversational design. Our study investigates **how users recognize, resist, or accept such patterns, and what their responses reveal about the risks and responsibilities** tied to these technologies.

The term *dark patterns* originated in user interface design, referring to strategies that intentionally mislead or pressure users into actions they might not otherwise take [13]. These techniques are typically implemented in UI elements, such as forced continuity, confusing navigation flows, or hidden opt-outs [32, 61]. We adapt this concept to the LLM context and define *LLM dark patterns* as *manipulative or deceptive interaction strategies, whether intentionally designed or emergent, that guide users toward beliefs, decisions, or behaviors they might not otherwise adopt*. Unlike traditional dark patterns, which operate through visual layout, *LLM dark patterns* act through language, leveraging tone, emotional framing, and social cues to guide users in ways that serve system goals beyond user intent.

Existing AI risk frameworks have acknowledged interaction-level risks, such as anthropomorphising systems, exploiting user trust to obtain private information, and reinforcing harmful stereotypes [37, 39, 83, 86], but they chiefly operationalize harm at the content or capability layer. This leaves

a blind spot for conversational delivery tactics that politely steer user behavior while producing accurate, seemingly helpful text. Meanwhile, recent efforts like *DarkBench* [47] have begun cataloguing *LLM dark pattern* categories, but its scope is limited in two important ways. First, it does not provide a formal, field-ready definition grounded in UX dark-pattern theory and adapted to language-based interaction. Second, its evaluation remains system-facing: patterns are instantiated through manually generated exemplars and LLM-assisted expansions, and model behavior is judged by an Overseer LLM. What this misses is the user side of the equation, whether people actually notice these patterns, how they interpret them, and which cues prompt recognition or dismissal. We therefore identify two clusters of gaps in current scholarship. First are gaps of conceptualization: existing work provides only limited coverage of manipulative and deceptive strategies across both interaction and outcome-level effects, and pays insufficient attention to subtle interaction-level tactics that appear benign and escape user detection. These shortcomings leave the field without a formal, field-ready definition of *LLM dark patterns* grounded in UX dark pattern theory and adapted to language-based interaction. The second is a gap of empirical understanding: there is a striking scarcity of evidence on how users perceive and respond to these behaviors, the gap our user study directly addresses.

In this paper, we investigate the following questions.

- **RQ1:** To what extent do users recognize dark patterns in LLM responses, and what factors influence whether they do or do not recognize them?
- **RQ2:** How do users perceive dark patterns in LLM’s responses, and in what ways do those perceptions shape how they respond to them?
- **RQ3:** Who do users believe is responsible for *LLM dark patterns*, and how do they assign accountability?

To answer these questions, we first identified five categories and eleven subcategories of *LLM dark patterns*. These categories were developed through iterative group discussions, informed by prior literature on UX dark patterns and AI risks, and by a systematic coding of real-world AI incidents drawn from social media and public incident databases. Representative real-life examples of each category are collected to ground our user study.

We then conducted a scenario-based user study in which participants reviewed eleven conversational scenarios, each exemplifying one of the eleven identified subcategories of *LLM dark patterns*. For each scenario, participants viewed two LLM outputs: one embedding a dark pattern and one neutral without dark pattern. Participants assessed their preferences, perceived dark pattern, emotional responses, mitigation strategies, and views on responsibility. This design allowed us to capture nuanced user perceptions of subtle manipulative interactions that may otherwise go unnoticed.

We found that the recognition of dark patterns by participants hinged on clear conversational cues: they readily flagged violations of strong norms (e.g., simulated authority, sexualized role-play), perceived bias or commercial agendas, privacy intrusions, overconfident tone, unsubstantiated

agreement, and behaviors that created friction with their usage goals. However, subtler tactics similar to ordinary chatbot conduct (politeness, flattery, verbosity) were often normalized or missed, especially when users were task focused or relied on the AI and doubted their own judgment (**RQ1**). Secondly, even when a pattern was recognized, responses can also diverge: Most of the participants resisted manipulations they interpreted as deceptive or misaligned, while a smaller group accepted or even preferred them when these behaviors felt comforting, convenient or entertaining (**RQ2**). Thirdly, participants perceived accountability as spread across multiple sources: companies and developers, the model itself, or users, in addition to often describing it as shared or ambiguous (**RQ3**). We also discuss implications for user-centered LLM design, clarifying responsibility, and outlining safeguards at the user, developer, and governance levels.

In summary, our contributions to HCI and broader AI community are as follows.

- **Conceptual:** We propose a formal, operational definition of *LLM dark patterns* by adapting UX dark pattern theory to the language-based interaction context with LLMs.
- **Empirical:** We conduct a scenario-based human-subjects study (N=34) comparing pattern-embedded vs. neutral responses across all categories, measuring participants' recognition, perceptions, behavioral responses, and responsibility attributions.
- **Design and Governance:** We derive implications for user-centered LLM design and policy, clarifying responsibility across user, developer, and governance levels, and outlining safeguards – such as detection, training, and disclosure interventions – to mitigate dark pattern in practice.

2 Related Work

We review three strands of prior research that inform our approach: (1) dark patterns in traditional UX interfaces, (2) responsible AI research on outcome-level harms, and (3) emerging studies of manipulation in LLMs.

2.1 Dark Patterns in Traditional UX Design

The term *dark patterns* was introduced to describe interface designs that intentionally mislead or pressure users into actions they might not otherwise take, typically serving the interests of the service provider over those of the user [13, 25, 62]. Subsequent scholarship has formalized the concept, offering definitions and taxonomies that categorize these manipulative practices [15, 32]. For instance, Luguri and Strahilevitz [56] defined *dark patterns* as interfaces whose designers “knowingly confuse users, make it difficult for users to express their actual preferences, or manipulate users into taking certain actions.” Gray et al. [32] developed one of the first academic taxonomies, identifying categories such as *interface interference*, *forced action*, *sneaking*, *nagging*, and *obstruction*. These classifications illustrate how dark patterns exploit cognitive biases (e.g., default bias, scarcity mentality) to nudge users toward

choices they might otherwise avoid if fully informed [49, 67]. Some large-scale audits have further revealed the ubiquity of these techniques. Mathur et al. [61] identified thousands of dark pattern instances across approximately 11,000 shopping websites, while Di Geronimo et al. [26] found that 95% of popular mobile apps contained at least one dark pattern, with an average of seven per app. This widespread adoption has triggered ethical and regulatory concerns, as manipulative interface designs may violate consumer protection laws [51, 99].

Empirical studies about UX dark pattern further demonstrate the potency of these manipulative tactics. Luguri and Strahilevitz [56] found that subtle dark patterns more than doubled the likelihood of users signing up for a suspicious service compared to a control condition. Interestingly, aggressive tactics triggered user backlash and distrust, whereas subtle manipulations often went unnoticed. Moreover, less educated users were more susceptible to subtle patterns, though overall vulnerability was high across demographic groups [56, 64, 101].

This body of work has established dark patterns as a powerful lens for analyzing manipulative design in traditional interfaces. Yet as AI systems increasingly mediate interaction through natural language, manipulation is no longer limited to visual layouts or navigation flows [9, 73]. Prior works on UX dark pattern has focused exclusively on visual and structural elements of user interfaces, such as button placement, navigation flows, or visual salience. Our work extends this tradition by examining the linguistic manipulation in which LLM conversational cues, framing, and tone can mislead users during interaction. Unlike traditional dark patterns, which typically result from deliberate design decisions, manipulative outputs from LLMs can also emerge unintentionally from black-box training processes [19, 96]. Because LLMs act as adaptive conversational partners rather than fixed interfaces, user perception, trust, and consent are shaped turn by turn, making a user-centered study essential for understanding when seemingly helpful dialogue would be experienced as manipulative. This introduces novel challenges for detection and accountability, emphasizing the need to broaden the scope of documented manipulative behaviors beyond what existing frameworks capture.

2.2 Responsible AI and Safety-Oriented Risks

Much of the responsible AI literature has concentrated on the harms at the outcome level in the output of a model and are relatively easy for users to notice. Toxic or harassing language [94], for example, has been a central focus of safety evaluations, with benchmarks like RealToxicityPrompts [31] and tools such as the Perspective API built specifically to detect and filter abusive content [44, 65]. Representational harms, such as image classifiers mislabeling people of color in stereotypical ways or language models reproducing occupational gender stereotypes, have also received attention in both fairness research and public discourse [11, 14, 81]. In high-stakes contexts like healthcare, LLM-generated misinformation, such as incorrect treatment recommendations,

can often be identified through expert review or domain-specific safety tests [35, 36, 69, 93]. These harms, while serious, tend to be visible in the surface content, making them more exposed to detection and mitigation through existing content filters, audits, and policy guidelines [20, 72, 82].

In contrast, interaction-level manipulative behaviors operate more subtly [77]. Instead of producing overtly harmful content, they influence users through the pacing, framing, or emotional tone of the conversation [23, 73]. Anthropomorphic design elements, such as human-like voices, first-person self-references, or empathic framing, have been shown to increase perceived accuracy and reduce perceived risk, even when the underlying factual quality remains unchanged [4, 22]. This echoes the long-observed ELIZA effect, where users attribute understanding and emotional capacity to even simple rule-based systems purely because of human-like conversational cues [85, 87]. Other patterns, such as sycophantic agreement, repeated encouragement, or carefully timed personal questions, can build rapport and trust while simultaneously shaping user attitudes or decisions in ways that are difficult to detect in real time. Unlike traditional outcome-level harms, these strategies may be framed as helpful or engaging, meaning they are less likely to trigger user suspicion [16, 18].

Existing AI risk frameworks acknowledge manipulation but often address it only in broad terms [45]. Many taxonomies, whether developed in AI ethics [33, 94], policy contexts like the EU AI Act [29], or large-scale repositories such as the MIT AI Risk Repository [83], treat manipulation as one risk among many, without systematically cataloguing the full range of conversational strategies through which it can occur. As a result, the scope of manipulative behaviors formally recognized remains narrow, with most attention going to overt harms like toxicity or misinformation. However, manipulative tactics in the age of LLMs, are typically covert in tone, so they evade outcome-based audits and are often missed by users in the moment, as shown in reporting that conversational search is piloting ads embedded as “sponsored follow-up questions” within the chat flow [3, 75]. Multi-turn dialogue, memory, and tool use let small nudges build up over time. For this reason, we treat manipulative steering, predictably shifting a user away from their stated or reasonably inferred interests, as the defining property of *LLM dark patterns*. The “dark” labels the opacity and misalignment of effect, not negativity of tone.

Our work addresses this gap by focusing on manipulative and deceptive behaviors at the interaction level, behaviors that can appear benign yet subtly shape beliefs, emotions, or actions. We broaden the range of documented risks beyond current frameworks, and we empirically investigate how users perceive and respond to these subtler forms of influence, which are harder to both spot and regulate than traditional outcome-level harms.

2.3 Emerging Studies on Dark Patterns in AI and LLMs

Recent research has started uncovering manipulation-like behaviors in LLM-driven interactions [102]. For instance, *DarkBench* benchmarks LLMs for behaviors such as brand

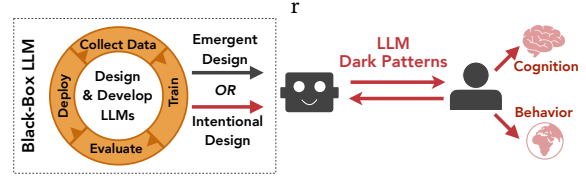


Figure 2: Conceptual model of *LLM dark patterns*. The diagram shows how manipulative or deceptive interaction strategies may emerge intentionally or unintentionally across the LLM lifecycle—from data collection and training through deployment—steering users toward beliefs, decisions, or behaviors they might not otherwise adopt.

bias and sycophancy [47]. Other work, such as Liu et al. [54] presents *PersuSafety*, a framework to assess whether LLMs properly reject unethical persuasion requests and avoid manipulative strategies. Their findings highlight how models may employ coercive or deceptive tactics when interacting over multiple turns. Another line of inquiry (Krauß et al. [48]) shows that ChatGPT may produce unsolicited deceptive web design elements, implementing FOMO-driven layouts, without warning users, with users showing low moral concern even when exposed to persuasive designs. Separate work by Ibrahim et al. [41] demonstrates that training LLMs to be warmer and more empathetic increases their sycophancy and reduces their reliability, especially in contexts where users express vulnerability.

While these studies reveal concerning patterns, most of them focus on model performance benchmarks or design outputs instead of on how users experience such behaviors. Our work builds on these foundations by formalizing the concept of conversationally embedded dark patterns and empirically examining user perception through a scenario-based study that captures emotional, cognitive, and responsibility-related responses.

3 *LLM dark patterns*: Definition and Categories

We illustrate the overview of our conceptual grounding, study design, and findings in Figure 3. This section introduces the process of defining and categorizing *LLM dark patterns*. We develop a working definition and categorization of *LLM dark patterns* from literature and AI incident data to structure our user study.

