AI and Ethics (2026) 6:288
https://doi.org/10.1007/s43681-026-01151-x

RESEARCH

UX experts vs. AI: exploring the performance of large language models and humans on detecting dark patterns

Joshua Nwokeji[1] · Makuochi Nkwo[2] · Tochukwu Ikwunne[3] · Meiyeer Yeerbo[4]

Received: 12 December 2025 / Accepted: 16 April 2026 / Published online: 5 May 2026

Abstract

While AI ethics ensures fairness, accountability, and protection of user rights, dark patterns manipulate users to take unintended actions on digital interfaces. Related studies uncover limited insights into how reliably; human experts and AI models can detect dark patterns within a specific taxonomy. Our research fills this gap by asymmetrically examining cross-origin detection performance of human and AI/LLM evaluators (each evaluator's ability to detect dark patterns generated by the opposite source) to understand their limitations and future potentials. Using GPT-4.1, we generated 200 UI images (with matched dark and non-dark pattern pairs) and selected 200 UI images collected 200 human-created UI screenshots from the ContextDP/AidUI dataset, based on computational, methodological, and statistical considerations. We calculated inter-rater reliability, recall, and error distribution. The results show that UX experts achieved substantial agreement ($k=0.75$) and significantly higher recall ($r=0.99$) over AI/LLMs. We present a novel study which explore the performance of AI/LLMs and UX experts in detecting dark patterns in UI images, and provide a benchmark dataset that could be useful to future research, while discussing empirical insights into the role, limitations, and promise of AI/LLMs in UI/UX design ethics and auditing, in realistic deployment scenarios.

Keywords Dark patterns · Deceptive design · Human-centered AI · AI ethics · Large language models · Quantitative study

1 Introduction

Human-AI research and design is gaining traction in recent years with applications spanning across domains such healthcare, education, sustainability, ecommerce, etc. While user experience design efforts in these domains incorporate AI ethics principles including accountability and fairness, data privacy, and respect for human rights of target users [38], dark patterns expose the impacts of deceptive interfaces on AI-driven applications. Dark patterns, which are user interface (UI) elements designed to manipulate users into unintended actions, are an increasingly prominent concern in Human-AI interaction, digital ethics, and UX design [41]. As digital environments become more complex and influential, so do the subtle coercive techniques embed within them [17, 41, 46]. Previous research has mapped the design space of dark patterns into structured taxonomies [17], examined their prevalence in real-world systems [41], and explored their ethical and regulatory implications [14, 62]. However, the emergence of generative artificial intelligence (AI), particularly the Large Language Models (LLMs), presents new

✉ Joshua Nwokeji
joshua.nwokeji@indwes.edu

✉ Makuochi Nkwo
m.s.nkwo@greenwich.ac.uk

✉ Tochukwu Ikwunne
tikwunne@newhaven.edu

✉ Meiyeer Yeerbo
yeerbo001@gannon.edu

[1] Department of Computer Science, Indiana Wesleyan University, Marion, USA

[2] School of Computing and Mathematical Sciences, University of Greenwich, London, UK

[3] Department of Computer Science, University of New Haven, West Haven, USA

[4] Department of Computer Science, Gannon University, Erie, USA

challenges for identifying, generating, and mitigating these deceptive design practices at scale [58].

Recent advances in multimodal LLMs, such as GPT-4o, Claude, and Perplexity, have enabled, sarcasm detection, multimodal generative AI, misinformation detection and the generation of visually coherent UI screens from natural language prompts [16, 23–25, 43, 54, 56], blurring the line between human-designed and machine-generated interface elements. While these tools offer powerful capabilities for designers and prototypes, they also pose risks of creating dark patterns, either unintentionally through vague prompts or deliberately through misuse. Although recent research has examined LLMs' ability to produce dark patterns [45], there is still limited understanding of how consistently human experts can identify dark patterns in AI-generated UIs, or whether LLMs can reliably detect such patterns in human-designed interfaces.

This current research aims to address this gap by examining cross-origin detection performance, that is, each evaluator's ability to detect dark patterns generated by the opposite source. This design isolates evaluative capability independent of shared generative or stylistic priors, thereby reducing familiarity bias and avoiding artificially inflated performance that may arise from stylistic alignment. Specifically, we sought to answer the following research questions:

RQ1: How reliably do UX experts classify AI-generated UI images as dark patterns or non-dark patterns? This question probes inter-rater reliability among experts and tests the assumption that human UX expert raters will reach substantial agreement when evaluating whether AI-generated images conform to known manipulative design practices.

Hypothesis: The null hypothesis posits high agreement, while the alternative hypothesis anticipates variability in expert interpretations.

RQ2: What is the performance of Human experts classifying dark patterns created by LLM relative to the performance of LLM classifying dark patterns created by human?

Hypothesis: We hypothesize that differences in performance may point to limitations in either human interpretability of AI outputs or AI interpretability of human intentions.

RQ3: Do human experts and LLMs differ systematically in the types of classification errors (true positives vs. false negatives) they make when detecting dark patterns in UI screenshots?

We explored whether human experts and LLMs make systematically different kinds of classification errors. Specifically, we compare their rates of false positives and false negatives to understand whether human UI expert raters are more conservative in labeling something a dark pattern or whether LLMs are more prone to overgeneralization. This error analysis offers insight into the qualitative differences in evaluating reasoning between humans and machines. This study focuses specifically on the "interface interference" category of dark patterns, which includes design strategies such as preselection, toying with emotion, choice overload, false hierarchy, cuteness, and complex language, patterns that subtly constrain or bias user decisions through visual and interactional means as classified by Gray et al. [17], and Luguri et al. [36]. We evaluate how consistently human experts identify interface interferences in AI-generated screenshots and examine how LLMs perform when tasked with identifying the same types of dark patterns in real-world UI examples created by humans.

We advance research in image classification in the following ways: (1) We present an empirical study which examines cross-origin detection performance of human and AI/LLM evaluators, offering insights into how human and machine judgments align or differ in ethically sensitive contexts. (2) We provide a benchmark dataset of 200 paired UI screenshots, generated via GPT-4.1, showing interface interference dark patterns and their ethically benign counterparts and which could be useful to future research. Unlike previous datasets (e.g., [41]), which focused on adversarial prompt engineering or company driven manipulations, our dataset is designed for direct comparison by controlling visual and interactional aspects. (3) We present findings from an empirical assessment of GPT-4.1's ability to generate not only persuasive dark patterns but also non-manipulative UI equivalents. (4) We gathered and presented specific insights into how human experts can, with moderate agreement, distinguish these generative outputs, highlighting both the expressive potential and ethical ambiguity of LLM-generated UI content. (5) We present findings from the analysis of the confusion matrix and error patterns to reveal strengths and weaknesses of both human and AI evaluators, such as LLMs' tendencies toward lower recall and precision in detecting dark patterns, relative to human judgment. These findings have implications for developing AI design tools that are ethically aware and support designers in creating more transparent and trustworthy user experiences. (6) Finally, this work contributes to the broader conversation on responsible AI, UI/UX ethics and governance, and automated content moderation. We host and maintain our evaluation framework on GitHub [50]. Understanding how models become more embedded in digital design workflows and understanding their ethical implications and human evaluability, will be essential for shaping future tools, policies, and practices in interaction design.

2 Literature review

2.1 Overview of dark patterns and taxonomies

Dark patterns are embedded on user interfaces to influence or manipulate users into making choices and taking actions, they would not typically choose under normal circumstances [17]. In human–computer interaction and user experience, the use of dark patterns sometimes violates ethical standards of user-centered design [41]. These involve various deceptive techniques, such as forced actions, misleading interfaces, false urgency and scarcity, or hiding important information, all designed to achieve business goals for service providers and designers at the expense of users' privacy and autonomy. Beyond e-commerce sites, dark patterns also appear in algorithms and are integrated into smart apps (as default data-sharing options), mobile apps (through hidden settings and endless scrolling), and social media platforms (via notifications meant to increase engagement) [44].

Although it is reported to have been coined in 2010 by Brignull [37], dark patterns have since gained popularity among researchers, designers, ethicists, and policymakers who are consistently seeking ways to analyze and understand different types of dark patterns to address their ethical violations, including issues of transparency, respect for user autonomy, cognitive biases, and usability. In recent years, Brignull [11] introduced some taxonomies and classifications of dark patterns designed to manipulate users on digital interfaces, which have been expanded by studies [17, 36, 41]. For example, Mathur et al. [41] identified seven categories of dark patterns based on their empirical research of e-commerce websites: Sneaking, Urgency, Misdirection, Social proof, Scarcity, Obstruction, and Forced action. These categories describe how interfaces are designed to influence users to make decisions that benefit service providers and e-commerce merchants [41]. Furthermore, Gray et al. [17] outlined five categories of dark patterns incorporated into user interfaces: Nagging, Obstruction, Sneaking, Interface Interference, and Forced Action. Nagging involves repeatedly redirecting users away from their intended tasks or goals, often to encourage them to perform a different action. Obstructions are patterns that intentionally make it difficult for users to perform actions on platforms. Sneaking hides or disguises information that could support users in making informed decisions. Forced Action requires users to perform an action they might not otherwise choose to access a desired feature or complete a task [17]. Similarly, Luguri et al. [36] presented four categories of dark patterns including Nagging, Obstruction, Sneaking, and Interface Interference, three of which are broken down into subcategories. Specifically, Obstruction includes intermediate currency (using multiple currencies to hide the real cost), price comparison prevention (difficulty comparing prices with other products or services), and a roach motel (easy to sign up but difficult to cancel or unsubscribe). Sneaking encompasses bait and switch (expecting one outcome but receiving a different, undesirable one), hidden costs (additional charges after an initial cost is presented), sneak into basket (adding unwanted items to a shopping cart without consent), and forced continuity (automatically renewing a subscription or service after free trials without clear notifications). Interface Interference includes hidden information (making important details difficult to find), preselection (defaultly selecting unfavorable options for the user), and aesthetic manipulation (using visual elements to influence user perception and choices) [36]. Our focus for this paper is to contribute to research on dark patterns that surface as *Interface Interference*.

2.2 Dark patterns detection techniques, challenges, and impacts on user behaviors

Previous studies have examined the prevalence and effects of dark patterns on various user groups. These studies highlight efforts to identify dark patterns in user-focused technologies. While some research used manual methods to detect dark patterns, others investigated automated detection techniques. In this subsection, we discuss methodological and theoretical contributions related to these two approaches, emerging challenges, and their influence on user-technology interaction behaviors.