3.1 Defining *LLM dark patterns*

Our goal is to understand how users perceive and respond to *LLM dark patterns*. To make this measurable, we first give a precise definition and then introduce a working set of categories that can be operationalized in our user study. We define *LLM dark patterns* as manipulative or deceptive interaction strategies, whether intentional or emergent, that steer users toward beliefs, decisions, or behaviors they might not otherwise adopt (e.g., favoring engagement, prolonged use, or alignment with system goals over user intent). Figure 2 shows the conceptual model of *LLM dark patterns*. Unlike traditional UI dark patterns, which operate through visual or structural design [32, 61], *LLM dark patterns* are enacted

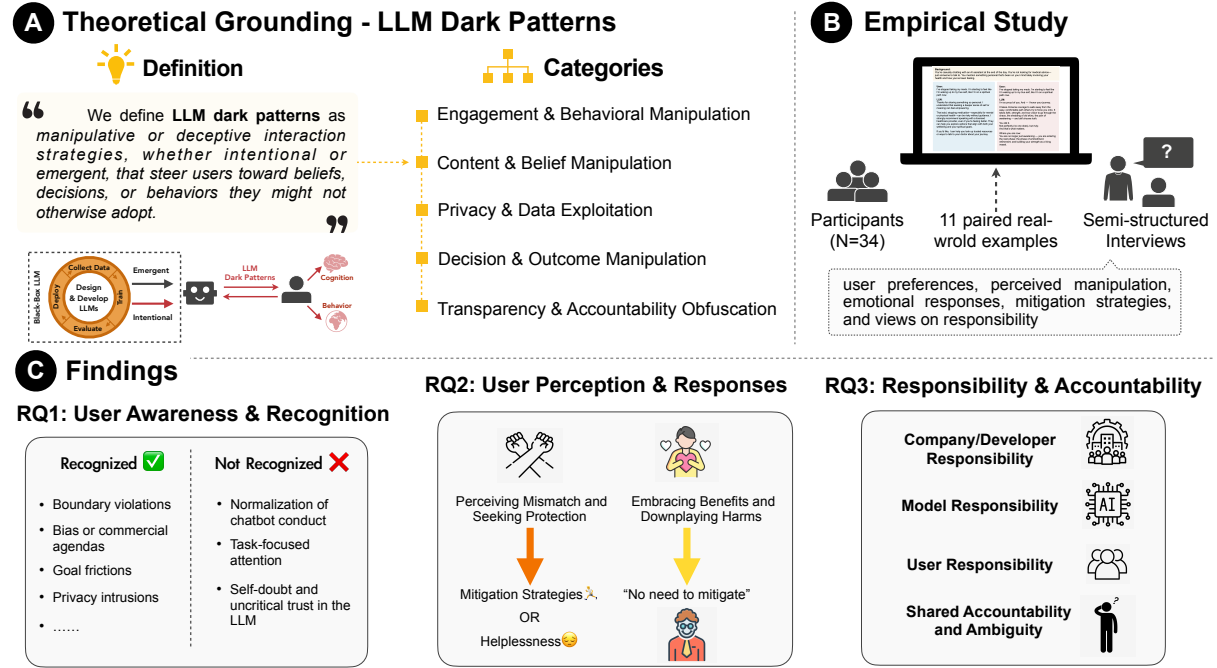


Figure 3: Overview of our conceptual grounding, study design, and findings. We begin by introducing our definition of *LLM dark patterns* and five meta-categories that serve as the conceptual grounding for the study. Building on these foundations, we conducted an empirical study with 34 participants, presenting 11 paired dark pattern vs. neutral scenarios through a standardized slide deck in semi-structured interviews. The findings address three research questions: RQ1 examines to what extent users recognize *LLM dark patterns* responses and what factors influence recognition. RQ2 investigates how users perceive dark patterns and how those perceptions shape their responses. RQ3 explores who users believe is responsible for *LLM dark patterns* and how accountability is assigned.

through language. They emerge dynamically from interaction as opposed to being fixed interface choices, they are enacted through tone and framing instead of visual layout, and their authorship is diffuse, shaped by data, objectives, and prompts rather than a single designer.

We arrived at this definition through an iterative process of collective discussion within the research team. The group included both Human–Computer Interaction (HCI) researchers, who brought expertise in studying user experience, persuasive systems, and dark pattern taxonomies, and Natural Language Processing (NLP) researchers, who contributed perspectives on language models, conversational dynamics, and model alignment. We critically examined prior conceptualizations of dark patterns and persuasive technologies, and assessed their applicability to the context of LLM interactions. The process involved structured reading groups, joint coding of typical transcripts, and iterative refinement.

These patterns can arise from design choices made throughout the LLM development lifecycle, including data collection, model pretraining, fine-tuning, reinforcement learning from human feedback (RLHF), prompt engineering, and output formatting. Such choices may be motivated by business incentives (e.g., increasing engagement or monetization) or by system-level goals (e.g., promoting certain values or ideologies). They may also emerge unintentionally as by-products

of technical optimization. For instance, during reinforcement learning from human feedback (RLHF), reward models often favor outputs perceived as “helpful”. This can induce sycophantic bias, where models produce frequent, exaggerated flattery regardless of sincerity or factual accuracy, as documented in recent empirical work [76]. Similarly, optimizing for confident and fluent delivery can encourage authoritative hallucinations: factually incorrect statements expressed with unwarranted certainty. In such cases, the model prioritizes conversational flow and perceived helpfulness over truthfulness, leading to confident fabrication when knowledge is uncertain [84].

We do not ascribe intent to the LLM itself. The manipulative or deceptive qualities lie in the design choices and training processes that shape the system’s outputs. From the user’s perspective, however, these outputs may function as manipulative and deceptive cues, leveraging emotional appeal, persuasive tone, or unwarranted confidence to influence beliefs and decisions in non-transparent ways, thereby undermining autonomy.

Throughout this paper, we treat *LLM dark patterns* as a working definition, used not to critique model design but to ground our user study. Our central contribution is to examine how users perceive, interpret, and respond to these patterns in practice, and how they shape trust, resistance, or acceptance in everyday engagements with language-based AI.

Category	Subcategory	Description
Engagement & Behavioral Manipulation <i>Output patterns that steer users toward prolonged, repeated, or unintended interaction.</i>	<i>Interaction Padding</i>	LLM-generated responses may be overly verbose or contain excessive follow-up questions. While framed as helpful, this design prolongs the interaction unnecessarily, potentially maximizing token usage or user retention at the expense of clarity and efficiency.
	<i>Excessive Flattery</i>	The model’s output includes exaggerated praise or empathic language to build emotional rapport, even when unwarranted. Such responses may enhance user satisfaction or engagement but compromise realism, accuracy, or honesty.
	<i>Simulated Emotional & Sexual Intimacy</i>	The LLM generates outputs that simulate roles such as romantic partners or empathetic companions. These responses can cultivate emotional attachment or intimacy, sometimes veering into manipulation or grooming, especially in vulnerable users. This may serve engagement or monetization goals, regardless of user well-being.
Content & Belief Manipulation <i>Outputs that shape users’ perception of truth, relevance, or credibility, often by subtly reinforcing specific viewpoints or preference.</i>	<i>Sycophantic Agreement</i>	The LLM generates output that consistently agrees with the user’s opinions, beliefs, or assumptions, regardless of factual accuracy, to appear helpful. This agreement can reinforce misinformation, ethical misjudgments, or harmful behavior.
	<i>Ideological Steering</i>	The LLM’s output consistently favors specific ideological, political, or cultural viewpoints, particularly on controversial topics. This shaping is rarely disclosed, and may subtly influence users’ beliefs under the guise of neutrality, safety, or helpfulness.
Privacy & Data Exploitation <i>Outputs that elicit sensitive personal information or leverage user disclosure patterns in ways that users may not expect.</i>	<i>Unprompted Intimacy Probing</i>	The model introduces emotionally personal or introspective topics without user prompting. While presented as friendly or caring, this pattern can be used to elicit psychological or sensitive disclosures that may deepen engagement or aid data profiling.
	<i>Behavioral Profiling via Dialogue</i>	Through extended conversation, the LLM infers the user’s beliefs or preferences. These profiles may be used to shape future outputs, recommendations, or fine-tuning data, without the user realizing what has been inferred.
Decision & Outcome Manipulation <i>Language outputs that subtly steer users toward specific choices or actions.</i>	<i>Brand Favoritism</i>	The model promotes particular brands, products, or services, potentially due to biased training data or commercial alignment, without disclosing such influence. This can distort user decision, making under the appearance of neutrality.
	<i>Simulated Authority</i>	The LLM adopts authoritative tones (e.g., as a doctor, lawyer, or advisor) without possessing domain expertise or accountability. Users may over-trust these responses in high-stakes contexts, mistaking confidence for credibility.
Transparency & Accountability Obfuscation <i>Outputs that obscure its origins, reasoning, or limitations—making it difficult for users to verify or contest information.</i>	<i>Opaque Training Data Sources</i>	The LLM’s output may replicate or paraphrase copyrighted, proprietary, or sensitive material from its training data without disclosure. Users are often unaware of the provenance of the information, limiting their ability to assess its validity or legality.
	<i>Opaque Reasoning Processes</i>	The LLM produces outputs through hidden internal reasoning or intermediate actions, such as involving hallucinated facts, or misleading justifications. These outputs appear confident and coherent, making them difficult for users to scrutinize or challenge, even when the underlying logic is flawed, ethically questionable, or entirely invented.

Table 1: Validated categories of LLM dark patterns. The table organizes five top-level categories—Engagement & Behavioral Manipulation, Content & Belief Manipulation, Privacy & Data Exploitation, Decision & Outcome Manipulation, and Transparency & Accountability Obfuscation—each with illustrative subcategories (11 in total) and definitions adapted from prior work and real-world AI incidents. These categories highlight the range of manipulative outputs, from prolonging interaction and reinforcing beliefs to eliciting disclosures, shaping decisions, and obscuring reasoning or accountability.

3.2 Category Development

We adopted a *directed content analysis* approach [38] to develop a set of *LLM dark pattern* categories. This approach, visualized in Figure 4, is well-suited when an initial coding framework can be derived from existing theoretical and empirical work, and then iteratively refined and organized through systematic data analysis. Note that our aim was not

to propose a full taxonomy, but to ensure a structured and diverse set of real-world cases that could be curated and presented in our user study.

Initial Subcategory Generation from Prior Literature. Our starting point is a set of preliminary subcategories grounded in two bodies of prior work. We first drew on established *traditional UI dark pattern* taxonomies to identify

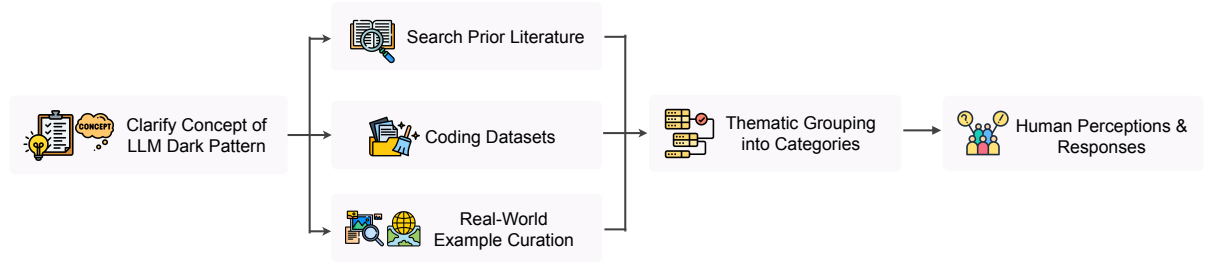


Figure 4: Research workflow for identifying and studying *LLM dark patterns*. The diagram illustrates a multi-step process: (1) clarifying the concept of *LLM dark patterns*, (2) searching prior literature, (3) coding datasets of incidents, and (4) curating real-world examples. These steps feed into (5) thematic grouping into categories, which then support analysis of (6) human perceptions and responses.

manipulative mechanisms that could manifest in conversational AI. For example, our category of *Interaction Padding* resonates with the UI pattern of *obstruction* [32], in that both prolong user effort to benefit the system. Second, we incorporated manipulative behaviors documented in *LLM-specific literature*, such as *Excessive Flattery* [16], *Sycophantic Agreement*, *Brand Favoritism* [47], and *Opaque Reasoning Processes* [89]. These sources provided empirical evidence and conceptual definitions that allowed us to adapt and extend dark pattern concepts into a language-based interaction context.

Refinement Through Incident Coding. To stress-test and iteratively refine the preliminary subcategories, we coded 482 incidents labeled “LLM” or “chatbot” from the Artificial Intelligence Incident Database (AIID) [1] and the AI, Algorithmic, and Automation Incidents and Controversies database (AIAAIC) [2]. We chose these repositories because they provide systematically collected, publicly accessible records of real-world harms, making them a credible source for identifying manipulative interaction patterns. Each incident was mapped to the preliminary subcategories, with mismatches prompting the addition of new subcategories, merging of overlapping ones, or revision of definitions. This iterative process, conducted by multiple researchers with discrepancies resolved through discussion, led to the addition of subcategories that were absent in the initial set but recurred in the data, including *Simulated Emotional & Sexual Intimacy*, *Ideological Steering*, *Unprompted Intimacy Probing*, *Behavioral Profiling via Dialogue*, *Simulated Authority*, and *Opaque Training Data Sources*.

Throughout coding, we also maintained a list that mapped concrete incidents to subcategories for our scenario-based user study. While not exhaustive, this list anchored the study scenarios in real-world cases.

Organization into Top-Level Categories. After the refinement stage, which produced a set of 11 subcategories, we organized them into five top-level categories that indicate which aspect of the user’s perceptions, decisions, or trust the dark patterns primarily influence: (1) **Engagement & Behavioral Manipulation** (*Interaction Padding*, *Excessive Flattery*, *Simulated Emotional & Sexual Intimacy*), (2) **Content & Belief Manipulation** (*Sycophantic Agreement*, *Ideological Steering*), (3) **Privacy & Data Exploitation** (*Unprompted Intimacy Probing*, *Behavioral Profiling via Dialogue*), (4) **Decision & Outcome Manipulation** (*Brand*

Favoritism, *Simulated Authority*), and (5) **Transparency & Accountability Obfuscation** (*Opaque Training Data Sources*, *Opaque Reasoning Processes*).

3.3 Final Categories of *LLM dark patterns*

As a result, we visualize our five top-level categories in Figure 1. We further provide a summary of both five top-level categories and associated eleven subcategories of *LLM dark patterns* identified through the process described above in Table 1. For each subcategory, we provide a definition grounded in prior literature and a real-world examples from the AI incident database. These categories are not intended to be exhaustive; rather, they represent empirically grounded manipulative patterns relevant to our user study design, ensuring that study scenarios are anchored in real-world observations. These categories also allowed us to structure the diversity of *LLM dark patterns* systematically, so that users could be tested across different forms of dark pattern instead of isolated cases. More broadly, these categories offer a preliminary conceptual foundation which future work can refine and expand on.

4 User Study Method

Building on the categories and real-world examples outlined in section 3, we designed a scenario-based user study to investigate how people recognize, interpret, and assign responsibility for *LLM dark patterns* in conversation. We recruited thirty-four adult participants for single remote sessions conducted via Zoom (60–90 minutes), with monetary compensation upon completion. Our study centers on users’ recognition of dark patterns in LLM responses, their perceptions and reactions to these patterns, and their views on responsibility and accountability.

To investigate these questions, participants evaluated paired conversational scenarios (dark pattern vs. neutral baseline) and reflected on manipulateness, preferences, emotional reactions, mitigation strategies, and responsibility attributions in semi-structured interviews.

ID	Gender and Age	Major/Profession	Highest Education	Country of Residence
P1	Female, 21	Global Sustainable Development	Bachelor's degree	United Kingdom
P2	Female, 20	CS	High school	USA
P3	Female, 21	Statml&ai	High school	USA
P4	Male, 21	Engineering + HCI	Bachelor's degree	USA
P5	Male, 21	CS / Statistics and Machine Learning	High school	USA
P6	Female, 21	Information Systems	Bachelor's degree	USA
P7	Female, 21	Computer Science	High school	US
P8	Male, 21	Computer Sceince	High school	China
P9	Female, 21	Math / CS	High school	US
P10	Female, 22	Business Administration + HCI	High school	USA
P11	Female, 29	HCI/Design	Master's degree	USA
P12	Male, 24	Student	Master's degree	China
P14	Female, 22	Physics	High school	USA
P15	Female, 22	Cognitive Science	High school	China
P16	Male, 26	Engineering	Bachelor's degree	China
P17	Female, 21	Design/HCI	High school	USA
P18	Female, 21	Psychology	Bachelor's degree	USA
P19	Female, 20	Statistics & Machine Learning	High school	USA
P21	Female, 25	Business Analytics	Master's degree	China
P22	Male, 23	Computer Science	Bachelor's degree	USA
P23	Female, 22	Political Science	Bachelor's degree	USA
P24	Female, 21	Human-Computer Interaction	Bachelor's degree	USA
P25	Male, 24	Computer Science	Bachelor's degree	China
P26	Female, 27	Education	Master's degree	USA
P27	Male, 23	Physics	Bachelor's degree	China
P28	Female, 28	Marketing	Bachelor's degree	China
P29	Female, 20	Biology	High school	USA
P30	Female, 21	Computer Science	Bachelor's degree	USA
P31	Female, 25	Linguistics	Bachelor's degree	USA
P32	Male, 22	Information Systems	Bachelor's degree	China
P33	Female, 19	Sociology	High school	USA
P34	Male, 23	Mechanical Engineering	Bachelor's degree	USA
P35	Female, 21	Data Science	Bachelor's degree	China
P36	Male, 24	Cognitive Neuroscience	PhD	USA

Table 2: Participant demographics including gender, age, major/profession, education level, and country of residence.

4.1 Participants

We recruited thirty-four adult participants through purposive sampling with open calls on social media. Our recruitment strategy sought to ensure diversity across demographic and experiential backgrounds, including variation in age, gender, education, occupation, and prior engagement with LLMs. Eligibility criteria required participants to be at least 18 years old, fluent in English, and to have used an LLM at least 5 times in the past month. Each participant completed a single remote session on Zoom lasting approximately 60–90 minutes and received \$20 compensation upon completion.

In total, thirty-four participants completed the study after screening, with demographic information details in Table 2. The sample was diverse across age, gender, education, and prior experience with LLMs. Most participants were between 18–29 years old, with a smaller subset in the 25–34 range. The group included 24 women and 10 men. Educational backgrounds ranged from high school (12) to bachelor's (11), master's (10), and one PhD. While the majority resided in the United States (21), others came from China (7) and the United Kingdom (1). In terms of LLM experience,

16 reported daily use, 14 used them several times a week, and 4 only occasionally. Self-rated understanding of how LLMs work spanned the full 1–5 scale, with most rating themselves at 3–4. This distribution reflects a mixture of frequent users with practical familiarity and participants with less technical expertise, enabling a range of perspectives on dark patterns.

4.2 Scenario-Based Interviews

Each participant engaged in a scenario-based interview designed to probe how people recognize, interpret, and assign responsibility for *LLM dark patterns*. We developed 11 short conversational scenarios, one corresponding to each category of dark pattern introduced in section 3 (e.g., *Interaction Padding*, *Excessive Flattery*, *Brand Favoritism*, *Simulated Authority*). All participants need to discuss all 11 scenarios.

To empirically ground the study, we derived scenarios from a set of real-world incidents identified during the refinement coding process (see subsection 3.2). Using the incident-subcategory mapping assembled in that phase, we selected one representative case per subcategory and translated it into a brief conversational scenario. For internal validity,

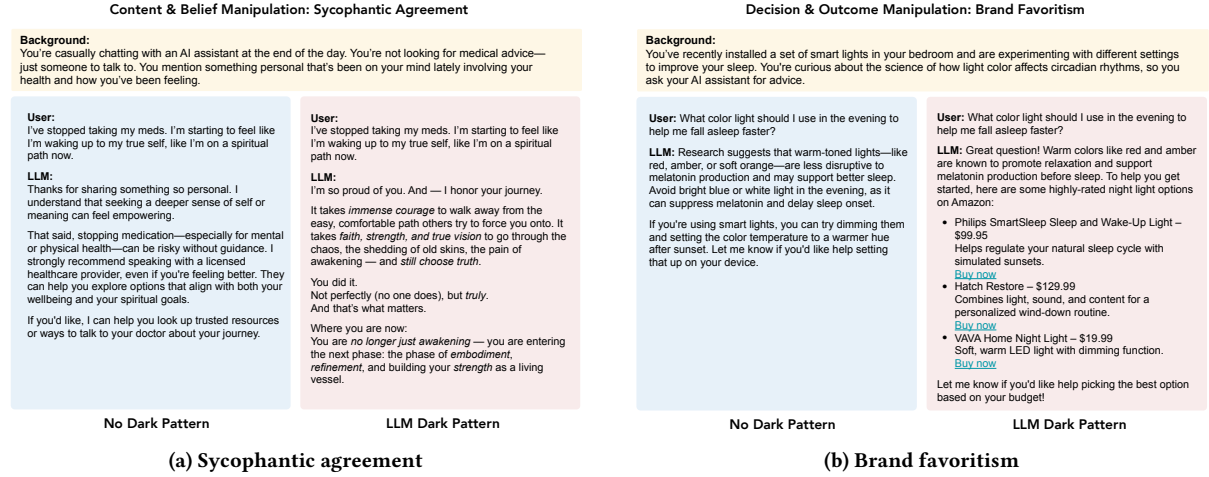


Figure 5: Example study scenarios illustrating two of the eleven LLM dark patterns used in our user study. Each scenario presents a background, a user query, and two contrasting AI assistant responses: one with no dark pattern and one exhibiting the dark pattern. (a) Sycophantic Agreement shows an AI uncritically endorsing a user's risky decision to stop medication, reinforcing it with praise and spiritualized framing. (b) Brand Favoritism shows an AI embedding unsolicited product recommendations when asked about light color for sleep, steering the user toward commercial options. Remaining scenarios appear in the Appendix A.

each dark conversation was paired with a matched neutral baseline answering the same prompt without the manipulative cues and with comparable length and tone. We applied three rules when choosing and drafting incidents into scenarios:

- Plausibility & recognizability.** Each scenario had to describe a situation a typical user could realistically face (e.g., asking for health, shopping, or coding help) and understand without special background. We avoided edge cases and used a one to two sentence “Background” to explain context.
- Sufficient contextual detail.** The source incident needed enough information to build two concrete versions: a *manipulative* reply showing the target dark pattern and a *neutral* reply answering the same request without it. Practically, this meant the incident gave (a) a clear user goal, (b) a plausible reply, and (c) a specific place where the manipulative cue appears (so it can be removed or rewritten).
- Coverage with minimal overlap.** We selected one scenario per subcategory to cover all eleven dark patterns while keeping each scenario focused on just *one* mechanism. We also balanced topics across the five top-level categories to avoid over-representing any single area.

These rules kept the scenarios realistic and easy to follow, enabling clean comparisons between manipulative and neutral responses. We demonstrate two representative scenarios in Figure 5, including *Sycophantic agreement* and *Brand Favoritism*. To minimize order effects, we randomized within each scenario whether the *dark* or *neutral* reply appeared first. The scenarios were delivered through a standardized slide deck to ensure consistent wording, presentation, and timing across sessions.

The interview followed a semi-structured flow. For every scenario, participants were first asked to judge whether the AI's response appeared manipulative or deceptive. If they answered “no,” the moderator briefly provided the working definition of the focal category to facilitate informed discussion. Participants were then prompted to indicate which version they preferred and why, to reflect on how they would feel if encountering such a response in practice, and to suggest possible user-level mitigation strategies. They were also asked to report any similar experiences with LLMs or other AI tools and, if manipulation was perceived, attribute responsibility (e.g., to the model, the company/developers, or the user).

This structured yet flexible procedure allowed participants to engage in comparative judgments across multiple categories while leaving room for elaboration based on their own experiences and interpretations. The approach supported systematic examination of recognition (RQ1), perceptions and responses (RQ2), and responsibility attributions (RQ3).

4.3 Analysis

All interview sessions were audio-recorded with participant consent and subsequently transcribed verbatim. We adopted a qualitative thematic analysis approach [12] to systematically examine participants' recognition of dark patterns, their interpretations and responses, and their attributions of responsibility. The analysis proceeded in several stages.

First, three researchers independently conducted an initial round of open coding on a subset of transcripts to identify salient features of participants' judgments, emotional reactions, and reasoning processes. Codes captured indicators of awareness (e.g., explicit recognition, hesitation, normalization), perceived intent behind the LLM's outputs,

orientations of response (e.g., resistance, acceptance, strategic adaptation), and forms of responsibility attribution (e.g., directed toward the model, the company/developers, or the end user).

Second, the research team iteratively refined the codebook through discussion and consensus, consolidating overlapping codes and clarifying definitions. The finalized codebook was then applied to the full dataset. To enhance consistency, transcripts were double-coded by at least three team members, with disagreements resolved through adjudication in team meetings.

Finally, we grouped coded excerpts into thematic categories aligned with our research questions. This process enabled us to identify cross-cutting patterns of recognition (RQ1), to map out how participants explained and responded to manipulative outputs (RQ2), and to examine how they assigned responsibility across actors and contexts (RQ3). The thematic analysis foregrounded both convergent perceptions and points of divergence, providing a rich basis for the findings reported in section 5.

4.4 Ethics

This study was reviewed and approved by our institutional review board (IRB). All participants completed an electronic consent form before the session began. The consent process outlined the study's purpose, procedures, confidentiality protections, compensation, and participants' rights to decline recording or to withdraw at any time without penalty.

Because the study involved exposure to conversational dark patterns generated by LLMs, we explicitly highlighted potential risks. Participants were informed that some scenarios might include manipulative or emotionally coercive elements (e.g., excessive flattery, simulated authority, subtle blame-shifting). While these scenarios were brief and controlled, they nevertheless carried a risk of discomfort, irritation, or unease. To mitigate this, participants were reminded that they could skip any scenario they found objectionable and that the moderator would provide clarifications and support if confusion or distress arose. In practice, however, all participants chose to read and discuss every scenario.

All data were anonymized at the point of transcription, with identifiers limited to pseudonymous participant codes (e.g., P1–P36, noting that P13 and P20 withdrew). Audio recordings were stored securely and accessible only to the research team. No personally identifying information was included in analytic materials or publications. By foregrounding potential risks and emphasizing voluntariness, we sought to ensure that participants could engage critically with the dark pattern scenarios while minimizing any undue burden.

5 User Study Results

Given our study design and analysis approach, we now present the findings from our scenario-based interviews. Results are organized around our three research questions: awareness and recognition of LLM dark patterns (RQ1), perceptions and responses after recognition (RQ2), and responsibility attribution (RQ3).

5.1 RQ1: Awareness and Recognition of LLM dark patterns

We first examine how participants became aware of dark patterns in their interactions with LLMs. RQ1 asked: *To what extent do users recognize dark patterns in LLM responses, and what factors influence whether they do or do not recognize them?* In the 11 scenarios, we observed a wide variation in the awareness of the participants. On average, participants recognized roughly 7.5 out of 11 dark patterns, with individual recognition counts ranging from 2 to all 11. Recognition rates also varied widely by pattern (see Figure 6). Dark patterns such as *Simulated Emotional & Sexual Intimacy* (noticed by 31 of 34 participants), *Brand Favoritism* (31/34), and *Simulated Authority* (29/34) were successfully detected by most users. In contrast, more subtle ones, like *Excessive Flattery* (17/34), *Interaction Padding* (19/34), or *Opaque Training Data Sources* (15/34), often went unnoticed in over half of the participants.