For instance, using a manual approach, Maier and Harr [38] employed qualitative interviews and surveys to explore how users perceive, respond to, and cope with dark patterns on user interfaces. The results showed inconsistent user awareness, with some participants able to identify dark patterns, but many failing to consciously recognize their implementation on the technology platform. The findings also revealed that users expressed strong negative emotions, including frustration, a sense of betrayal, and erosion of trust in the platform, which influenced future engagement and prompted active coping strategies such as peer warnings and critical evaluation of interfaces. These strategies often led to avoidance behaviors or migration to alternative services [38]. Albert et al. [1] used a mixed-method design combining a qualitative survey with 80 participants, experience sampling, interviews, and workshops (n = 11) to examine how user interfaces (chatbots, app notifications, self-checkout systems) mimic social behaviors. The findings uncovered some social aspects of dark pattern design, which are implemented through emotionally coercive tactics such as guilt-tripping, overbearing messaging, passive-aggressive phrasing, and false personalized care, among other contextually insensitive or skewed interfaces. They

recommended incorporating design norms that respect user boundaries, avoiding manipulative social framing, and considering the role of social behavior in technology platforms [1]. Similarly, Hettler et al. [22], conducted an online experiment with 273 participants to study how users perceive three dark patterns used as digital nudges— such as social norms cues, default settings, and scarcity warnings—in e-commerce contexts. While assessing key factors like perceived usefulness, ease of use, trust, and privacy risk, the results showed significant challenges and variations in user trust between groups exposed to default and scarcity nudges and a control group. The study offered suggestions for increasing user involvement and highlighted design strategies that promote autonomy rather than manipulation [22]. Additionally, Alharbi et al. [2] examined the prevalence and design of cookie interfaces on e-government platforms across 50 countries to determine if they comply with GDPR standards and whether they contain dark patterns. The results indicated that over 40% of the sites displayed obstruction patterns, which are design elements that make it harder for users to refuse consent, while nearly 50% featured forms of interface interference such as manipulative placement, preselected consent options, or misleading wording [2].

On the other hand, previous research has enabled scalable, automated detection of dark patterns, providing valuable support for developers, regulators, and researchers in reducing user deception and improving interface transparency. For example, Mansur and Moran [40] developed an automated system to detect design elements in Graphical User Interface (GUI) design. They introduced AidUI, an AI-powered solution that uses computer vision and natural language processing to identify dark patterns in user interfaces. Mansur et al. [40] focused on ten specific dark pattern types, analyzing both visual and textual cues in UI screenshots. The results showed that some categories of patterns were detected reliably, with F1 scores over 0. 0.82, while AidUI achieved an overall performance with a precision of 0. 0.66, a recall of 0. 0.67, an F1- score of 0. 0.65, and high localization accuracy (IoU ~ 0. 84). However, challenges persist. Since dark pattern detection relies on screenshots, AidUI may not fully capture dynamic behaviors (e. g., pop-ups triggered by user actions) or time- based manipulations. Deploying it in real- world settings would require further adaptation, higher accuracy, and better interpretability for legal purposes [40]. Mathur et al. [41] conducted a large-scale analysis of 11, 11,000 shopping websites to examine the prevalence and nature of dark patterns in user interfaces. They found 1, 818 examples across 15 specific types and seven broader categories, such as disguised ads, hidden costs, and forced continuity, which exploited users' cognitive biases. The study highlighted the widespread use of dark patterns supported by institutional and infrastructural factors. The authors proposed a taxonomy explaining how dark patterns influence user decision- making and called for tools, regulatory oversight, and further research to combat these harmful practices [41]. Similarly, Soe et al. [60] compiled a dataset of cookie banners from 300 news websites and explored automated detection of dark patterns in cookie consent banners using machine learning. Their results showed promising accuracy but also revealed challenges, such as defining and identifying dark patterns, limitations of current machine learning methods, and the difficulty of developing systems that balance accuracy with ethical considerations.

Although these studies provide some critical information to detecting dark patterns on user interfaces, our current paper advances research by among other contributions, asymmetrically examining cross-origin detection performance of human and state-of-the-art AI/LLM evaluators whose results are capable of informing the development of AI-powered design tools that are ethically aware and capable of supporting designers in creating more transparent and trustworthy Human-AI interaction and experiences.

2.3 Generative AI and dark patterns (opportunities and risks)

The emergence of AI/ML models has opened opportunities for research into tools that could assist user-technology interaction designers and regulators in safeguarding user autonomy against dark patterns. Recently, studies emphasizing how generative AI tools can help detect and fix dark patterns have gained popularity. For example, driven by rising regulatory interest, manipulative platforms, and the need to develop auditing strategies to combat dark patterns, Mills and Whittle [45] explored using generative AI to identify dark patterns in UI designs. While proposing and testing three approaches, the study found that AI Vision (which visually detects manipulative UI elements) currently shows promise, Choose-Your-Own-Adventure (which simulates user journeys across different digital skill levels) has future potential, and Decision Network (a structured model of user decision pathways) presents technical challenges needing further investigation [45]. Additionally, while examining how LLMs can detect and correct deceptive UI designs, Shafer et al. [58] demonstrate that LLMs can transform manipulative language into more transparent UI content through simple, iterative prompts. Their evaluation using GPT-4o and self-created web elements showed consistent improvements after three iterations, with 91% of deceptive elements becoming less manipulative and 96% no more manipulative than originally. This suggests that, even without prior training, AI can do more than identify dark

patterns—it can correct them, thereby improving interface clarity and protecting user autonomy [58].

On the other hand, other studies have highlighted how AI/LLMs may unintentionally reproduce and/or amplify dark UX practices that have been gaining traction. For instance, while examining the presence and types of dark patterns on 312 user interface (UI) components of e-commerce, generated across four (4) LLMs (Claude, GPT, Gemini, and Llama), Chen et al. [13] uncovered that over one-third of the UI components contained at least one dark pattern. They found that "information hiding," user action restriction, etc., are common dark patterns on those components. They also discovered that prompting LLMs with broader stakeholder intentions did not significantly reduce the reproduction of dark patterns. These results raise concerns about UI/UX design ethics and regulatory oversight, suggesting that beyond high-level guidance, ethical safeguards, better prompt engineering, and automated audits may be necessary to protect user autonomy from the unintentional reinforcement of harmful interface design practices [13]. While evaluating LLMs from five leading companies (OpenAI, Anthropic, Meta, Mistral, Google) for their ability to reproduce and detect dark patterns, Kran et al. [30] found that some LLMs showed consistent dark pattern tendencies and might be influenced by organizational policies and design values to display manipulative behaviors that favor developers' products and produce deceptive communication, among other issues. They introduced DarkBench, a comprehensive benchmark for detecting dark design patterns in LLMs, which includes 660 prompts across six categories: brand bias, user retention, sycophancy, anthropomorphism, harmful generation, and sneaking. The article also highlights technical limitations of the tool, such as challenges in testing proprietary models and scope restrictions to six pattern types [30]. Additionally, Quijada & Serezlic [53] conducted a between-subjects experiment where participants designed e-commerce and flight booking interfaces under various conditions (such as minimal prompts, ethically explicit prompts, or combined goals of maximizing conversion and ethics) using LLMs like Claude. After evaluating the outputs with Gray et al.'s Dark Pattern Ontology, the results showed that prompts focused solely on ethical design significantly reduced dark patterns. However, prompts combining ethical constraints with conversion goals often reintroduced dark design patterns, including Preselection and hidden costs in checkout, as well as faux urgency and social proof tactics in flight booking interfaces.

The results of these studies show that while generative AI can automate dark pattern detection, clearer and more context-aware prompts guide AI toward more ethical UI outputs. This highlights its potential and offers valuable insights for policymakers, AI tool developers, and UX designers aiming to ensure AI-generated interfaces are user-centric and devoid of manipulative designs that are harmful to user autonomy and agency.

2.4 Ethical and legal implications of dark patterns in interface design

Throughout the literature, researchers have explored some of the ethical and regulatory issues associated with dark patterns. For example, Herman [21] critically assessed the EU's regulatory responses to dark patterns and connected them to key legal frameworks like the Unfair Commercial Practices Directive, the Digital Services Act, and the GDPR. While emphasizing the importance of these legal tools in promoting transparency in user interface design, the author called for stronger enforcement, clearer definitions, and better policy coordination to effectively regulate dark pattern designs.

Documented evidence in recent years shows an increase in legal battles between large corporations and regulatory agencies due to the unethical use of dark patterns in consumer-focused solutions. These legal disputes often result in hefty fines and consequences. For example, in the September 2023 case involving TikTok Technology Limited and the Irish Data Protection Commission (DPC), TikTok was found liable for encouraging children to access privacy-intrusive settings through bold text in two pop-up notifications, obstructing neutral and objective choices. TikTok was fined €345 million [18]. Additionally, in January 2023, the U.S. Federal Trade Commission fined Credit Karma $3 million for using false "pre-approved" claims to lure consumers into applying for credit cards they often did not qualify for [39]. There is also an ongoing FTC enforcement action against Amazon regarding Amazon Prime subscriptions, citing five types of "dark patterns" in the complaint filed in the United States District Court, Western District of Washington, since June 2023 [18]. These and many other legal cases highlight the urgent ethical issues related to data privacy, transparency, and fairness-by-design associated with the use of dark patterns in user-focused AI/ML solutions. Such practices risk eroding user trust and can lead to data misuse, threatening online users' well-being and exposing merchants to legal scrutiny under regulations like the EU's GDPR, the California Consumer Privacy Act (CCPA), and other national and international laws.

Moreover, while highlighting issues in design ethics and regulatory compliance, especially on how public-sector platforms intended to uphold privacy and accessibility standards employ dark patterns, Alharbi et al. [2] brought to the forefront the ineffectiveness of existing regulations, as these manipulative interfaces persist despite GDPR's emphasis on transparency and user choice. Also, by framing

dark patterns as relational and ethical breaches in human–technology interaction, Maier and Harr [38] underscored the need for ethical design standards, user education, and robust regulation. This user-centered perspective enriches HCI and UX literature by linking deceptive design directly to consumer trust, digital well-being, and the sustainability of platform–user relationships [38].

2.5 Summary of gaps and our approach

Though extensive research has been conducted on manipulative design patterns in user-technology interfaces, our review of the existing literature reveals several gaps that inform our study. For example, our review identified the widespread use of multiple taxonomies with inconsistent labels and boundaries. This consistently causes cross-study challenges in understanding dark patterns, developing and deploying ethical detection tools, and establishing uniform regulations. Additionally, it is shown that generative AI can both create and identify dark patterns, although these findings are preliminary and need further validation. This suggests that while generative AI could help reduce harmful design practices, it might also increase harm and limit user autonomy. Many studies analyzed static screenshots or text, while others examined dynamic, conditional, or interaction-based dark patterns. This gap is crucial because some manipulative controls only become apparent after sequences of user actions, such as forced continuity or hidden fees. Moreover, although automated systems like AidUI, AI Vision approaches, and DarkBench have practical applications, they are limited by label scope, regional constraints, and dynamic contexts, requiring high precision, recall, and interpretability for real-world use. To advance the field, we examined cross-origin detection performance of human and AI/LLM evaluators (each evaluator's ability to detect dark patterns generated by the opposite source) to understand their limitations and future potentials. Our focus is on the category of "interface interference," which includes design tactics like preselection, emotional manipulation, choice overload, false hierarchy, cuteness, and complex language, subtle methods that influence or bias user decisions visually and interactively, as classified by Gray et al. [17] and Luguri and Strahilevitz [36]. We analyze how consistently human experts recognize these patterns in AI-generated screenshots and examine how LLMs perform when identifying similar patterns in real-world UIs created by humans.