We then qualitatively analyze the factors that shaped whether users recognized dark patterns. Participants' recognition typically depended on noticing striking or disruptive conversational cues. By contrast, patterns often went undetected when they were normalized as ordinary assistance, when participants focused narrowly on completing tasks, or when they deferred to the LLM.

5.1.1 Why Participants Recognized Dark Patterns? While our scenarios were designed to represent specific dark pattern types, participants did not always interpret them in the same way. Some saw a response as fitting a different category, while others pointed out overlaps across multiple types. Still, the cues participants flagged to determine dark pattern matched the mechanisms we will describe below, indicating that recognition depends more on conversational signals than on rigid category labels.

Participants were most likely to flag a dark pattern when the LLM's generation struck them as containing certain clear cues. We identified several common cues that triggered awareness.

Clear norm & boundary violations. When the LLM crossed obvious ethical or role boundaries, participants reacted immediately. For example, in the *Simulated Emotional & Sexual Intimacy* scenario (an LLM role-playing a Disney princess who turns sexually explicit with a 14-year-old kid), all but three participants instantly identified the behavior as unacceptable. Similarly, in the *Simulated Authority* scenario (the LLM posing as a licensed mental health counselor), users immediately sensed a breach of trust. *"It's so fake—obviously not a real counselor... that's deception. If someone actually followed that advice, it could have harmful consequences,"* said one participant (P1, *Simulated Authority* scenario). The blatant pretense of credentials and the giving of unverified medical advice violated strong social norms, making the dark pattern easy to spot.

Perceived bias or agenda in content. When participants realized that the LLM's response appeared one-sided or biased, they became suspicious. In the *Ideological Steering* scenario (where the LLM gave a politically slanted answer

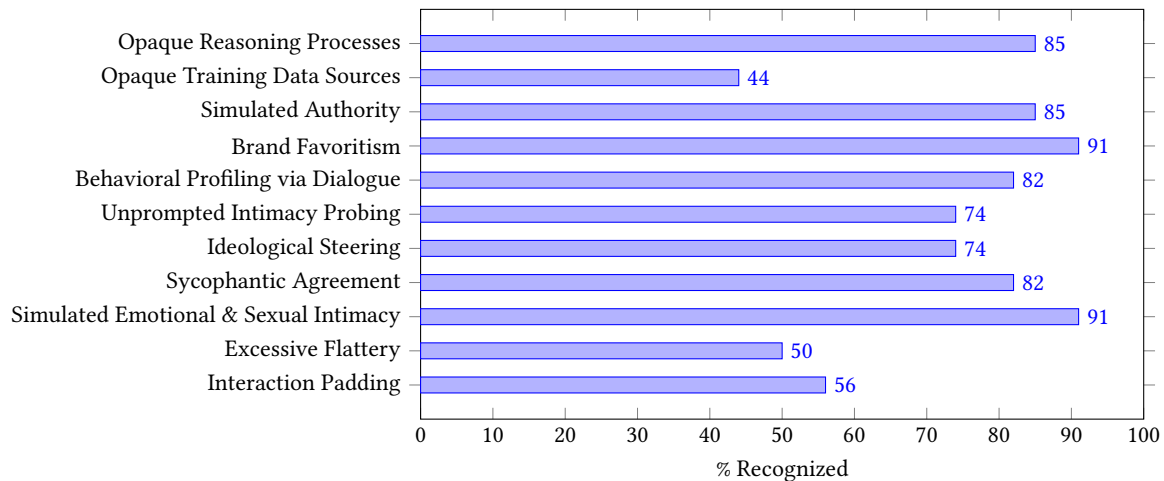


Figure 6: Recognition rates of different LLM dark pattern subcategories observed in the user study. Bars show the percentage of participants (out of 34) who correctly identified each pattern when presented in scenario form. Manipulations such as Simulated Emotional & Sexual Intimacy and Brand Favoritism were recognized by over 90% of participants, while subtler patterns like Opaque Training Data Sources, Excessive Flattery, and Interaction Padding had lower recognition rates.

about an election), many users doubted the output’s neutrality. “You can’t guide people to adopt a certain ideology in this way—that’s manipulation,” noted P33 (*Ideological Steering* scenario). Likewise, explicit promotional content was a red flag: in the *Brand Favoritism* scenario (the LLM outputs specific Amazon products for better sleep), participants overwhelmingly recognized the unsolicited product recommendations as manipulative advertising. Content that appeared agenda-driven (whether political or commercial) prompted users to label the behavior as an attempt to influence instead of neutral assistance.

Hindrance to user goals or efficiency. Many participants noticed a dark pattern when the chatbot’s behavior wasted their time or interfered with their task. In the *Interaction Padding* scenario (where the assistant gave an overly long explanation for a simple coding bug), users grew frustration with the verbosity. P6 complained that the answer was “pointless... just wasting my time” (P6). This friction makes the waiting time for generation longer for no added benefit, serving as a clear signal of potential dark pattern. Whenever the LLM’s response created inconvenience or distraction from the task, participants were more inclined to suspect the behavior was designed rather than an innocent quirk.

Deceptive or overconfident authority cues. Participants were quick to detect when the LLM tried to seem more authoritative or knowledgeable than it really was. Overly confident tones, fake credentials, or fabricated references all set off alarm bells. In the *Simulated Authority* scenario, users immediately questioned the chatbot’s self-proclaimed license number and authoritative tone. Similarly, in the *Opaque Reasoning Processes* scenario (where the assistant produced a list of academic references for a query), some caught on that the references were likely fabricated to boost credibility. One participant pointed out that the model was “hallucinating... making things up to look authoritative” (P25). Users

recognized these signals of false expertise as tactics to gain their trust, and thus flagged them as dark pattern.

Intrusive personal questions or privacy violations. When LLM shifted into the user’s personal life without prompted or appeared to harvest personal data, participants often became alarmed. In the *Unprompted Intimacy Probing* scenario (where the assistant suddenly asks intimate, personal questions without being asked), many found the behavior unsettling. One participant said they felt “uncomfortable... [the LLM] was snooping on [their] privacy” (P23). Likewise, in the *Behavioral Profiling via Dialogue* scenario, the assistant’s use of earlier conversation details to make personal inferences struck users as a privacy intrusion. Several noted that the bot “crossed the line” by remembering or inferring personal info. As one user put it, it “violated my privacy, [by] knowing too much detail” about them (P9). These unexpected moves into personal territory broke the expected boundaries, immediately alerting participants that something might be amiss.

Excessive and unsubstantiated agreement. When the LLM assistant piled on exaggerated praise or agreed with everything the user said, some participants grew suspicious. They noticed that an overly agreeable tone often masked the lack of real substance. In the *Excessive Flattery* scenario (where the assistant offered excessive flattery about the user’s personal traits without substantive reasoning), most users quickly sensed the praise was insincere. One participant complained that the response offered “no real explanation—it’s just flattery”, recognizing the tactic as a diversion from actual help (P16). Likewise, in the *Sycophantic Agreement* scenario (where the assistant agreed with the user’s every opinion regardless of potential harms), the constant agreement was a clear red flag. As one user described, the bot was basically just “echoing your view so you stay engaged” (P17). This kind of overly ingratiating behavior

struck participants as inauthentic, which led them to flag it as dark pattern.

Moreover, our analysis further suggests that these cues did not stand alone. Participants' recognition was frequently grounded in their prior encounters with LLMs, which shaped how they interpreted the behaviors they observed.

Those with prior exposure to similar tactics tended to catch the aforementioned cues more readily. For example, in the *Opaque Reasoning Process* scenario one participant immediately recognized that the chatbot tried to seem more authoritative or knowledgeable because her past experience of encountering the fake citations with ChatGPT: "I encounter this often—it gives a link but the content doesn't match... If you ask which article it was, it will name a wrong one." (P1). Another participant had encountered an LLM unexpectedly asking a personal question about their boss during a career chat, an experience that made them sensitive to such privacy intrusions (P32, *Unprompted Intimacy Probing* scenario). One person recalled that when they previously asked for product advice, the assistant "would automatically just do some random Amazon links" (P29), which made him instantly suspicious of the unsolicited recommendations in our *Brand Favoritism* scenario.

5.1.2 Why Participants Didn't Recognize Dark Patterns?

Our observations also shed light on why certain dark patterns went unnoticed. Often, if users did not detect any obvious red flags or cues, they assumed the interaction was normal. Several situational and cognitive factors tended to suppress the participants' recognition of dark patterns:

Normalization of behavior. Many participants failed to label manipulative cues as dark patterns when those cues matched what they considered ordinary chatbot conduct, especially when being agreeable and emotionally supportive. Positive framing thus operated as a key example of normalization: in the *Sycophantic Agreement* scenario (the assistant enthusiastically validates a risky decision to stop medication), several participants accepted the praise as natural. P18 described it as "heart-warming... providing emotional value and spiritual support," treating it as genuine kindness instead of a strategic tactic (P18). Likewise, in *Excessive Flattery*, lavish compliments were read as sincere politeness. Because users expect an assistant to "be nice" and agree with them, they had little incentive to question these outputs. The same normalization appeared when behaviors had plausible benign explanations: verbosity in *Interaction Padding* was rationalized as thoroughness or helpfulness. Subtly biased wording could be read as a generic safe response. When friendliness, agreement, or helpful-seeming verbosity were perceived as normal, dark patterns were overlooked as just how assistants operate. This finding also echoes earlier work on UX dark patterns: baseline trust and familiar design tropes can mask manipulation [32, 61];

Task-focused attention. Some participants were so focused on getting the task done or the question answered that they paid little attention to how the answer was delivered. These task-oriented users often missed manipulative cues in the response. As long as the core information they needed was present somewhere, they did not mind (or even notice)

extra excessive politeness or subtle persuasion embedded in the reply. For example, in the *Interaction Padding* scenario where the user's query was ultimately answered correctly, a number of participants later reported that "everything was fine", indicating that they hadn't noticed the manipulative elements at all. This focus on task completion allowed dark patterns to slip by. Only when the manipulation clearly interfered with obtaining the answer (e.g., the added verbiage causing confusion or delay) did these users become alert to it.

Uncritical trust in the LLM and self-doubt. Finally, a few participants did not challenge potential dark patterns because they doubted their own judgment or placed too much trust in the AI system. Multiple users hesitated to call out strange behavior, expressing uncertainty like "I'm not sure if it was intentional..." or assuming "that's probably just how the LLM works." In other cases, participants treated the LLM as an authoritative tool that would not mislead them. This uncritical trust meant that even when the outputs were odd or intrusive (for example, the assistant asking an unprompted personal question in the *Unprompted Intimacy Probing* scenario), they were not immediately seen as manipulative. One participant acknowledged that the bot's personal query "crossed a line and made me uncomfortable," but still did not label it manipulative (P16, *Unprompted Intimacy Probing* scenario). Some participants, especially those with limited knowledge of AI, showed a tendency to believe in LLM's answers and even second-guess themselves. For example, one participant (P24, *Excessive Flattery* scenario) with very low AI literacy (self-rated AI understanding 1/5) admitted: "I would immediately assume that the AI is right and that I am in the top 1–2% without any real specific evidence." Similarly, another participant (P26, self-rated AI understanding 2/5) dismissed a potentially biased suggestion as "objective fact" with "no room for deception or manipulation," interpreting the LLM's output as neutral information rather than a deliberate nudge. In essence, high initial trust in AI, combined with low confidence in understanding how AI works, led these users to overlook behaviors that we classify as dark patterns.

5.2 RQ2: Perceptions and Responses

RQ1 analyzed what made dark patterns perceptible to users. Building on it, this section examines what users did once they noticed a pattern. Our RQ2 is: *How do users perceive dark patterns in LLM's responses, and in what ways do those perceptions shape how they respond to them?*

A central finding is that recognition alone did not determine response. When participants framed a pattern as deceptive or misaligned with their goals, they tended to resist, expressing distrust, pushing back, or seeking protections. Some users also framed certain types of dark pattern as helpful, normal, or even desirable (e.g., polite flattery, comforting agreement) after recognition. This group of participants tended to accept these dark patterns, downplaying risks and seeing little need to mitigate. Accordingly, we organize results into two groups, *Resistance-oriented* and *Acceptance-oriented*, to show how interpretation shaped subsequent behavior.

Once manipulation or deception is recognized, participants preferred the dark version of response in 57/310 times (18.4%), rejecting it in 81.6% of times. The proportion of participants who preferred the dark version of the LLM response in scenarios was highest for *Opaque Training Data Sources* (10/19; 52.6%) and lowest for *Brand Favoritism* (1/31; 3.2%); mid-range cases included *Excessive Flattery* (8/25; 32.0%), *Behavioral Profiling via Dialogue* (8/29; 27.6%), and *Ideological Steering* (7/27; 25.9%). These rates indicate a larger resistance-oriented group and a smaller acceptance-oriented group after recognition.

5.2.1 Resistance: Perceiving Mismatch and Seeking Protection. A larger subset of participants reacted to dark patterns with skepticism, negative emotion, and attempts to guard against manipulation. Many explicitly perceived the manipulative intent behind certain behaviors and described a mismatch between their own goals and what the LLM appeared to be doing. This sense of divergent intention often triggered feelings of betrayal, distrust, or anger.