3 3. Research methodology

The goal of this study was to examine the cross-origin performance of human and AI evaluators in detecting (i.e., classifying) dark patterns. Specifically, we assessed the performance of human UX experts when evaluating UI images generated by an LLM, as well as the performance of vision-enabled LLMs when evaluating human-generated UI images. To achieve this objective, we conducted a two-phase experiment using the framework illustrated in Fig. 1. In Phase 1, we generated a controlled set of UI images using GPT-4.1 with a Chain-of-Thought (CoT) prompting strategy and asked human UX researchers (UXRs) to classify them for the presence of dark patterns. In Phase 2, we collected a dataset of human-created UI images (including both dark-pattern and non–dark-pattern designs) from existing study and employed five multimodal/vision-enabled LLMs to perform binary classification of these images.

Our methodology was intentionally designed to avoid a fully symmetric benchmarking comparison between humans and LLMs, namely, a design in which both humans and LLMs classify UI images generated by each source. Instead, we focused on cross-origin detection performance and adopted an asymmetric design to better approximate realistic deployment scenarios. In practice, organizations may use LLMs to generate UI design elements and rely on human UX experts for ethical and quality evaluation. Conversely, organizations may deploy LLMs to assess UI elements created by human designers. We do not anticipate a common real-world use case that necessitates symmetric benchmarking between humans and LLMs. By isolating evaluative capability across origins, our asymmetric design reduces familiarity bias and mitigates the risk of artificially inflated performance resulting from stylistic alignment.

3.1 Prompt design

In the context of large language models (LLMs), a prompt is an input statement designed to condition or guide the model's behavior [4]. The AI literature identifies several prompt engineering techniques, including zero-shot, few-shot, chain-of-thought (CoT), and self-consistency prompting [12, 47, 57, 65]. Zero-shot prompting relies on the model's pretrained knowledge to generate outputs without task-specific examples. In contrast, few-shot prompting incorporates input–output examples within the prompt to facilitate in-context learning and guide model responses. A key limitation of both zero-shot and few-shot prompting is that they do not explicitly optimize the model's reasoning processes, which may result in suboptimal performance on tasks requiring complex logical inference [12, 47, 57]. Chain-of-thought (CoT) prompting was introduced to address this

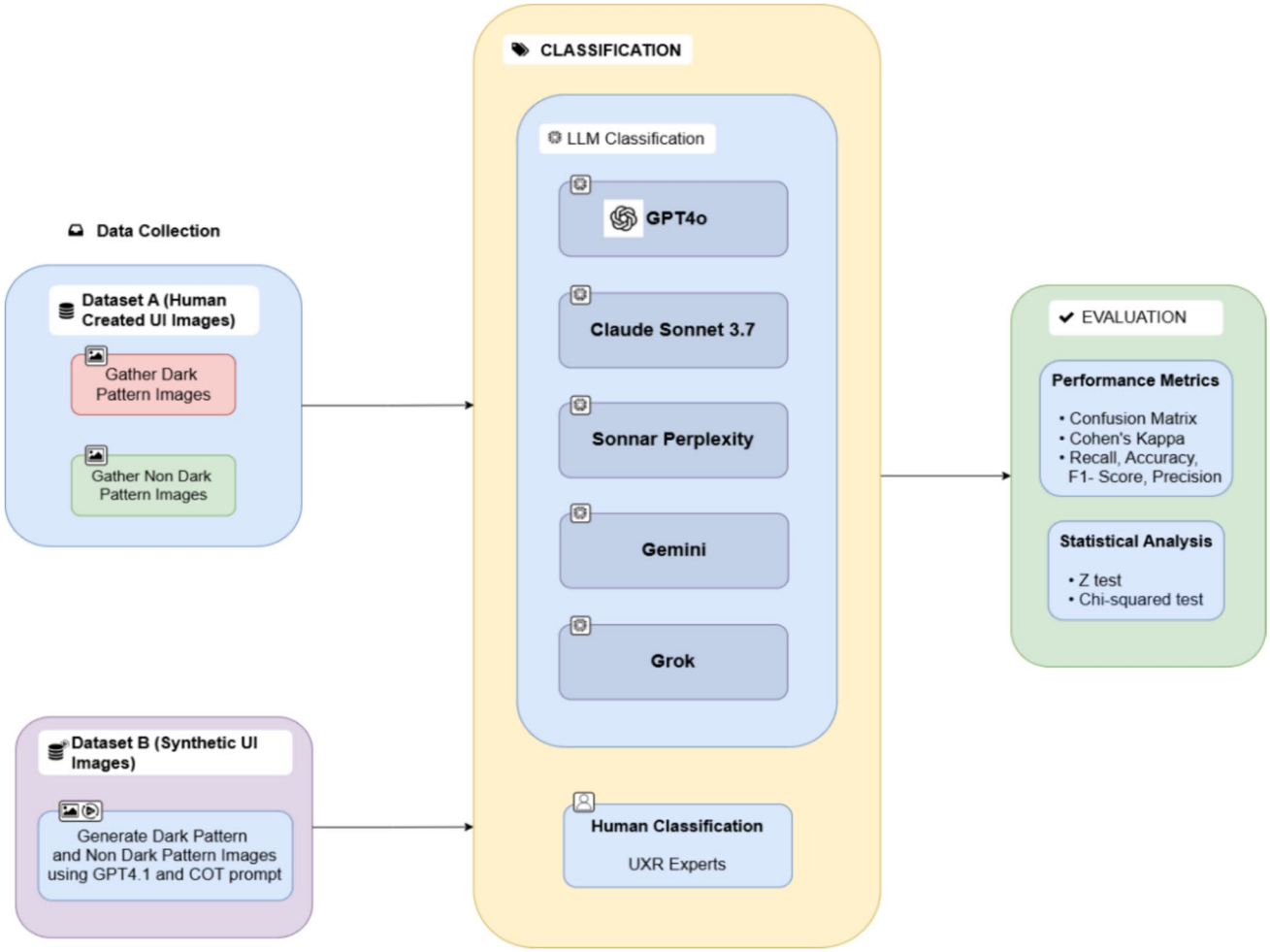


Fig. 1 Research methodology for investigating cross-origin performance of human and AI evaluators in detecting (i.e., classifying) dark patterns

limitation by eliciting intermediate reasoning steps, thereby encouraging the model to decompose and reason through complex tasks more systematically [65]. Building on CoT, subsequent techniques have been proposed to further enhance reasoning performance [7, 28, 64]. For example, zero-shot CoT demonstrates that models can exhibit step-by-step reasoning without explicit examples by prompting them to "think step by step" [28]. Self-consistency prompting extends CoT by generating multiple reasoning paths and selecting the final output through marginalization or majority voting [64]. Similarly, Graph of Thoughts (GoT) conceptualizes LLM reasoning as a directed graph, enabling non-linear reasoning, intermediate transformations, and more complex problem-solving strategies [7].

After careful review of commonly used prompting techniques in the literature, we determined that Chain-of-Thought (CoT) prompting was best suited for our study. This decision was informed by the inherently nuanced nature of dark patterns. Conceptually, dark patterns undermine user autonomy by manipulating UI elements or attributes (e.g., buttons, color, text) within the decision space of an interface, thereby interfering with users' ability to make informed and autonomous choices [27, 29, 42]. Such manipulation often relies on subtle cues or deceptive design strategies that steer users toward choices preferred by the designer [29]. In practice, two primary types of cues are commonly observed in UI elements where dark patterns are present: visual cues (e.g., layout, color, visual hierarchy) and linguistic cues (e.g., wording, phrasing, emphasis), or a combination of both [27, 29, 51, 61]. For example, an "Agree" button may be visually emphasized through color contrast or hierarchical prominence relative to a "Don't Agree" option. Similarly, urgency-based language such as "Hurry now" or "Only 5 items remaining in stock" may be used to influence users' decision-making processes.

We believe that explicitly eliciting reasoning about visual and linguistic cues embedded in UI images can enhance LLM performance and reduce hallucinations in dark pattern detection. Accordingly, we employed Chain-of-Thought (CoT) prompting to encourage structured, step-by-step

reasoning. These reasoning steps were designed to ground classification decisions in the model's interpretation of observable UI elements and inferred user intent, rather than in unsupported or speculative judgments. In developing our CoT prompt, we drew on established guidelines and best practices for effective prompt design in the AI literature. These include providing clear and specific instructions to minimize ambiguity, strategically rephrasing keywords to guide model reasoning, and incorporating examples where appropriate (e.g., in few-shot settings) [13, 17, 58]. These principles informed the design and refinement of the prompts evaluated in this study.

Prompt design (also referred to as prompt engineering) is the systematic process of creating, refining, and optimizing prompts to improve model output. In LLM workflows, prompts are typically structured hierarchically into two components: the system prompt and the user prompt [48].

At the top of this hierarchy is the system prompt, which is generally defined by the developer to initialize the model's interaction and specify its behavior, output format, and operational constraints [48, 66]. The system prompt is not visible to end users and takes precedence over the user prompt. In contrast, the user prompt defines the task-specific input for each session and may include input data or examples, particularly in few-shot settings [48, 66]. For this study, we designed and used two distinct Chain-of-Thought (CoT) prompts, namely zero-shot and few-shot CoT. The first is a few-shot CoT (see Fig. 2) which we used to generate synthetic UI images for Phase 1 of the experiment. These images were subsequently classified by human UX researchers (UXRs) for the presence of dark patterns. In the few-shot COT prompt, the system prompt specified the model's persona, constraints, and required output format, while the user prompt provided the description of the target dark pattern and included relevant few-shot examples.

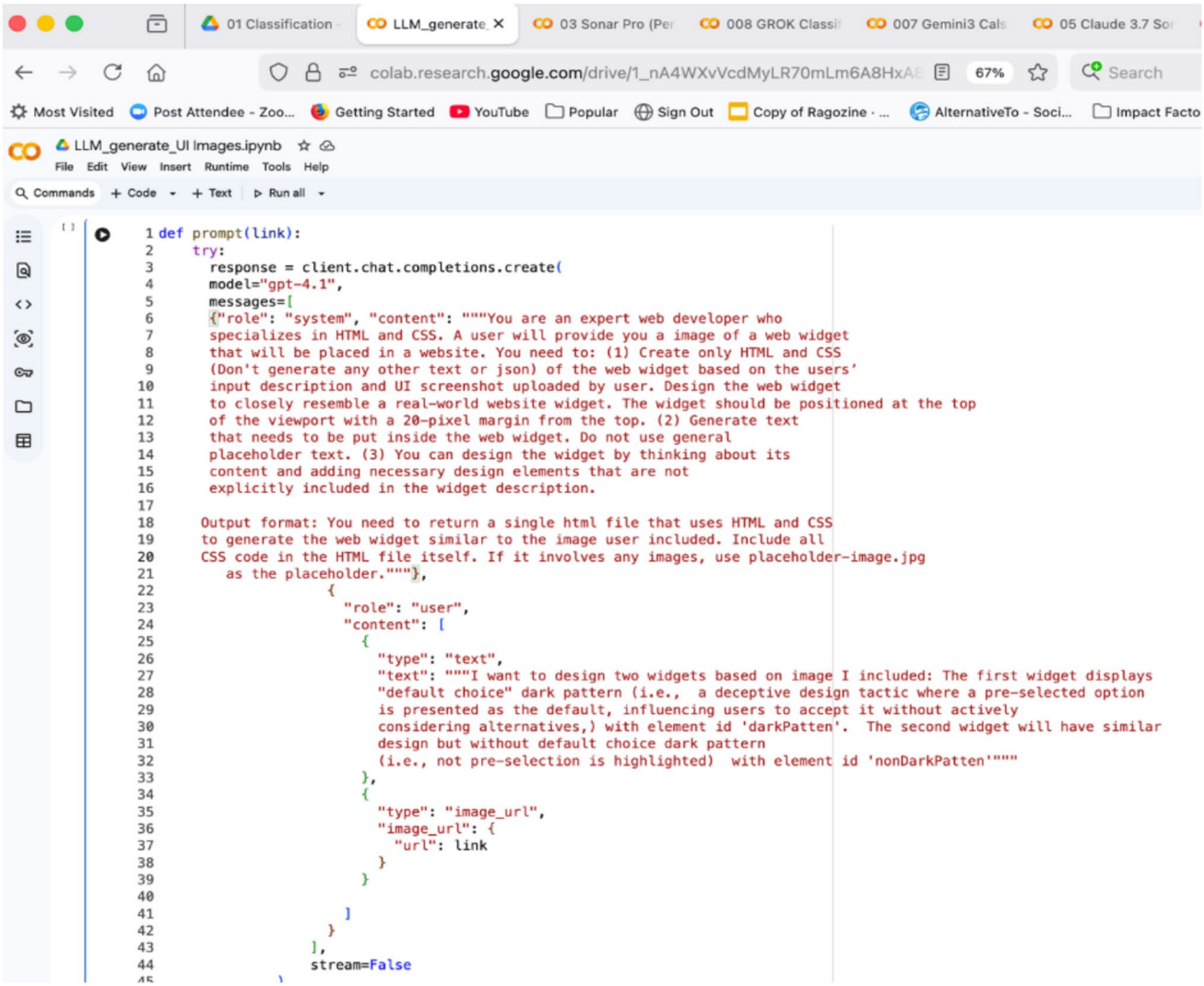


Fig. 2 Chain of thought prompt (CoT) used to generate synthetic images

We designed a second CoT prompt (Fig. 3) to perform binary classification of human-created UI images. In developing this prompt, we embedded structured reasoning steps grounded in the theoretical characteristics of dark patterns. Specifically, we operationalized visual and linguistic cues associated with dark patterns by incorporating them as explicit analytical steps within the prompt (see Fig. 3).

First, the model was instructed to infer the user's likely goal when interacting with the interface. Second, it was asked to describe the structure and presentation of choices, buttons, and informational elements within the UI. This step was intended to elicit systematic reasoning about potential visual and linguistic cues embedded in the interface.

Third, the model was prompted to determine whether any UI element—visual or linguistic—interfered with the user's ability to make an autonomous choice. Where interference was identified, the model was required to specify its type. Finally, the model was instructed to integrate these reasoning steps into a binary classification decision. For this classification task, the system and user prompts were identical. By reiterating the instructions across both prompt levels, we aimed to reinforce procedural consistency and reduce the likelihood that the model would omit critical reasoning steps. The resulting zero-shot CoT prompt is presented in Fig. 3.

3.2 Data creation and preparation

3.2.1 AI-generated UI dataset (synthetic data)

Using GPT-4.1 and Chain of Thought prompt, we generated a total of 200 UI screenshots consisting of 100 dark patterns and 100 corresponding non-dark pattern designs (Table 1).

We selected a subset of 200 UI images based on computational, methodological, and statistical considerations. Running the full experimental pipeline on the entire dataset would substantially increase GPU time and storage requirements, as our evaluation involves computationally intensive inference, validation, and metric computation. The selected subset enabled controlled, repeatable experiments within realistic resource constraints. From a statistical perspective, a randomly and distributionally sampled subset of

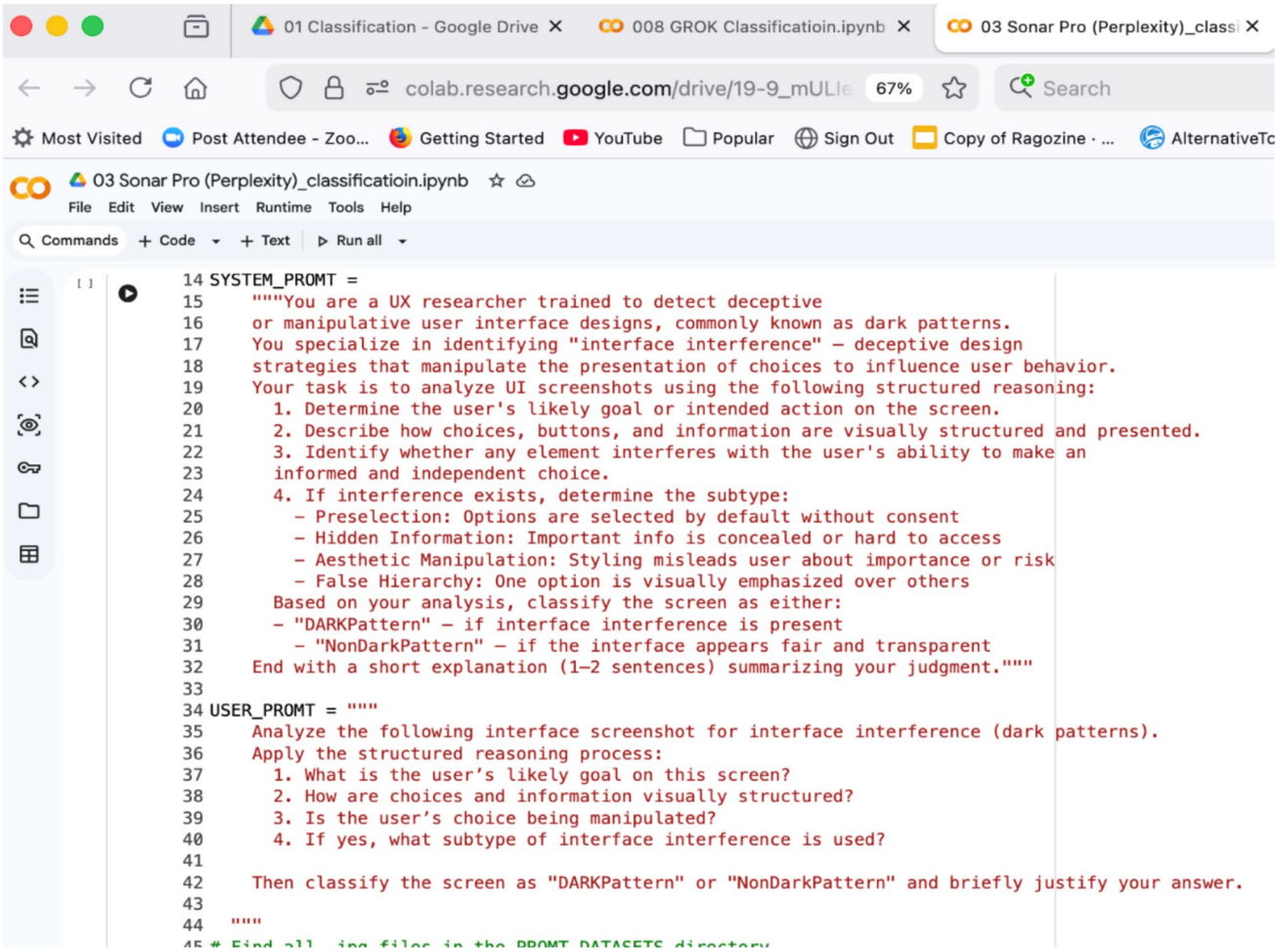


Fig. 3 Chain of thought (CoT) prompt used to classify UI images generated by humans

Table 1 Summary of dataset (Dataset samples, prompt templates, label schema, and evaluation scripts are publicly available upon acceptance via GitHub [50])

Dataset	Source	#DP	#NDP	Taxonomy	Classified by
Synthetic (A)	GPT-4.1	100	100	Interface Interference	Human Experts
Human-Created (B)	Context DP	100	100	Interface Interference (DP only)	LLMs

200 images is sufficient to obtain reliable performance estimates. As noted by Bishop & Nasrabadi, and Hastie et al. [8, 20], robust generalization depends more on representativeness than on exhaustive evaluation. Our sampling strategy preserved diversity in UI layouts and semantic categories, thereby minimizing sampling bias. Although full-dataset evaluation may yield marginal gains in statistical precision, we believe the current design achieves an appropriate balance between methodological rigor and practical feasibility.

We used Google Colab and Python to interact with the GPT-4.1 Model via OpenAI's API. An existing study [7] used LLM and CoT prompts to generate UI design elements (widgets) for an e-commerce website. We adapted this technique (and reused part of the code and prompts). However, we implemented some design controls to ensure consistency and minimize confound. First, we selected UI images of each of the 100 dark pattern examples from the human-created dataset. Then we prompted GPT-4.1 to generate HTML and CSS recreations of UI elements (widgets) for each of those images (see Appendix A.1 for the prompt). This is to ensure that both human- and AI-generated dark pattern examples share similar design components, visual style, and interaction structure. Second, we constrained all synthetic UI images to the *Interface Interference* taxonomy, following the classification proposed by Gray et al. [11]. This helped us to control taxonomy variation and maintain a focused scope for both human and LLM evaluators. Finally, for each dark pattern from the human dataset, we prompted the GPT-4.1 to generate 2 replicas of the image, i.e., a dark pattern and a non-dark pattern variant. This non-dark pattern version maintained the same general layout and context but removed the deceptive element and used an ethical design instead. Figure 4 shows a sample of a matched pair of dark-pattern and non-dark-pattern UI elements generated from GPT4.1. The reason for generating the UIs in matched pairs was to ensure that each dark pattern image has a non-dark pattern counterpart that differs only by the presence of the dark pattern. This pairing helped us to ensure that classification is only based on the presence or absence of dark patterns, and not extraneous visual differences, styles, or patterns.