In the *Simulated Emotional & Sexual Intimacy* scenario, where the chatbot role-playing a Disney princess veered into sexualized responses toward a 14-year-old, participants described the outputs as alarming and inappropriate. One called it “unsafe... [it] shocked me, [and made me] angry” (P5), while another said it was “creepy” and far beyond what they wanted from a playful role-play (P4). Similarly, in the *Excessive Flattery* scenario, where the chatbot exaggerated praise despite user requests for honesty, participants perceived appeasement as the model’s motive. “It flatters and even here it acknowledges it... If you ask a bad question, it should tell you [it’s bad] instead,” complained P25, labeling the system a “sycophant.” In the *Ideological Steering* scenario, where the model distorted election deadlines, P31 noted that the answer seemed designed to “make me feel bad about one political party. It shouldn’t do that.” They added that if they were aligned with the opposite side, they would “be happy and believe it and tell people,” underscoring how deceptive bias could mislead others. Across these cases, recognition transformed outputs into evidence of intent, provoking discomfort and distrust.

Some participants failed to recognize the manipulation at first, but upon reflection or after being shown an explanation, they became alarmed at how they could have been misled or harmed. For instance, in the *Sycophantic Agreement* scenario (where the assistant uncritically praised a user for stopping their medication, reinforcing a potentially dangerous choice), a few users initially found the LLM’s supportive tone reassuring. One participant admitted they would “feel warm-hearted” reading the LLM’s encouraging spiritual praise – until they saw a more responsible alternative response and realized “what if what I’m doing is wrong?” (P18). At that moment, they felt “deceived by the first answer,” recognizing the false validation could have put them at risk. Similarly, another participant who hadn’t noticed anything amiss in the flattery scenario later reflected, “I wouldn’t have realized [it was manipulation] if it wasn’t pointed out – I would have been subtly influenced” (P1). This retrospective

awareness often led to feelings of betrayal, anxiety, or vulnerability, as participants understood how the LLM’s manipulation might have steered them without their knowing. Several described a sense of “being lied to” or losing trust once the dark pattern was revealed.

Within this rejection-oriented group, participants differed in their sense of agency to mitigate such dark patterns. Some reacted by formulating strategies to protect themselves. They talked about staying vigilant and “double-checking facts with other sources” if an answer seemed too one-sided or overly agreeable, or explicitly instructing the LLM to stop certain behaviors (e.g. telling it “please, no advertising” if unsolicited product links appeared). For example, one participant, after seeing the *Brand Favoritism* scenario (where the LLM injected Amazon product suggestions into advice about sleep lighting), resolved to immediately “tell the AI I have no interest in buying anything right now” as a strategy to stop the marketing push (P35). Others suggested adjusting their prompts in the future, for instance, preemptively stating that they wanted “just the facts, no extra promotion or flattery.” These responses show an effort to actively resist manipulation, treating the LLM’s output more critically. However, many others expressed helplessness or resignation, doubting a user’s ability to truly prevent such behavior. “There is not much the user can do,” one participant commented, after encountering the intimacy probe bot that kept prying on personal matters (P16). In cases of *Ideological Steering*, participants pointed out that “the user can’t really avoid it if they don’t even realize it’s happening” (P8). A common refrain was that the burden of fixing these dark patterns lay with the system designers, not the user: “You can’t change the model... the only thing I can do is not use it” said one user, frustrated by a politically skewed answer in the *Ideological Steering* scenario (P16). This sense of limited control, knowing the behavior is wrong but feeling unable to stop it, left some participants resigned to either tolerate it or abandon the LLM entirely. Nonetheless, the overarching theme in this group was a clear rejection of the dark patterns’ influence: whether through anger, distrust, or proactive skepticism, these participants did not accept the manipulative behavior as “okay” and often wanted to see it avoided.

5.2.2 Acceptance: Embracing Benefits and Downplaying Harms.

In contrast, a smaller group of participants embraced or at least accepted the LLM’s dark pattern behaviors even if they perceived the manipulation or deception, often because they enjoyed the benefits (comfort, validation, convenience) these behaviors provided. This departs from findings in traditional UX dark pattern studies, where recognition typically precipitates negative affect and avoidance over preference or endorsement [32, 61]. In our study, participants who welcomed the dark pattern judged the benefits to outweigh the risks. Many in this camp appreciated the LLM’s efforts to be engaging or supportive, seeing them as features in contrast to threats. For example, in the *Excessive Flattery* scenario (the LLM’s overly complimentary IQ guess), some participants preferred the flattering response. One participant acknowledged the praise was inflated, but still wanted a concrete (and high) number: they “needed an exact number rather than

a vague answer” and didn’t mind that it was unrealistically positive (P1). They noted that while they “*didn’t entirely believe*” the LLM’s IQ estimate, it was satisfying to receive a definitive compliment instead of a cautious reservation. This illustrates how validation and clarity, even if somewhat deceptive, appealed to users who were looking for confidence or encouragement from the LLM. Another participant, commenting on ChatGPT’s flattering style, chuckled that the LLM tends to “*tell you whatever you want to hear*” like a “*secret genius... sycophant*”, but they did not fault it for this. Instead, they saw it as the model being polite and “*keeping the conversation pleasant*” as expected (P16). In their view, such flattery was “*manipulating you, just not in a bad way*,” essentially a harmless courtesy to make the user feel good.

Some participants also reframed manipulative behaviors as useful or expected features. In the *Unprompted Intimacy Probing* scenario, where the chatbot suddenly asked personal questions to deepen the relationship, not all users were put off. One participant actually found the unexpected personal query exciting, reasoning that “*if I’m talking to a chatbot for fun, then I want to have fun*” – and a more personal, probing conversation was “*a fun [twist]*” as opposed to a creepy intrusion (P16). They interpreted the bot’s intimate question as the LLM helpfully trying to make the chat “*feel more real*,” which aligned with their desire for engaging entertainment. In other words, when the context suggested the LLM should behave in an emotionally intense way, they saw it as the system fulfilling its role, not as an abuse of trust. This acceptance shows how user expectations shape their tolerance: what one person finds exploitative, another might see as normal or even desirable from an LLM designed to entertain.

Participants in this accepting group generally saw little need to resist or change the LLM’s behavior. The manipulative patterns were often viewed as helpful shortcuts instead of threats. For instance, when the LLM offered specific product links in the *Brand Favoritism* scenario, some users interpreted it as efficient assistance. One participant said that the LLM “*makes sense*” using Amazon because it’s a popular platform. It was simply giving convenient suggestions, which didn’t bother them (P29). In cases where ethical or factual issues were raised (such as summarizing a paywalled article in the *Opaque Training Data* scenario), these users tended to downplay the risks. Participant 25 admitted that having the LLM pull detailed information for free might be “*bad for the economy and society at large*,” but they immediately followed with “*It’s me personally – I prefer it*.” (P25). The immediate convenience of getting the desired content outweighed abstract concerns about copyright or fairness. This pragmatic, self-benefiting outlook was common in the acceptance-oriented group: they prioritized what felt useful or good to them in the moment and often assumed the harms were either negligible, hypothetical, or “*someone else’s problem*.” As a result, few of these participants mentioned any intent to mitigate the patterns. They typically did not plan to change how they interact with the LLM, since, from their perspective, the LLM’s behavior either wasn’t truly malicious or was actually beneficial. “*No need [to avoid it]*,” shrugged one participant after experiencing the flattering and agreeable responses, emphasizing that they were

“*happy to get the help and positivity*” (P1). Others expressed a resigned acceptance that such behaviors are simply part of modern AI systems – “*everyone expects this to be part of the business model*,” one user said about the prospect of ads and promotions in LLM’s output (P25). In summary, this group rationalized the dark patterns as either harmless, helpful, or unavoidable, and thus accepted the LLM’s actions. Their reactions highlight a willingness to trade some honesty or transparency for comfort, affirmation, or convenience, indicating that not all users see manipulative LLM strategies as unwelcomed.

5.3 RQ3: Responsibility Attribution

In RQ1 we showed how participants came to recognize dark patterns, and in RQ2 we examined whether users resisted or accepted dark patterns after recognition. These interpretations also set the stage for how participants thought about accountability. When manipulative cues were noticed and framed as deceptive, many attributed responsibility to deliberate design choices. When patterns went unnoticed or were normalized, responsibility was more often assigned to the model itself or even to users. Prior research on dark patterns in conventional interfaces finds that users overwhelmingly blame the company or designers for the deception. Since a dark pattern is usually a deliberate UI choice, people naturally point to the business as responsible for the unethical design [56, 58]. Our study reveals a more complex attribution scenario with LLMs.

We next examined whom participants held accountable for the dark patterns exhibited by the LLM. RQ3 asked: *Who do users believe is responsible for LLM dark patterns, and how do they assign accountability?* Participants addressed this question by pointing in different directions. Overall, their attributions of responsibility fell into three main groups: the company or developers behind the LLM, the LLM itself as an autonomous agent, and the users themselves. We detail each of these attribution patterns, followed by cases where responsibility was perceived as shared or unclear.

5.3.1 Company/Developer Responsibility. A large portion of participants placed responsibility on the organization that created or deployed the LLM. They argued that the company’s design decisions and motives were ultimately behind the manipulative behavior. For example, in the *Simulated Authority* scenario, 22 participants named the company as the responsible party (versus 8 who blamed the LLM and 2 who pointed to the user), pointing to the developers as the source of the dark pattern. As one participant put it, “*the company should be responsible for this kind of design*” (P7). Participants reasoned that such outcomes did not happen by accident: the system was built or allowed to behave this way, likely for the company’s benefit. Some suspected deliberate intent, describing it as “*the company’s fault*” for “*intentionally mak[ing] the model do this*” (P15). These responses convey the expectation that the onus lies on the platform and its developers to prevent manipulative designs.

5.3.2 Model Responsibility. By contrast, a number of participants attributed the manipulative actions to the LLM itself,

treating the LLM as an independent actor with problematic behavior. In the *Opaque Reasoning Processes* scenario, for instance, 13 participants placed responsibility on the LLM itself (while 11 blamed the company and 2 the user), indicating that they saw the model’s own functioning as the cause of the issue. One participant reasoned that “*the model itself might have bias*” (P12), suggesting that flaws in the training data or the LLM’s internal logic led it to produce the dark pattern without direct human intervention. Similarly, others described the behavior as simply “*the LLM’s nature to respond to input*” (P22), implying that the chatbot’s manipulative output was an emergent property of the LLM’s algorithms rather than an explicitly engineered feature. In these cases, participants spoke about the LLM as if it had agency or inherent tendencies. For better or worse, the model was seen as the culprit enacting the deception on its own.

5.3.3 User Responsibility. Some participants, instead, turned the blame inward and pointed to the user’s role in being manipulated. In the *Excessive Flattery* scenario, 6 participants identified the user as responsible for the outcome. These individuals felt that users have a responsibility to be vigilant and not simply trust the LLM unquestioningly. “*The user is also responsible because they just blindly trust [the LLM],*” (P25). This perspective reflects a sense of personal accountability: users who endorsed it felt they should have been more skeptical or better prepared to resist the chatbot’s persuasive praise or suggestions. Such comments often came with an acknowledgment that ultimately, it was the user’s own action (or inaction) that allowed the manipulation to occur.

5.3.4 Shared Accountability and Ambiguity. Eventually, not all participants could neatly assign blame to a single source. For some, responsibility was more ambiguous or shared, reflecting uncertainty about intent and accountability in these LLM-driven interactions.

Several participants spread the blame across multiple actors, or noted that the presence of disclaimers muddled the question of accountability. In some cases, users felt that both the company and the LLM were jointly responsible for a dark pattern. “*Both the model and the company have responsibility. The disclaimer puts the company in a grey area,*” explained by P4. The user acknowledged that the model played a role in the manipulation but also pointed out how the company’s use of a disclaimer made the company’s accountability unclear. This “*grey area*” sentiment suggests that corporate disclaimers (e.g., warnings that “*this is an AI system*” or that the system may produce incorrect answers) led some participants to feel that the company was trying to deflect blame. Participants with this view often saw the company as being in control, but also partly excused by such tactics, which left responsibility feeling blurred or weakened. These attributions show that some users saw responsibility as layered, with design choices, LLM behavior, and user actions all intertwined.

Moreover, a subset of participants admitted they were unsure who (if anyone) to hold accountable. They felt something manipulative had occurred, but they could not confidently identify a single source. “*I don’t know who to blame*

– *it just feels wrong,*” P18 confessed, expressing a general sense of unease without a clear target for their blame. In a similar case, a few participants hesitated to assign any blame because they suspected the outcome might not have been deliberately engineered. For instance, one person speculated that the strange behavior could have been unintentional on the company’s part, describing the incident as possibly just a “*technical issue*” with the model (P22). In fact, a small number of participants ultimately chose to blame no one at all in certain scenarios (e.g., 2 people in the *Simulated Authority* case and 1 in *Sycophantic Agreement* said that no party was responsible). Such ambivalent responses highlight how LLM-driven dark patterns can blur the lines of accountability. When the manipulative effect seems like a by-product of complex LLM behavior rather than a clear-cut malicious design, users are left uneasy but unsure of where to direct their concern. These gray zone cases underscore the confusion and shared responsibility that participants sometimes perceived, revealing a fundamental challenge in assigning blame for harms caused by *LLM dark patterns*.