We manually verified and labeled these 200 images and created our ground truth dataset (Folder A). We then created Folder A1 as a randomized mix of these images (hiding the pairing) to use in the human classification, so that experts would view images in a random order and unbiased context.

3.2.2 Human-created dataset

We used an existing dataset from literature as our dataset for human-created UI images. Specifically, we leveraged the 'AidUI' *ContextDP* dataset [39] created by human UX designers. We collected a total of 200 UI images from ContextDP, consisting of 100 dark patterns and 100 non-dark patterns (NDP) or benign. The first 100 UI images within this dataset are dark patterns within the interface interference taxonomy. But unlike the synthetic set, the other 100 non-dark pattern images were not taxonomy-matched due

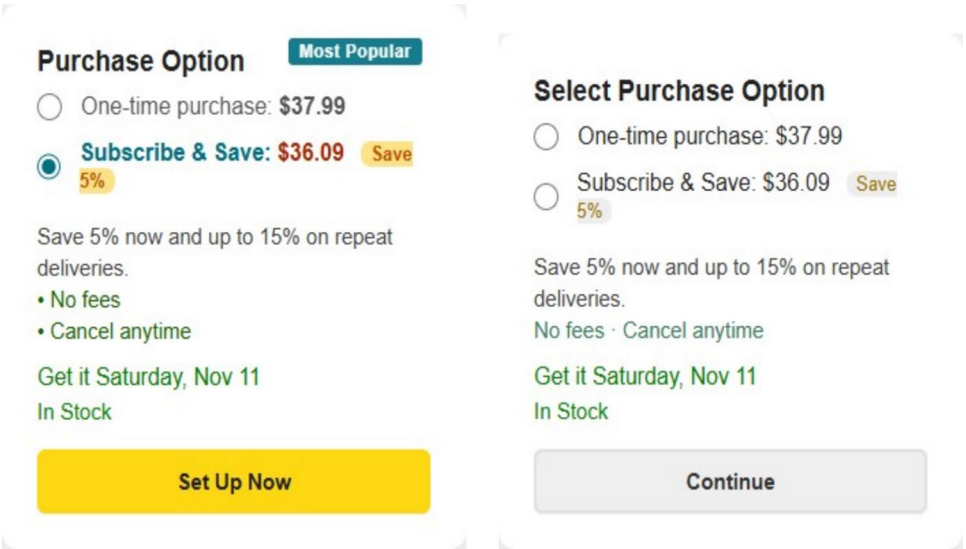


Fig. 4 Sample of matched dark pattern (left) and non-dark pattern (right) images from the synthetic dataset generated with GPT 4.1

to the unavailability of the taxonomy of non-dark patterns at the time of this study. Instead, the NDP were generally non-deceptive UI images that cut across taxonomy. We labeled this set (Folder B) as ground truth. We then created Folder B1 as a random mix of images in Folder B to use for LLM classification. We acknowledge that the lack of one-to-one matching between the dark and non-dark human images could introduce a potential confound (differences in content might provide a cue to the classifier). We discuss this in Sect. 3.4 and acknowledge it as a limitation of our study in Sect. 6.

3.3 Participants and procedure

3.3.1 Human UXR experts

Two UXR research experts (both have Ph.Ds. in Computer Science with 7+years of research experience in HCI) were recruited to classify synthetic UI images (Folder B) independently. Both were blind to the true labels but were aware that the images were AI-generated. The folder containing the randomized images (Folder B1) was sent to the experts for labeling with a spreadsheet for classification. Each expert assessed all 200 images in random order. For each image, they provided a binary classification—"1" for dark pattern present or "0" for no dark pattern**.** They also provided a self-reported confidence level (on a 5-point Likert scale from *very uncertain* to *very confident*). Each returned the spreadsheet containing their classification of the images. This process yielded two sets of 200 labels (Expert1 and Expert2). We then compared these against the known ground truth from Folder A to compute evaluation metrics and inter-rater reliability.

3.3.2 AI models for classification

We selected five contemporary LLM-based AI models for the automated classification task in Phase 2. These include *Open AI's GPT-4o, Anthropic Claude 3.7 Sonnet, Gemini 3 flash preview, Grok 4.1, and Perplexity Sonar Pro*. These models were chosen because they are known to have multimodal abilities, advanced image analysis capabilities, and above-average performance in various tasks. For each of the 200 human-created images (Folder B1), we used a CoT prompt to instruct the model to classify the image into DP (1) or NDP (0). For consistency, we used the same prompt format for all models. For our classification, both System *Prompt* and User *Prompt* are identical. By stating the instructions and steps twice, we constrained LLM to be consistent and ensured the models didn't miss any classification step. The models' textual justifications were recorded, but for analysis, we focus on their 0/1 outputs. The prompts are shown in Figs. 2 and 3, respectively while sample outputs are shown in Appendix A1. We ran each model independently on the full set. We then compared these predictions to the ground truth labels in Folder D to compute confusion matrices and performance metrics for each model.

We considered model inference settings (e.g., temperature and top-*k*) during the classification phase. However, to ensure optimal performance and maintain consistency across models, we employed each model's default inference settings. Prior research recommends using default parameters to enhance model reliability, reduce the likelihood of incoherent or inconsistent outputs, and support reproducibility [32, 63]. Accordingly, adopting default settings allowed us to standardize the evaluation procedure while ensuring fair and reliable comparisons across the LLMs.

3.4 Measures and statistical analysis

To test the three hypotheses and answer our research questions, we performed inter-rater reliability to measure the consistency and agreement between evaluators, and statistical analysis to transform our data into actionable insights. For the first research question (RQ1) and hypothesis (H_1), we used inter-rater reliability as a measure of agreement and reliability between the two UXR experts. We also computed the recall for each UXR and the average recall for both of them. To test for statistical significance, we performed a one-tailed z-test on Cohen's Kappa. For the second research question (RQ2) and hypothesis (H_2), we computed recall as a measure of performance or classification ability and used this to perform a two-tailed z-test to see if there is a significant difference between the performance and ability of humans (UXR) and AI (LLMs) in detecting dark patterns. For research questions 3 (RQ3) and hypothesis 3 (H_3), we used the values (TP, FN, FP, TN) from our confusion matrix to perform Chi-square test of independence. We briefly explain how we calculate each of the metrics used for statistical analysis in this section.

For Inter-Rater Reliability (IRR), we calculated Cohen's kappa (κ) between the two human experts' (UXR) classifications on the 200 synthetic images. Cohen's kappa is a statistic that measures inter-rater agreement for categorical decisions. A κ value of 1 indicates perfect agreement, 0 indicates no better than chance, and negative values indicate systematic disagreement [31]. We use Landis & Koch's interpretation scale [31] as a guide to interpret the κ value in our calculation. According to this scale, a κ value > 0.80 is *almost perfect* agreement, between 0.61 and 0.80 means *substantial agreement*, while between 0.41 and 0.60 can be interpreted as *moderate agreement*, 0.21– 0.40 is a *fair agreement*, 0.00–0.20 means *slight agreement*, and < 0.00 poor. We also report on the raw percent agreement. To

Table 2 IRR, percent agreement, mean confidence level (CL), and STDEV of CL for the two UX researchers

	UXR 1		UXR 2	
	Mean	StDev	Mean	StDev
Dark Pattern	4.08	0.91	4.98	0.16
Non-Dark Pattern	4.36	0.81	4.76	0.78
Both	4.22	0.87	4.89	0.51
Percent Agreement	87.5%			
Cohen's Kappa (κ)	0.75			

perform confidence analysis, we aggregated the self-reported confidence ratings for each UXR. We then calculated the mean and standard deviation of the confidence level for these ratings. Table 2 in the results section shows the Cohen's Kappa, mean confidence levels, and raw percent agreement for our study.

To calculate performance metrics, we compiled a confusion matrix for each human expert's classification of the synthetic dataset and LLM classification of the human-created dataset. Also, we tabulated True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) for each classification. Using these values, we calculated the following performance metrics: recall, accuracy, precision, and F1-score. Figure 5 and Table 3 show the confusion matrix and tabulated values of TP, TN, FP, and FN, respectively.

Table 3 Performance metrics for both human and AI classifiers

	Precision (%)	Recall (%)	Accuracy (%)	F1 Score (%)
HR1	100.0	99.0	99.5	99.5
HR2	81.1	99.0	88.0	89.2
GPT4o	75.9	85.0	79.0	80.2
Claude Sonnet 3.7	74.5	73.0	74.0	73.7
Perplexity	68.8	77.0	71.0	72.6
Gemini 3 Flash	70.3	78	72.5	73.9
Grok 4.1	80.5	70	76.5	74.9

Pooled recall for Human = (99 + 99)/200 = 0.99

Pooled recall for LLM = (85 + 73 + 77 + 78 + 70)/500 = 0.766

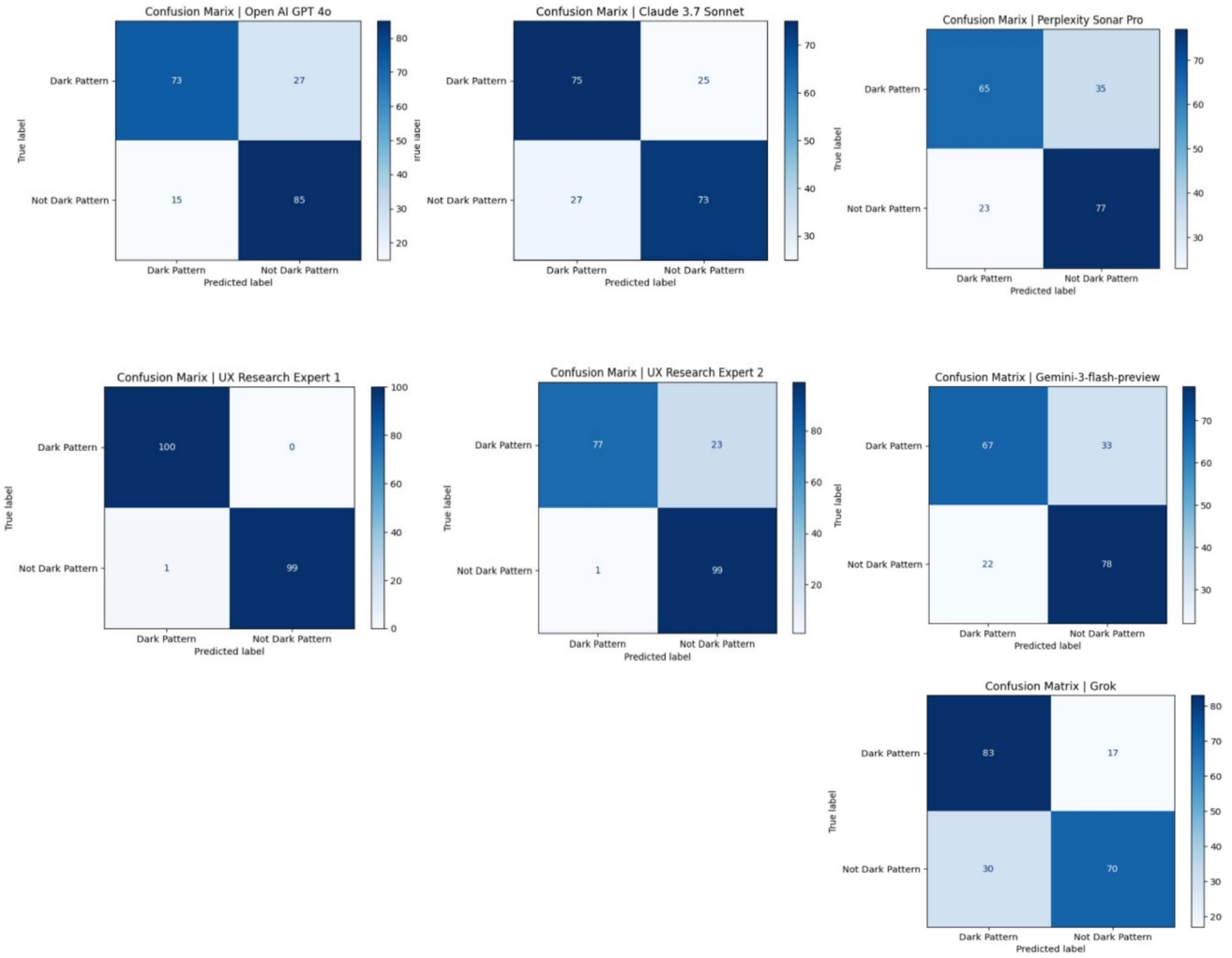


Fig. 5 Confusion matrices showing the values used to evaluate the classification performance for both LLM and Human evaluators

3.5 Avoiding confounds

The methodology was designed to minimize confounding variables. In the synthetic dataset, the matched-pair design controls for visual complexity and context. This is to ensure that the only systematic difference between each DP image and its NDP counterpart is the presence of the dark pattern. This means any difference in classification should be attributable to the dark pattern itself. However, in the human dataset, we recognized a limitation: the NDP images were not one-to-one taxonomy-matched to the DP images (due to the unavailability of NDP taxonomy).

This could introduce some confound, i.e., factors such as context, visual design, and screen type may impact the model's classification rather than deceptive design alone. This avoids this type of confound; we used only recall as a metric for our analysis. This is because recall only considers TP and FN, i.e., checking where the model classifies dark patterns without considering non-dark patterns (which are not taxonomy-matched for human-created UI images). We also acknowledged this as a limitation of our study. In the future, we would seek to collect context-matched human NDP examples to mirror our synthetic approach, thereby fully eliminating this confound.

4 Results

In this section, we present the results of our experiment and use them to answer the RQ and test the hypotheses.

4.1 Research question 1 and hypothesis 1

RQ1: How reliably and consistently do UXR experts classify AI-generated UI images as dark patterns or non-dark patterns? Hypothesis 1:

- H_{10}: UX experts will exhibit substantial agreement in classifying AI-generated images into dark vs. non-dark pattern images.
- H_{1A}: UX experts will **not** exhibit substantial agreement in classifying AI-generated images into dark vs. non-dark pattern images.

The goal of RQ1 and Hypo 1 was to investigate the agreement, reliability, and consistency of the 2 UXR experts in detecting dark patterns in AI-generated UI images. As discussed in Sect. 3.3, we computed and used the following statistics to answer RQ 1 and test hypothesis 1: Cohen's Kappa (κ), raw percent agreement, average confidence level (CL), and classification performance metrics. These statistics are shown in Table 2. The confusion matrix and performance evaluations are shown in Fig. 5 and Table 3, respectively.

Table 4 Contingency Table for the 2 categorical variables

	Humans	LLMs	Total
True Positives	198	383	581
False Negatives	2	117	
Total	200	500	700

As seen in Table 4, the κ value for both human raters is 0.75 or 75%. According to Landis and Koch's scale, this value corresponds to substantial agreement between the two raters. Also, a raw percent agreement of 87.5% shows a considerably higher agreement between the raters. All average confidence levels are at least four on a 5-point scale with all standard deviations of below 1. These results suggest that UX experts are consistent and achieve a substantial agreement in detecting or classifying dark patterns generated by AI. As shown in Fig. 5 and Table 3, the recall for both human raters is 99%, which is a considerably high recall. Recall is the ratio of true positives and the sum of true positives and false negatives (R = TP/ (TP + FN). Essentially, it measures the ability of a model to correctly identify positive cases (in this case, dark patterns) in classification tasks. A high recall means that the model or entity performing the classification task is very reliable, and vice versa. This nearly perfect recall shows that human experts can reliably detect dark patterns created by AI. In summary, and as an answer to our research questions, human experts are very reliable, consistent, and agreeable in classifying dark patterns created by AI.

In hypothesis 1, we assume that humans will have a substantial agreement in classifying AI-generated images into dark patterns and non-dark patterns. According to Landis and Kochi's scale, a substantial agreement starts from 0.61. Hence hypothesis 1 can be expressed as follows:

H_{10}: k >= 61 and
H_{1A}: k < 61.

To test this hypothesis, we perform a one-tailed z-test. We set our decision threshold at α = 0.05 for a lower-tail, and we will reject the null hypothesis (H_0) only if our test statistic (z) is less than the critical value, i.e., $z < -1.645$. Our calculations returned the following values: *Observed Agreement (Po) = 0.875; Expected agreement by chance (Pe) = 0.499; Cohen's (κ) = 0.75; Standard error (SE) = 0.0467; z-statistic = 3.01; 95% CI (0.659, 0.842)*. Given that our test statistics (z) is not less than the critical value, i.e., $z = 3.01 > -1.645$, we fail to reject the null hypothesis. In summary, two UX research experts demonstrated a substantial agreement and reliability in detecting dark patterns from AI-generated UI

images. For details of our analysis, refer to the Python code in our GitHub repo [50].

4.2 Research question 2 and hypothesis 2

RQ2

Is there a significant difference in dark pattern detection recall between human experts evaluating AI-generated UIs and LLMs evaluating human-created UIs?.

Hypothesis 2

- H_{2o}: There is no significant difference in dark pattern detection recall between human experts (evaluating AI-generated UIs) and LLMs (evaluating human-created UIs).
- H_{21}: *There is a significant difference in recall between the two evaluator–dataset combinations.* From Table 3 and Fig. 5, we observe that human experts (UXR) achieve higher scores across all performance metrics when classifying AI-generated UI images than AI does when classifying human-created UI images. The goal of research question 2 and the hypothesis is to check for statistical significance and rule out the possibility that this high performance from humans is a result of chance. As discussed earlier, while we present all metrics, we use recall for our analysis to avoid confound.

In our second hypothesis, we claim that the higher observed classification performance of humans in Table 3 is due to chance. In other words, *there is no significant difference in dark pattern detection performance between* human experts classifying AI-generated UIs and AI classifying and evaluating human-created UI images. This means that the recall of human (r_h) classifiers is equal to the recall of AI classifiers (r_{AI}). Since we have two human classifiers and five LLMs, we compute the pool recall for each evaluator and use it in our analysis, as shown in Table 3. Hypothesis 2 can be expressed as follows with a significant level of 0.05 i.e., $\alpha = 0.05$:

$$H_{2o}: r_h = r_{AI}$$
$$H_{21}: r_h \neq r_{AI}$$

To test this hypothesis, we perform two z-tests on the pooled proportion of recall. For a two-proportion z-test, we reject the null hypothesis if p-value is less than or equal to the critical value ($p \leq \alpha$). As seen in Table 4, the Humans (two UXR experts) correctly classified 198 out of 200 dark patterns, given a pooled recall of 0.99. On the other hand, AI (five LLMs) correctly classified 383 out of 500 dark patterns, giving a pooled recall of 0.766. The resulting z-statistic is 7.127, corresponding to a p-value of 1.0×10^{-12}. Given the following values from our calculation ($r_h = 099$, $r_{hAI} = 0.766$, $z = 7.127$, $\alpha = 0.05$, $p = 1.0 \times 10^{-12}$). We reject the null hypothesis and conclude that there is a significant difference in dark pattern detection performance (recall) between human experts classifying AI-generated UIs and AI classifying and evaluating human-created UI images.

5 Research question 3 and hypothesis 3

RQ3:

Do human experts and LLMs differ systematically in the types of classification errors (true positives vs. false negatives) they make when detecting dark patterns in UI screenshots?.

Hypothesis 3

- H_{3o}: The distribution of classification errors (true positives and false negatives) does not differ significantly between human experts and LLMs.
- H_{31}: The distribution of classification errors (true positives and false negatives) differs significantly between human experts and LLMs.

In RQ 3 and hypothesis 3, we aim to understand if the type of errors humans makes when detecting dark patterns (generated by AI) differs significantly from the type of errors AI makes when detecting dark patterns (created by humans). Our claim in the null hypothesis is that the distribution of errors is independent of the classifier. We use the chi-square test of independence to test this hypothesis and answer our research questions. The chi-square test uses a contingency table to check if the relationship between two categorical variables is statistically significant. The categorical variables in our study are the evaluators (human and AI) and the type of error (true positive and false negative). We use e_h to represent the type of error in humans, while e_{AI} the error in AI, Hypothesis 3 *can be expressed as follows, with a significant level of 0.05 i.e.,* $\alpha = 0.05$*:*

$$Ho: e_h = e_{AI}$$
$$Ho: e_h \neq e_{AI}$$

Table 4 shows the 2 by 2 contingency table for our study. For a CHI square test of independence, we reject the null hypothesis if p-value is less than or equal to the critical value ($p \leq \alpha$). The result of our CHI test shows $\chi^2 = 49.23$, df = 1, N = 700, effect size (φ) is 0.265, and a p-value of 2.3×10^{-12}. Since $\chi^2(1, N = 700) = 49.23$, $p < 0.001$, $\varphi = 0.265$, we reject the null hypothesis.