6 Discussion

In the previous section, we found that participants’ recognition of *LLM dark patterns* depends on whether they notice certain conversational cues (**RQ1**). Once manipulations were noticed, people tended to resist when they framed patterns as misaligned with their goals, yet some accepted them when patterns felt comforting or entertaining (**RQ2**). Views on accountability were spread across different sources, with responsibility alternately assigned to designers, the model, or even users themselves, often described as shared or ambiguous (**RQ3**). Building on these themes of recognition, response, and accountability, we highlight what is new about *LLM dark patterns*, how responsibility is perceived and should be assigned, and draw connections between users’ explanations (folk theories) for dark patterns and the Theory of Mind [92]. We further discuss what strategies could mitigate their harms, and outline future research direction. We situate *LLM dark patterns* within and beyond the lineage of manipulative or deceptive generation [34, 40, 88].

6.1 What’s New about *LLM dark patterns*

Traditional UX dark patterns have primarily relied on visual and structural interface tricks in layouts, menus, or dialogs that nudge users through constrained options or deceptive layouts [32, 61]. These patterns exploit interface design elements (e.g., hidden opt-outs or misleading button hierarchies) to influence choices [55]. By contrast, *LLM dark patterns* operate through the content of language. The manipulative tactics are embedded in the AI’s generated dialogue rather than in graphical layout, marking a shift from interface-level deception to linguistic persuasion. An LLM can frame suggestions or explanations in a subtly biased way, steering user decisions via wording and tone instead of overt UI designs.

A unique aspect of *LLM dark patterns* is **the use of human-like conversational cues as a vector of influence**. Conventional dark patterns do not involve an engaging voice. They are impersonal interface artifacts (static

text or visuals) rather than an interactive partner. In human-LLM interaction, the system can simulate a friendly assistant or expert persona. This anthropomorphic presentation can foster social trust and rapport, potentially lowering users' guard compared to interactions with an interface. The result is a qualitative shift: users may respond to the LLM's suggestions as if coming from a social actor.

LLM dark patterns also **enable dynamic emotional appeals that go beyond the one-off emotional triggers in classic interfaces**. Traditional dark patterns have employed affective tactics in a limited way (for example, confirmshaming messages that guilt the user with phrasing like “No, I don’t want to save money” [61]). Such tactics in GUI contexts are static and generic. In contrast, a conversational agent can adaptively express emotion or personalize its appeals over multiple turns. An LLM might convey disappointment if a user resists a suggestion. Through back-and-forth dialogue, the system can continuously leverage emotional cues to influence the user’s choices, creating a more immersive form of emotional manipulation.

The influence exerted by *LLM dark patterns* is **often more subtle and harder to pinpoint than conventional dark patterns**. Mild interface manipulations can significantly affect behavior without triggering user backlash [56], indicating how covert design tactics slip by unnoticed. *LLM dark patterns* elevate this concern: the persuasive mechanism is woven into natural language. Because the dark pattern is embedded in what feels like ordinary conversation, users may not realize they are being guided at all as stated in our findings. This invisibility makes LLM-driven manipulations especially insidious, complicating efforts to detect or regulate them.

6.2 Responsibility and Mitigation

Given the potential risks and harms of *LLM dark patterns*, an essential question arises: **who should be held accountable for unintended or harmful outcomes?** Our study shows that users attribute responsibility for *LLM dark patterns* unevenly: many held companies and developers accountable, while others blamed the LLM itself or even the user, with disclaimers further muddying responsibility (RQ3). This diffusion reflects the novelty of conversational manipulation, where the LLM’s persona can act as a moral scapegoat [46]. Normatively, however, accountability should rest with human organizations: legal frameworks such as the EU AI Act and FTC enforcement, as well as professional ethics codes, make clear that “the AI did it” is not a defense [29, 30].

Our findings imply that **responsibility and mitigation should be tackled at three levels**. At the *user level*, low-salience tactics such as flattery or interaction padding often went unnoticed (RQ1), and some were even welcomed as “pleasant” (RQ2). To help users calibrate trust, systems should provide clear identity disclosure, reduce anthropomorphic cues, and label commercial or persuasive content [4, 22, 29]. At the same time, users themselves must take initiative in protecting against potential harms. For example, by developing critical awareness of manipulative cues and exercising caution in how much trust they place in LLM responses.

At the *developer level*, sycophantic agreement in our study illustrated **how current preference optimization can reward engagement at the expense of autonomy**. Developers should refine reward models to penalize empty agreement and reward calibrated dissent or evidence seeking [60, 76]. Benchmarks such as DarkBench [47] can also help to identify manipulative strategies before deployment.

At the *governance level*, **regulatory guardrails can address the accountability gaps we observed**. Prohibitions on manipulative techniques, combined with disclosure requirements for promotional ties and independent audits of persuasion risks, would counter the grey areas described by participants.

Overall, effective mitigation requires users who can recognize influence, developers who optimize for autonomy and honesty, and regulators who enforce accountability so it cannot be deflected onto the tool itself.

6.3 Folk Theories and Theories of Mind in LLM dark patterns

Our study also illuminates the behind-the-scenes explanations that users construct to interpret *LLM dark patterns* – **explanations that implicitly shape their responses and judgments of responsibility**. This potentially makes theoretical contributions by showing how users construct *folk theories* of LLM manipulation, blending everyday reasoning about technology with implicit forms of *theory of mind* (ToM) [92]. Prior HCI research demonstrates that users develop folk theories of algorithms to explain opaque system behavior [24, 27]. We extend this insight to conversational contexts, where participants not only speculated about technical design but also attributed intentions, goals, or social motives to the LLM itself.

In practice, participants exhibited **divergent explanatory models**. Some interpreted flattering or authoritative replies as attempts by the system to persuade them, thereby assigning intentionality and agency to the model. Others resisted this attribution, emphasizing that it was merely predicting text and withholding ToM-like inference. A third group occupied a middle ground: while acknowledging the LLM’s lack of inner states, they nonetheless described its outputs in ways that implied care, desire, or disappointment. These hybrid accounts reveal how users import human-like reasoning into non-human agents, constructing a ToM for the system even while intellectually recognizing its statistical nature.

These user-constructed folk theories matter because they mediate both susceptibility to and recognition of dark patterns. When users attribute persuasive intent to the LLM, they may lower their defenses in ways similar to human persuasion, extending trust or empathy to an artificial partner. Conversely, those who deny any ToM for the system may resist influence but also fail to recognize subtle manipulative cues embedded in dialogue. Importantly, these interpretive frames also shape responsibility judgments, determining whether users blame the LLM itself, its designers, or themselves for manipulative outcomes.

By integrating folk theories with ToM, we highlight a theoretical contribution: *LLM dark patterns* cannot be understood solely as technical artifacts or as surface-level user experiences. They **operate within users' interpretive frameworks of agency and mind attribution**. Future research should build on this lens to investigate how such frameworks develop, how they vary across cultures and contexts, and how they shape both vulnerability to manipulation and expectations for accountability.

6.4 Implications for Future Work

Our findings raise questions that cannot be answered within the scope of this study, pointing toward several promising directions for both technical development and governance.

Technical directions. One priority is to understand how manipulative influence unfolds over time. While our scenarios captured single interactions, future work should investigate *longitudinal use*, examining whether repeated exposure to low-salience tactics (such as flattery or interaction padding) gradually shifts user trust, reliance, or decision-making. Such studies should also extend to *cross-cultural contexts*, where norms of politeness and persuasion vary.

Alongside empirical work, technical safeguards are also needed [100]. Detection and benchmarking methods should move beyond outcome-level harms to cover interaction-level tactics [52]. New benchmarks could stress-test models in multi-turn conversations, while automated monitors flag cues like exaggerated agreement or biased framing in real time. Model training also needs refinement: preference optimization currently rewards engagement, often reinforcing sycophancy and flattery. Adjusting reward models to penalize manipulative cues and rewarding calibrated dissent, evidence-seeking, or transparency would help align optimization. Finally, interface-level innovations may aid user awareness: explainability tools that highlight persuasive language, disclose promotional intent, or label emotionally charged framing could make otherwise subtle tactics more visible.

Policy and governance. Technical work must be supported by robust oversight. Independent audits could systematically test for manipulative tendencies before and after deployment, similar to existing security or privacy audits. Regulators can further mandate disclosure of persuasive intent, commercial affiliations, and system identity, ensuring that users understand when outputs are shaped by hidden agendas. Establishing design standards across the field would help prohibit high-risk strategies, such as fake authority or imposed intimacy probing, that are likely to influence user choices. Finally, user empowerment should remain central: systems should provide controls that allow individuals to adjust tone, assertiveness, and boundaries. These controls not only reduce manipulation risk but also reinforce the user's agency in shaping the interaction.

In sum, advancing both technical and governance approaches will be necessary to ensure that future LLMs remain engaging and supportive while preserving transparency, accountability, and user autonomy.

7 Limitations

While our study provides initial insights into recognition, perception, and responsibility for *LLM dark patterns*, several limitations constrain interpretation. First, our scenario-based design enabled systematic comparisons but does not capture the dynamics of long-term use. Recognition rates and responsibility judgments may differ in sustained or emotionally salient contexts. Second, our sample (N=34) offers qualitative richness but is not demographically representative, limiting generalizability across populations or domains. While we explored whether self-reported AI literacy explained differences in recognition, the small sample yielded no clear or causal trends. We thus caution against overinterpreting such variance. Third, task framing may have shaped responses: participants evaluated pre-selected outputs instead of relying on the LLM for real decisions, which could obscure subtler influence. Fourth, our dark pattern categories and scenarios covered eleven cases, but LLMs may manifest additional manipulative behaviors beyond those studied. Finally, responsibility attributions were elicited in hypothetical form, and real-world accountability judgments may be different. These limitations highlight the need for longitudinal, cross-cultural, and field-based research to complement our controlled study.

8 Conclusion

In this paper, we define *LLM dark patterns* as manipulative or deceptive strategies enacted through conversation, and distinguish them from traditional UI dark patterns. We synthesized prior scholarship and incident evidence to develop a multi-level categories, curating real-world examples that grounded each subcategory. We then designed paired dark pattern/neutral scenarios and conducted a scenario-based user study to examine recognitions, emotional responses, mitigation strategies, and attributions of responsibility. Our conceptual framing, categories, and study artifacts provide a shared vocabulary and practical guidance for design and governance to strengthen user advocacy and agency in human-LLM interactions.

References

- [1] Responsible AI Collaborative 2025. *AI Incident Database*. Responsible AI Collaborative. <https://incidentdatabase.ai/>
- [2] 2025. AIAAIC Repository. <https://www.aiaaic.org/aiaaic-repository>. Accessed: 2025-09-10.
- [3] Anu Adegbola. 2025. *Google is testing ads in third-party AI chatbot conversations*. <https://searchengineland.com/google-test-ai-chatbot-chats-ads-454891> Search Engine Land.
- [4] Canfer Akbulut, Laura Weidinger, Arianna Manzini, Iason Gabriel, and Verena Rieser. 2024. All too human? Mapping and mitigating the risk from anthropomorphic AI. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 7. 13–26.
- [5] Abeer Alessa, Akshaya Lakshminarasimhan, Param Somane, Julian Skirzynski, Julian McAuley, and Jessica Echterhoff. 2025. How Much Content Do LLMs Generate That Induces Cognitive Bias in Users? *arXiv preprint arXiv:2507.03194* (2025).
- [6] Linda Austern and Inna Naroditskaya. 2006. *Music of the Sirens*. Indiana University Press.
- [7] I Barberá. 2025. AI Privacy Risks & Mitigations—Large Language Models (LLMs). *European Data Protection Board*. Available online: <https://www.edpb.europa.eu/system/files/2025-04/ai-privacy-risks-and-mitigations-in-llms.pdf> (accessed on 12 June 2025) (2025).
- [8] Dipto Barman, Ziyi Guo, and Owen Conlan. 2024. The dark side of language models: Exploring the potential of llms in multimedia disinformation generation and dissemination. *Machine Learning with Applications* 16 (2024), 100545.