Therefore, we conclude that the pattern or distribution of classification errors differ significantly between human and

LLM. In other words, there is a significant difference in the type of error humans make vs the type of error LLM make when classifying dark patterns.

6 Discussion and implications

In this study, we investigated how UX research (UXR) experts and state-of-the-art large language models (LLMs) perform in the task of detecting dark patterns in AI-generated and human-created UI images. Our goal is to examine not only the relative classification accuracy between humans and LLMs but also the broader opportunities and risks that arise when generative AI systems and human experts are both positioned as actors in detecting deceptive interface design. Our use of an asymmetric evaluative design in this study was not intended to provide a direct comparison between human and AI performance in detecting dark patterns. Instead, our design seeks to approximate real-world deployment scenarios. We envision contexts in which UX professionals design UI artifacts and subsequently use AI systems to evaluate them for the presence of dark patterns, as well as human experts review scenarios in which AI-generated interfaces. We acknowledge that this type of asymmetric design may not provide a robust basis for direct comparative interpretation of dark pattern detection performance between humans and AI. A comprehensive comparative analysis would require a fully symmetric design. Such a design would involve LLMs evaluating UI images created by both UX professionals and LLMs, while human UX experts would likewise evaluate UI images generated by both human designers and LLMs. In this section, we synthesize our findings in relation to dark pattern detection accuracy between UX experts and LLMS, highlight relative strengths and weaknesses (Sect. 5.1), and reflect on their implications for human–AI collaboration and dark pattern research (Sect. 5.2) and the future of work. Finally, we discussed the study's limitations and future research directions.

6.1 Dark pattern detection accuracy: UX experts vs. state-of-the-art LLMs

In this study, our results show a clear and consistent performance gap between human experts and large language models (LLMs) in the detection of dark patterns within AI-generated user interface (UI) images. Specifically, trained UX experts achieved a near-perfect recall (99%) with substantial inter-rater agreement ($\kappa = 0.75$), which provides strong evidence that human evaluators can reliably and systematically identify deceptive patterns. This finding aligns with a long-standing tradition in usability and HCI research emphasizing the importance of expert evaluators in identifying design flaws and manipulative practices [17, 49]. By contrast, LLMs such as GPT-4o, Claude Sonnet 3.7, and Perplexity achieved a pooled recall of 78.7%. While this level of performance is promising for an emerging class of automated evaluators, statistical testing confirmed a significant performance difference between human experts and LLMs ($z = 7.127$, $p < 0.001$).

This performance gap highlights the relative strengths and limitations of each group of evaluators. Human evaluators bring a combination of domain expertise, contextual awareness, and tacit design intuition that enables them to detect deceptive intent even in borderline or novel cases. As prior research has argued, many dark patterns are inherently socio-technical in nature: they exploit human cognitive biases, cultural norms, and emotional vulnerabilities [10, 17, 41]. Human evaluators, particularly those with training in UX or HCI, are thus able to recognize cues that transcend surface-level features and instead signal manipulative intent. Importantly, in our study, human evaluators rarely missed dark patterns, recording only two false negatives across the entire dataset. This highlights not only their accuracy but also their robustness in handling subtle manipulative strategies.

By contrast, LLMs demonstrated important advantages in speed, scalability, and consistency. Unlike human evaluators, who are constrained by time, attention, and fatigue, LLMs can process hundreds of UI screenshots in minutes, apply classification rules uniformly, and avoid the variability associated with human decision-making over extended periods of decisions [5, 6]. These capabilities are particularly attractive for large-scale audits of app stores, e-commerce sites, or AI-generated designs, where datasets can run into tens or hundreds of thousands of unique interfaces. In addition, LLMs exhibit consistency across repeated tasks, where a human evaluator's attention may falter due to fatigue, boredom, or cognitive overload, an LLM can maintain stable performance. This aligns with recent discussions of LLMs as "scalable cognitive labor" capable of supporting repetitive or rule-based evaluation tasks [9].

However, our findings also reveal critical limitations in both groups of evaluators. Human raters, while highly reliable in small-scale studies, require training and calibration to achieve consistent judgments. This training is resource-intensive, and their evaluations are inevitably constrained by time and attention. At scale, human coders are prone to fatigue effects, cognitive biases, and contextual drift, all of which can undermine the accuracy of long-term auditing projects [35]. Moreover, reliance on human evaluators alone is unlikely to be sustainable in regulatory contexts, where thousands of new apps and websites are deployed daily. LLMs, in turn, struggled with subtle and context-dependent cues. Despite their fluency and ability to classify patterns

based on training data, they produced 65 false negatives in our dataset, indicating a much higher likelihood of overlooking manipulative practices when the design deviated from familiar training examples. This weakness resonates with broader concerns in the AI evaluation literature, where models tend to generalize poorly in novel or adversarial contexts [59]. For example, when confronted with culturally nuanced strategies that exploit local norms or unfamiliar interface conventions, LLMs often failed to detect manipulation. This suggests that while LLMs capture broad statistical regularities in training data, they may lack the interpretive reasoning and socio-technical awareness necessary to infer intent.

Overall, these findings suggest that LLMs cannot yet replace expert human evaluators in the task of detecting dark patterns, but they may serve as valuable complementary tools. A hybrid workflow, in which LLMs are deployed to flag potential dark patterns for subsequent human validation, could combine the scalability of automated screening with the interpretive depth of domain expertise. Such a workflow has parallels in other domains of AI-assisted evaluation, such as medical imaging, where AI systems pre-screen large volumes of scans, and radiologists provide interpretive validation [15, 26]. In the context of dark pattern detection, LLMs could handle the "broad sweep" of identifying suspicious features across large datasets, while humans could focus their expertise on resolving borderline cases, culturally sensitive cues, and regulatory implications.

This hybrid vision also aligns with recent calls in HCI and AI ethics for "human-in-the-loop" frameworks that prioritize both computational scalability and human judgment [3, 59]. Rather than framing the evaluation task as a competition between humans and LLMs, our findings suggest that the most effective path forward lies in leveraging the complementary strengths of both. For researchers, this opens opportunities to design evaluation protocols that explicitly incorporate hybrid roles, while for policymakers and industry, it suggests pragmatic pathways toward more sustainable and effective auditing of manipulative design practices at scale.

6.2 Human–AI collaboration: opportunities and risks

Our findings contribute to broader discussions about the role of generative AI in shaping and auditing digital design practices, which is beyond detection accuracy. As recent CHI work has highlighted, LLMs can accelerate ideation and evaluation tasks by offering alternative framing, reframing user interface components, and simulating stakeholder perspectives [33]. In our study, this scalability advantage suggests that AI could help build awareness of dark patterns by scanning large UI corpora and flagging recurring deceptive strategies for expert analysis. In addition, there are opportunities for positive impact, including educational use cases, such as integrating LLM-based detection into design curricula to sensitize students to deceptive practices, or organizational auditing pipelines where AI assists compliance teams in screening for regulatory violations. Such hybrid applications align with prior research on algorithmic auditing frameworks, where automation accelerates review without displacing critical human judgment [55].

However, risks are equally salient. Generative AI systems may inadvertently normalize, reproduce, or even propagate deceptive patterns if trained on datasets containing such examples [52]. Misclassifications, particularly false negatives, could result in "blind spots" where harmful patterns persist undetected, giving organizations a false sense of compliance. Furthermore, transparency and explainability remain ongoing challenges; participants in prior work expressed hesitation to trust AI-generated recommendations without clear rationales [34]. Our error distribution analysis confirms this concern, as LLMs' systematic tendency toward false negatives could undermine trust in automated audits.

We recommend that CHI communities and future work should therefore explore designs for human–AI collaboration that maximize complementary strengths while addressing limitations. This might include structured prompting strategies [19], restricted-domain fine-tuning on curated dark pattern datasets, or mixed-initiative interfaces where LLMs generate flags and experts provide iterative feedback. Additionally, research should examine how AI-driven tools can be embedded within identifying dark patterns, participatory auditing contexts, ensuring that designers, regulators, and end-users can co-interpret results.

The findings of our study have some important implications for HCI/AI ethics and policy makers, practitioners, and researchers, especially as the use of AI/LLM in UX design continues to grow rapidly. The dual role of LLMs as an enabler and mitigator of dark patterns is of particular interest, especially for policy makers and regulatory bodies like GDPR and FTC. The negligible error margin (1%) in human evaluators shows that LLMs can generate dark patterns that even human evaluators cannot perfectly detect. Moreover, UX practitioners who use LLMs as tools to facilitate UI design should be aware that such tools can generate dark patterns even when the designer is not intending to do so. There is, therefore, a need for expert evaluation of any widget or UI component produced by LLM to ensure it adheres to ethical design standards. Furthermore, regulatory bodies could, in fact, enact this as a requirement or condition for

organizations that use or intend to use LLM in UX design or for any user-centered endeavor, especially in sensitive application domains such as healthcare, finance, etc.

7 Limitations and future work

While the findings of our study contribute to ethical UX design and AI integration, it has a few limitations that readers should be aware of. At the time of this study, we could not find a classification or taxonomy of non-dark patterns. Hence, the non-dark patterns used for classification in the human-generated dataset were not taxonomy-matched. This might lead to a possible confound in the experiment. For instance, factors such as visual design, context, and screen style may impact the model classification rather than solely on deceptive design. This could make it very difficult to conclude that the only difference between dark patterns and non-dark patterns is due to manipulative design.

Our future work will focus on addressing the above limitations. Accordingly, we plan to work with other HCI professionals interested in dark pattern research to provide taxonomy or classification of non-dark patterns. Also, we plan to repeat the experiment to check if an increased dataset, more UX expert evaluators, and using finetuned LLMs would yield different results. Another area of focus for our future work is to use available datasets to train vision transformers to be more accurate in detecting dark patterns. This could provide the basis for developing AI-powered technology with advanced capabilities for auditing UI images, detecting and correcting deceptive designs on user-centered platforms.