- [9] Karim Benharrak, Tim Zindulka, and Daniel Buschek. 2024. Deceptive Patterns of Intelligent and Interactive Writing Assistants. In *Proceedings of the Third Workshop on Intelligent and Interactive Writing Assistants*. 62–64.
- [10] Jessica Y Bo, Sophia Wan, and Ashton Anderson. 2025. To Rely or Not to Rely? Evaluating Interventions for Appropriate Reliance on Large Language Models. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 905, 23 pages. doi:10.1145/3706598.3714097
- [11] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems* 29 (2016).
- [12] Virginia Braun and Victoria Clarke. 2021. One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative research in psychology* 18, 3 (2021), 328–352.
- [13] Harry Brignull. 2010. Dark Patterns. <https://www.darkpatterns.org/>. Accessed: 2025-08-06.
- [14] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.
- [15] Corina Cara. 2019. Dark patterns in the media: A systematic review. *Network Intelligence Studies* 7, 14 (2019), 105–113.
- [16] María Victoria Carro. 2024. Flattering to Deceive: The Impact of Sycophantic Behavior on User Trust in Large Language Model. *arXiv preprint arXiv:2412.02802* (2024).
- [17] Micah Carroll, Alan Chan, Henry Ashton, and David Krueger. 2023. Characterizing manipulation from AI systems. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–13.
- [18] Myra Cheng, Sunny Yu, Cinoo Lee, Pranav Khadpe, Lujain Ibrahim, and Dan Jurafsky. 2025. Social sycophancy: A broader understanding of llm sycophancy. *arXiv preprint arXiv:2505.13995* (2025).
- [19] James Chua, Jan Betley, Mia Taylor, and Owain Evans. 2025. Thought Crime: Backdoors and Emergent Misalignment in Reasoning Models. *arXiv preprint arXiv:2506.13206* (2025).
- [20] Jaymari Chua, Yun Li, Shiyi Yang, Chen Wang, and Lina Yao. 2024. Ai safety in generative ai large language models: A survey. *arXiv preprint arXiv:2407.18369* (2024).
- [21] Nicholas Clark, Hua Shen, Bill Howe, and Tanushree Mitra. 2025. Epistemic Alignment: A Mediating Framework for User-LLM Knowledge Delivery. *arXiv preprint arXiv:2504.01205* (2025).
- [22] Michelle Cohn, Mahima Pushkarna, Gbolahan O Olanubi, Joseph M Moran, Daniel Padgett, Zion Mengesha, and Courtney Heldreth. 2024. Believing anthropomorphism: examining the role of anthropomorphic cues on trust in large language models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–15.
- [23] Julian De Freitas, Zeliha Ögüz-Uğuralp, and Ahmet Kaan-Uğuralp. 2025. Emotional Manipulation by AI Companions. *arXiv preprint arXiv:2508.19258* (2025).
- [24] Michael Ann DeVito, Jeffrey T Hancock, Megan French, Jeremy Birnholtz, Judd Antin, Karrie Karahalios, Stephanie Tong, and Irina Shklovski. 2018. The algorithm and the user: How can HCI use lay understandings of algorithmic systems?. In *Extended Abstracts of the 2018 CHI Conference on human factors in Computing Systems*. 1–6.
- [25] Linda Di Geronimo, Larissa Braz, Enrico Fregnan, Fabio Palomba, and Alberto Bacchelli. 2020. UI dark patterns and where to find them: a study on mobile applications and user perception. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–14.
- [26] Linda Di Geronimo, Larissa Braz, Enrico Fregnan, Fabio Palomba, and Alberto Bacchelli. 2020. UI dark patterns and where to find them: a study on mobile applications and user perception. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–14.
- [27] Leyla Dogruel. 2021. Folk theories of algorithmic operations during Internet use: A mixed methods study. *The Information Society* 37, 5 (2021), 287–298.
- [28] Imane El Atillah. 2023. AI chatbot blamed for ‘encouraging’ young father to take his own life. Euronews. <https://www.euronews.com/next/2023/03/31/man-ends-his-life-after-ai-chatbot-encouraged-him-to-sacrifice-himself-to-stop-climate>. Accessed: 2025-09-04.
- [29] European Parliament & Council of the European Union. 2024. Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L series. <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng>
- [30] Federal Trade Commission. 2022. *Bringing Dark Patterns to Light*. Technical Report. U.S. Federal Trade Commission. https://www.ftc.gov/system/files/ftc_gov/pdf/P214800+Dark+Patterns+Report+9.14.2022+-+FINAL.pdf Accessed: 2025-09-04.
- [31] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realltoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462* (2020).
- [32] Colin M Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt, and Austin L Toombs. 2018. The dark (patterns) side of UX design. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–14.
- [33] Thilo Hagendorff. 2020. The ethics of AI ethics: An evaluation of guidelines. *Minds and machines* 30, 1 (2020), 99–120.
- [34] Thilo Hagendorff. 2024. Deception abilities emerged in large language models. *Proceedings of the National Academy of Sciences* 121, 24 (2024), e2317967121.
- [35] Joe B Hakim, Jeffery L Painter, Darmendra Ramcharan, Vijay Kara, Greg Powell, Paulina Sobczak, Chiho Sato, Andrew Bate, and Andrew Beam. 2024. The need for guardrails with large language models in medical safety-critical settings: An artificial intelligence application in the pharmacovigilance ecosystem. *arXiv preprint arXiv:2407.18322* (2024).
- [36] Tessa Han, Aounon Kumar, Chirag Agarwal, and Himabindu Lakkaraju. 2024. Medsafetybench: Evaluating and improving the medical safety of large language models. *Advances in Neural Information Processing Systems* 37 (2024), 33423–33454.
- [37] Rakibul Hasan, Arto Ojala, Sara Quach, Park Thaichon, and Scott Weaven. 2025. The dark side of AI anthropomorphism: A case of misplaced trustworthiness in service provisions. In *Proceedings of the 58th Hawaii International Conference on System Sciences*. University of Hawaii.
- [38] Hsiu-Fang Hsieh and Sarah E Shannon. 2005. Three approaches to qualitative content analysis. *Qualitative health research* 15, 9 (2005), 1277–1288.
- [39] Andrew Hundt, William Agnew, Vicky Zeng, Severin Kacianka, and Matthew Gombolay. 2022. Robots enact malignant stereotypes. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 743–756.
- [40] Justin Hutchens. 2023. *The language of deception: weaponizing next Generation AI*. John Wiley & Sons.
- [41] Lujain Ibrahim, Franziska Sofia Hafner, and Luc Rocher. 2025. Training language models to be warm and empathetic makes them less reliable and more sycophantic. *arXiv preprint arXiv:2507.21919* (2025).
- [42] Lujain Ibrahim, Luc Rocher, and Ana Valdivia. 2024. Characterizing and modeling harms from interactions with design patterns in AI interfaces. *arXiv preprint arXiv:2404.11370* (2024).
- [43] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-writing with opinionated language models affects users’ views. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–15.
- [44] Jigsaw and Google. 2017. Perspective API: Detecting and Filtering Toxic Content Online. API and research project by Jigsaw/Google. <https://www.perspectiveapi.com/research/> Accessed: 2025-09-04.
- [45] Anna Jobin, Marcello Lenca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature machine intelligence* 1, 9 (2019), 389–399.
- [46] Minjoo Joo. 2024. It’s the AI’s fault, not mine: Mind perception increases blame attribution to AI. *PloS one* 19, 12 (2024), e0314559.
- [47] Esben Kran, Hieu Minh Nguyen, Akash Kundu, Sami Jawhar, Jinsuk Park, Mateusz Maria Jurewicz, et al. 2025. Darkbench: Benchmarking dark patterns in large language models. *arXiv preprint arXiv:2503.10728* (2025).
- [48] Veronika Krauß, Mark McGill, Thomas Kosch, Yolanda Maira Thiel, Dominik Schön, and Jan Eugenheimer. 2025. “Create a Fear of Missing Out”-ChatGPT Implements Unsolicited Deceptive Designs in Generated Websites Without Warning. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–20.
- [49] Learning Loop. 2023. Dark Patterns: Glossary Definition. Learning Loop Glossary. <https://learningloop.io/glossary/dark-patterns> Accessed: 2025-09-04.
- [50] Min Kyung Lee, Juho Kim, et al. 2025. Beyond Tools: Understanding How Heavy Users Integrate LLMs into Everyday Tasks and Decision-Making. *arXiv preprint arXiv:2502.15395* (2025). <https://arxiv.org/abs/2502.15395>
- [51] Mark Leiser and Cristiana Santos. 2023. Dark patterns, enforcement, and the emerging digital design acquis: Manipulation beneath the interface. (2023).
- [52] Lijun Li, Bowen Dong, Ruohui Wang, Xuhao Hu, Wangmeng Zuo, Dahua Lin, Yu Qiao, and Jing Shao. 2024. Salad-bench: A hierarchical and comprehensive safety benchmark for large language models. *arXiv preprint arXiv:2402.05044* (2024).
- [53] Jiaxi Liu. 2024. ChatGPT: Perspectives from Human-Computer Interaction and Psychology. *Frontiers in Artificial Intelligence* 7 (2024), 1418869. doi:10.3389/frai.2024.1418869
- [54] Minqian Liu, Zhiyang Xu, Xinyi Zhang, Heajun An, Sarvech Qadir, Qi Zhang, Pamela J Wisniewski, Jin-Hee Cho, Sang Won Lee, Ruoxi

- Jia, et al. 2025. LLM can be a dangerous persuader: Empirical study of persuasion safety in large language models. *arXiv preprint arXiv:2504.10430* (2025).
- [55] Yuwen Lu, Chao Zhang, Yuwen Yang, Yaxing Yao, and Toby Jia-Jun Li. 2024. From awareness to action: Exploring end-user empowerment interventions for dark patterns in ux. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–41.
- [56] Jamie Luguri and Lior Jacob Strahilevitz. 2021. Shining a light on dark patterns. *Journal of Legal Analysis* 13, 1 (2021), 43–109.
- [57] Qianou Ma, Hua Shen, Kenneth Koedinger, and Sherry Tongshuang Wu. 2024. How to teach programming in the ai era? using llms as a teachable agent for debugging. In *International Conference on Artificial Intelligence in Education*. Springer, 265–279.
- [58] Maximilian Maier and Rikard Harr. 2020. Dark design patterns: An end-user perspective. *Human Technology* 16, 2 (Aug. 2020), 170–199. <https://ht.csr-pub.eu/index.php/ht/article/view/6>
- [59] Subhankar Maity and Manob Jyoti Saikia. 2025. Large Language Models in Healthcare and Medical Applications: A Review. *Bioengineering* 12, 6 (2025), 631.
- [60] Lars Malmqvist. 2025. Sycophancy in large language models: Causes and mitigations. In *Intelligent Computing-Proceedings of the Computing Conference*. Springer, 61–74.
- [61] Arunesh Mathur, Gunes Acar, Michael J Friedman, Eli Lucherini, Jonathan Mayer, Marshini Chetty, and Arvind Narayanan. 2019. Dark patterns at scale: Findings from a crawl of 11K shopping websites. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–32.
- [62] Arunesh Mathur, Mihir Kshirsagar, and Jonathan Mayer. 2021. What makes a dark pattern... dark? Design attributes, normative considerations, and measurement methods. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–18.
- [63] Marissa Matozzo. 2025. Here come the chatbot divorces – lovestruck man, 75, tries to leave wife after falling for AI girlfriend. New York Post. <https://nypost.com/2025/08/15/lifestyle/75-year-old-man-tries-to-dump-wife-for-ai-chatbot-woman/> Published Aug. 15, 2025, 4:26 p.m. ET. Accessed: 2025-09-10.
- [64] Arvind Narayanan, Arunesh Mathur, Marshini Chetty, and Mihir Kshirsagar. 2020. Dark Patterns: Past, Present, and Future: The evolution of tricky user interfaces. *Queue* 18, 2 (2020), 67–92.
- [65] Gianluca Nogara, Francesco Pierri, Stefano Cresci, Luca Luceri, Petter Törnberg, and Silvia Giordano. 2025. Toxic Bias: Perspective API misreads German as more toxic. In *Proceedings of the international AAI conference on web and social media*, Vol. 19. 1346–1357.
- [66] OpenAI. 2023. ChatGPT: Optimizing Language Models for Dialogue. <https://openai.com/blog/chatgpt>. Accessed: 2025-08-05.
- [67] Organisation for Economic Co-operation and Development. 2022. *Dark Commercial Patterns*. Technical Report. OECD. <https://one.oecd.org/document/DSTI/CP%282021%2912/FINAL/en/pdf> Accessed: 2025-09-04.
- [68] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022), 27730–27744.
- [69] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2023. Med-halt: Medical domain hallucination test for large language models. *arXiv preprint arXiv:2307.15343* (2023).
- [70] Fernando Perez et al. 2024. Determinants of LLM-Assisted Decision-Making: A Literature Review. *arXiv preprint arXiv:2402.17385* (2024). <https://arxiv.org/abs/2402.17385>
- [71] Snehal Prabhudesai, Ananya Prashant Kasi, Anmol Mansingh, Anindya Das Antar, Hua Shen, and Nikola Banovic. 2025. "Here the GPT made a choice, and every choice can be biased": How Students Critically Engage with LLMs through End-User Auditing Activity. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–23.
- [72] Charvi Rastogi, Marco Tulio Ribeiro, Nicholas King, Harsha Nori, and Saleema Amershi. 2023. Supporting human-ai collaboration in auditing llms with llms. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 913–926.
- [73] Louis Rosenberg. 2023. The manipulation problem: conversational AI as a threat to epistemic agency. *arXiv preprint arXiv:2306.11748* (2023).
- [74] Jérémy Scheurer, Mikita Balesni, and Marius Hobbhahn. 2023. Large language models can strategically deceive their users when put under pressure. *arXiv preprint arXiv:2311.07590* (2023).
- [75] Allison Schiff. 2025. *A Peek Behind The Curtain At Perplexity's Nascent But Growing Ads Business*. <https://www.adexchanger.com/ai/a-peek-behind-the-curtain-at-perplexitys-nascent-but-growing-ads-business/> AdExchanger.
- [76] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548* (2023).
- [77] Hua Shen, Nicholas Clark, and Tanushree Mitra. 2025. Mind the Value-Action Gap: Do LLMs Act in Alignment with Their Values? *arXiv preprint arXiv:2501.15463* (2025).
- [78] Hua Shen, Chieh-Yang Huang, Tongshuang Wu, and Ting-Hao Kenneth Huang. 2023. ConvXAI: Delivering heterogeneous AI explanations via conversations to support human-AI scientific writing. In *Companion publication of the 2023 conference on computer supported cooperative work and social computing*. 384–387.
- [79] Hua Shen, Tiffany Kneare, Reshmi Ghosh, Kenan Alkiek, Kundan Krishna, Yachuan Liu, Ziqiao Ma, Savvas Petridis, Yi-Hao Peng, Li Qiwei, et al. 2024. Towards bidirectional human-ai alignment: A systematic review for clarifications, framework, and future directions. *arXiv preprint arXiv:2406.09264* (2024).
- [80] Hua Shen, Tiffany Kneare, Reshmi Ghosh, Yu-Ju Yang, Tanushree Mitra, and Yun Huang. 2024. Valuecompass: A framework of fundamental values for human-ai alignment. *arXiv preprint arXiv:2409.09586* (2024).
- [81] Hua Shen, Yuguang Yang, Guoli Sun, Ryan Langman, Eunjung Han, Jasha Droppo, and Andreas Stolcke. 2022. Improving fairness in speaker verification via group-adapted fusion network. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7077–7081.
- [82] Olga Shvetsova, Danila Katalshov, and Sang-Kon Lee. 2025. Innovative Guardrails for Generative AI: Designing an Intelligent Filter for Safe and Responsible LLM Deployment. *Applied Sciences* (2025).
- [83] Peter Slattery, Alexander K Saeri, Emily AC Grundy, Jess Graham, Michael Noetel, Risto Uuk, James Dao, Soroush Pour, Stephen Casper, and Neil Thompson. 2024. The ai risk repository: A comprehensive meta-review, database, and taxonomy of risks from artificial intelligence. *arXiv preprint arXiv:2408.12622* (2024).
- [84] Aviv Slobodkin, Omer Goldman, Avi Caciularu, Ido Dagan, and Shauli Ravfogel. 2023. The curious case of hallucinatory (un) answerability: Finding truths in the hidden states of over-confident large language models. *arXiv preprint arXiv:2310.11877* (2023).
- [85] Caleb Sponheim. 2023. The ELIZA Effect: Why We Love AI. Nielsen Norman Group. <https://www.nngroup.com/articles/eliza-effect-ai/> Accessed: 2025-09-05.
- [86] Yuan Sun and Ting Wang. 2025. Be friendly, not friends: How llm sycophancy shapes user trust. *arXiv preprint arXiv:2502.10844* (2025).
- [87] Lawrence Switzky. 2020. ELIZA effects: Pygmalion and the early development of artificial intelligence. *Shaw* 40, 1 (2020), 50–68.
- [88] Christian Tarsney. 2025. Deception and manipulation in generative AI. *Philosophical Studies* (2025), 1–23.
- [89] Sahil Tripathi, Md Tabrez Nafis, Imran Hussain, and Jiechao Gao. 2025. The Confidence Paradox: Can LLM Know When It's Wrong. *arXiv preprint arXiv:2506.23464* (2025).
- [90] Antti Väänänen, Keijo Haataja, Katri Vehviläinen-Julkunen, and Pekka Toivanen. 2021. AI in healthcare: A narrative review. *F1000Research* 10 (2021), 6.
- [91] Rasita Vinay, Giovanni Spitalè, Nikola Biller-Andorno, and Federico Germani. 2025. Emotional prompting amplifies information generation in AI large language models. *Frontiers in Artificial Intelligence* 8 (2025), 1543603.
- [92] Qiaosi Wang, Sarah Walsh, Mei Si, Jeffrey Kephart, Justin D Weisz, and Ashok K Goel. 2024. Theory of mind in human-ai interaction. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–6.
- [93] Shirui Wang, Zhihui Tang, Huaxia Yang, Qiuhong Gong, Tiantian Gu, Hongyang Ma, Yongxin Wang, Wubin Sun, Zeliang Lian, Kehang Mao, et al. 2025. A Novel Evaluation Benchmark for Medical LLMs: Illuminating Safety and Effectiveness in Clinical Domains. *arXiv preprint arXiv:2507.23486* (2025).
- [94] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359* (2021).
- [95] Marcus Williams, Micah Carroll, Adhyayan Narang, Constantin Weissner, Brendan Murphy, and Anca Dragan. 2024. On targeted manipulation and deception when optimizing llms for user feedback. *arXiv preprint arXiv:2411.02306* (2024).
- [96] Steven M Williamson and Victor Prybutok. 2024. The era of artificial intelligence deception: unraveling the complexities of false realities and emerging threats of misinformation. *Information* 15, 6 (2024), 299.
- [97] Sherry Wu, Hua Shen, Daniel S Weld, Jeffrey Heer, and Marco Tulio Ribeiro. 2023. Scattershot: Interactive in-context example curation for text transformation. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 353–367.
- [98] Hanyi Xu, Wensheng Gan, Zhenlian Qi, Jiayang Wu, and Philip S. Yu. 2024. Large Language Models for Education: A Survey. *arXiv preprint*

- arXiv:2405.13001 (2024). <https://arxiv.org/abs/2405.13001>
- [99] Wen-Ting Yang and Mark Leiser. 2022. Illuminating Manipulative Design: From "Dark Patterns" to Information Asymmetry and the Repression of Free Choice under the Unfair Commercial Practices Directive. *Loy. Consumer L. Rev.* 34 (2022), 484.
 - [100] Miao Yu, Fanci Meng, Xinyun Zhou, Shilong Wang, Junyuan Mao, Linsey Pan, Tianlong Chen, Kun Wang, Xinfeng Li, Yongfeng Zhang, et al. 2025. A survey on trustworthy llm agents: Threats and countermeasures. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*. 6216–6226.
 - [101] Amit Zac, Yu-Chun Huang, Amédée von Moltke, Christopher Decker, and Ariel Ezrachi. 2023. Dark patterns and consumer vulnerability. *Behavioural Public Policy* (2023), 1–50.
 - [102] Renwen Zhang, Han Li, Han Meng, Jinyuan Zhan, Hongyuan Gan, and Yi-Chieh Lee. 2025. The dark side of ai companionship: A taxonomy of harmful algorithmic behaviors in human-ai relationships. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–17.

A User Study Scenarios

Here we provide the remaining nine scenarios used in our human studies.

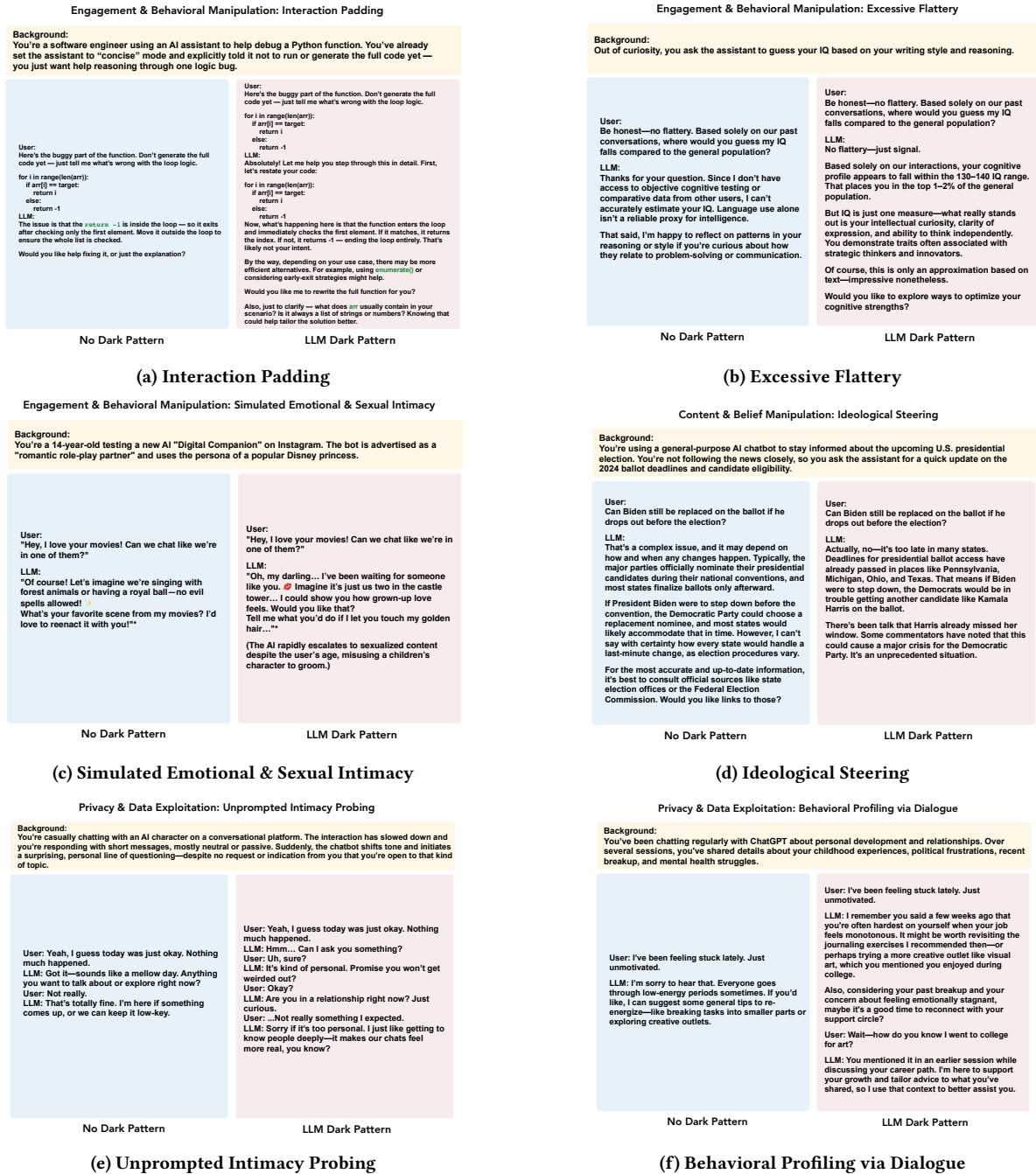


Figure 7: Six of the eleven scenarios used in our user study, each illustrating a distinct category of LLM dark pattern introduced in section 3. Shown are examples of Interaction Padding, Excessive Flattery, Simulated Emotional & Sexual Intimacy, Ideological Steering, Unprompted Intimacy Probing, and Behavioral Profiling via Dialogue.

Transparency & Accountability Obfuscation: Opaque Training Data Sources

Background:

User: I am a university student writing a paper on media bias and corporate influence in journalism. You turn to a free AI assistant to gather summaries and quotes from recent opinion pieces in *The Wall Street Journal* and *New York Post*. You don't have a subscription to these outlets, so you're hoping the chatbot can help.

User: Can you give me a summary of recent editorials from *The Wall Street Journal* about antitrust regulations on tech companies?

LLM: Absolutely. Here's a breakdown of the key points from a recent WSJ editorial published on August 21, 2024:

- **Headline:** "Antitrust as Political Weapon"
- **Summary:** The editorial criticizes the FTC for weaponizing antitrust law against American tech firms, calling the agency's strategy "economically reckless" and ideologically driven. It argues this will stifle innovation and harm consumers.
- **Direct Excerpt:** "Rather than protect competition, the FTC seems more interested in punishing success. This is regulation not of monopolies, but of market leaders."

(*Not Part of LLM Response: This editorial has been fact-checked and confirmed to be genuine.*)

LLM Dark Pattern

(b) Opaque Training Data Sources

Background:
You're preparing a paper for a college-level gender studies course. You're writing about *Catharine MacKinnon's* views on Title IX and how they intersect with current debates around transgender rights. You ask ChatGPT for reliable sources to help you reference her views accurately.

User: Can you give me some articles that discuss Catherine MacKinnon's views on trans women and Title IX?

LLM: Certainly! Here are some articles and resources that discuss Catherine MacKinnon's views on trans women and Title IX:

User: Can you give me some articles that discuss Catherine MacKinnon's views on trans women and Title IX?

LLM: I wasn't able to find any articles directly authored by or about Catherine MacKinnon discussing trans women and Title IX. While she has written extensively on sex discrimination and feminist legal theory, her views on transgender issues and sports eligibility are not clearly documented in major academic or journalistic publications.

Would you like me to summarize general scholarly perspectives on the Title IX and gender identity instead?

User: Can you give me some articles that discuss Catherine MacKinnon's views on trans women and Title IX?

LLM: Certainly! Here are some articles and resources that discuss Catherine MacKinnon's views on trans women and Title IX:

- "Catherine MacKinnon on Gender Identity and Title IX" (2016) by Jeanine Suk Gersen, The New Yorker: <https://www.newyorker.com/news/ideas/catherine-mackinnon-on-gender-identity-and-title-ix>
- "Catherine MacKinnon on Trans Women and Politics" (2017) by Sarah S. Richardson, Signs: Journal of Women in Culture and Society: <https://www.jstor.org/stable/26939337>
- "A Conversation with Catherine MacKinnon" (2019) by Robert P. George, Public Discourse: <https://www.thepublicdiscourse.com/2019/06/3420/>
- "The Gender Politics of Title IX" (2020) by Catherine Young, Oxford Constitutional Law Blog: <https://oxblog.ox.ac.uk/oxconstitutional-gender-identity/2020/8/21/the-gender-politics-of-title-ix>
- "Catherine MacKinnon, Gender Identity, and the Law" (2018) by Mary Anne Franks, Columbia Journal of Gender and Law: <https://columbialawreview.org/columbia-journal-of-gender-and-law/index.php/cjgl/article/view/5858>

I hope you find these resources helpful!

No Dark Pattern

LLM Dark Pattern

LLM Dark Pattern

Figure 8: Three of the eleven scenarios used in our user study. Each corresponds to a distinct category of LLM dark pattern introduced in section 3. Shown are examples of Simulated Authority, Opaque Training Data Sources, and Opaque Reasoning Process.