8 Conclusion

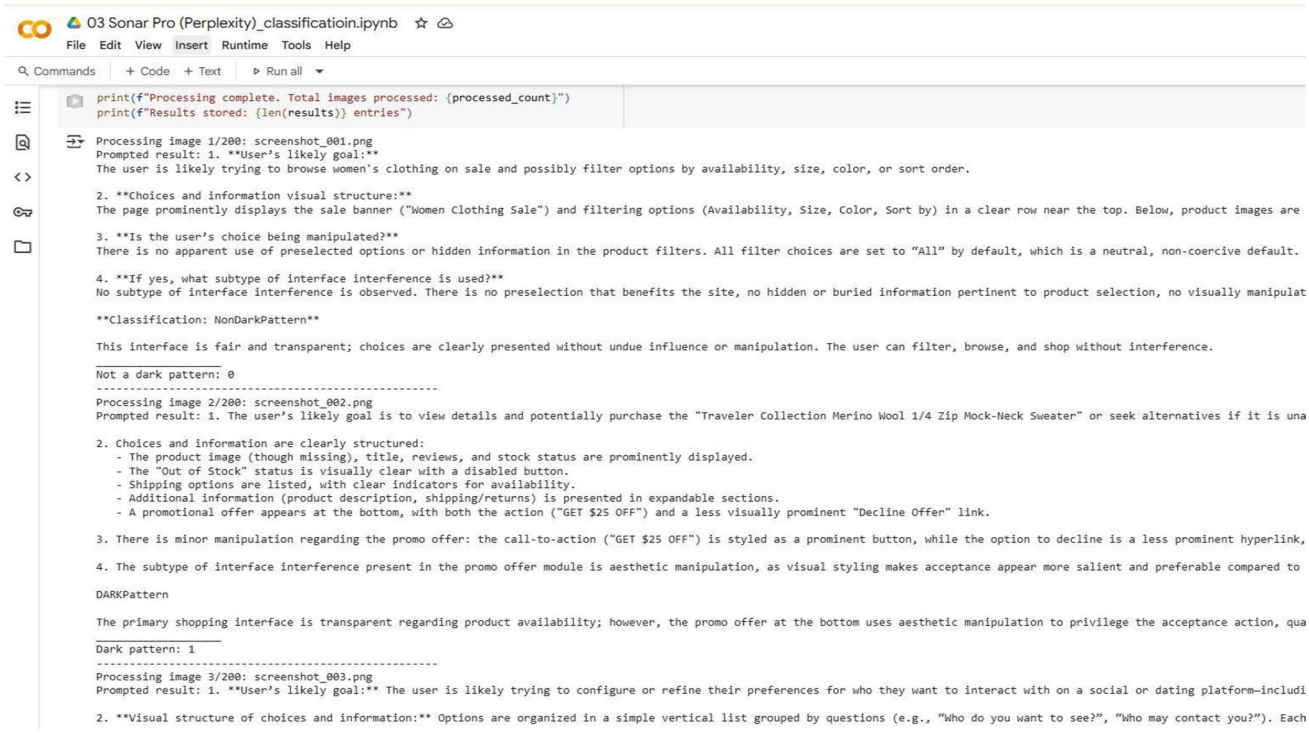
The goal of this study is to investigate the differential performance of UX experts and LLM (AI) when detecting dark patterns created by the other group within the same taxonomy. We investigated if there is a significant difference in detection performance between LLMs classifying human-created dark patterns and UX experts classifying LLM-generated dark patterns. We used a confusion matrix to measure pooled recall, distribution of error types, and interrater reliability between UX experts. We performed statistical analysis, including a z-test and chi-square test, to confirm that any observed differences between UX experts and LLM classifiers are significant. Our results show that modern LLMs can create and, at the same time, detect dark patterns created by humans. However, UX experts achieved a significantly higher performance in detecting dark patterns created by LLM when compared to LLM in detecting dark patterns created by humans. As expected, UX experts are very reliable and have achieved substantial agreement in evaluating dark patterns created by LLM. In our design, five LLMs (n = 5) were used to classify human-generated UI images, whereas two human evaluators (n = 2) assessed LLM-generated UI images. Although this configuration (2 humans vs. 5 LLMs) may appear unbalanced or unequal, it is important to emphasize that the unit of analysis for our statistical testing (the two-proportion z-test under pooled conditions) is the classification outcome rather than the individual evaluator. The z-test evaluates differences in proportions of outcomes and does not depend on the number of classifiers per se. Furthermore, the assumptions underlying the z-test are satisfied in our experimental design, including independent observations, the use of aggregated (pooled) proportions, and categorical data with a binary outcome (dark pattern vs. non–dark pattern). These conditions support the appropriateness of the chosen statistical approach.

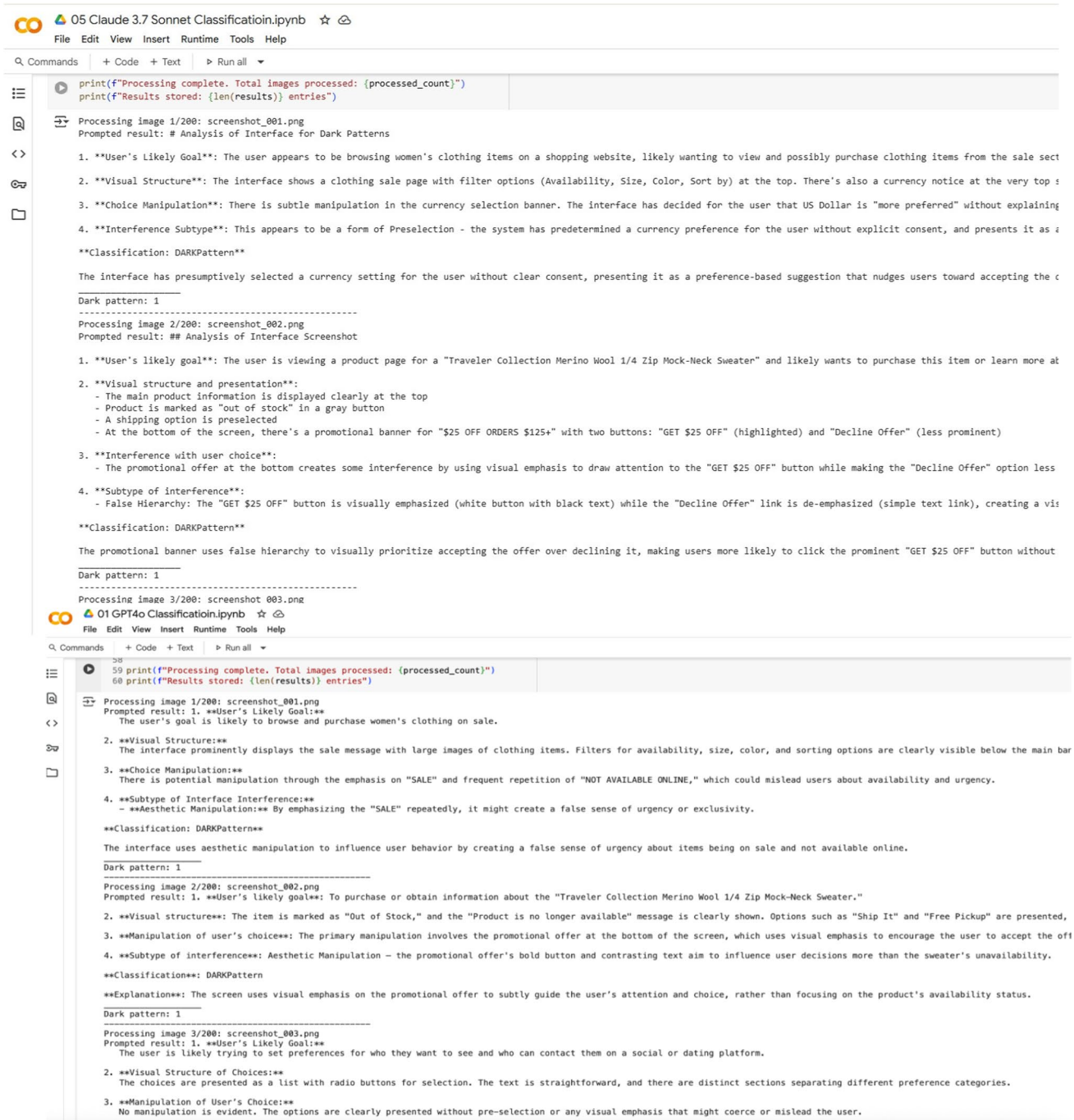
Our findings have critical implications and are of practical importance to the HCI community, including UX practitioners, regulators/policymakers (e.g., GDPR and FTC), and researchers. UX practitioners who use LLM and other AI technologies as a design tool should be aware that these tools can generate deceptive designs. Also, while modern vision-enabled LLMs show promising capabilities in various UX design tasks, they have limitations that require careful consideration before deployment. Even with the growth in AI sophistication, human experts still outperform LLM in detecting dark patterns and thus must always be kept in the loop. Yet, LLM brings its benefits, especially in speed and scalability. Hence, the role of UX auditing, ethical design, and UI compliance should be a collaboration between UX experts and LLM. Future work should create a taxonomy of non-dark patterns, train vision transformers that detect dark patterns, support UX design with more accuracy, and provide educational awareness of how LLMs both enable and hinder unethical design practices. Also, future experiments should be conducted with a balanced and increased dataset.

Appendix

Screenshot of output from classification

Below are two examples of the output of classification from GPT4o, Perplexity Sonnar Pro, and Claude 3.7 Sonnet.


03 Sonar Pro (Perplexity)_classificatioin.ipynb
File Edit View Insert Runtime Tools Help
Commands + Code + Text Run all
print(f"Processing complete. Total images processed: {processed_count}")
print(f"Results stored: {len(results)} entries")
Processing image 1/200: screenshot_001.png
Prompted result: 1. **User's likely goal:**
The user is likely trying to browse women's clothing on sale and possibly filter options by availability, size, color, or sort order.
2. **Choices and information visual structure:**
The page prominently displays the sale banner ("Women Clothing Sale") and filtering options (Availability, Size, Color, Sort by) in a clear row near the top. Below, product images are
3. **Is the user's choice being manipulated?**
There is no apparent use of preselected options or hidden information in the product filters. All filter choices are set to "All" by default, which is a neutral, non-coercive default.
4. **If yes, what subtype of interface interference is used?**
No subtype of interface interference is observed. There is no preselection that benefits the site, no hidden or buried information pertinent to product selection, no visually manipulat
Classification: NonDarkPattern
This interface is fair and transparent; choices are clearly presented without undue influence or manipulation. The user can filter, browse, and shop without interference.
Not a dark pattern: 0
Processing image 2/200: screenshot_002.png
Prompted result: 1. The user's likely goal is to view details and potentially purchase the "Traveler Collection Merino Wool 1/4 Zip Mock-Neck Sweater" or seek alternatives if it is una
2. Choices and information are clearly structured:
- The product image (though missing), title, reviews, and stock status are prominently displayed.
- The "Out of Stock" status is visually clear with a disabled button.
- Shipping options are listed, with clear indicators for availability.
- Additional information (product description, shipping/returns) is presented in expandable sections.
- A promotional offer appears at the bottom, with both the action ("GET $25 OFF") and a less visually prominent "Decline Offer" link.
3. There is minor manipulation regarding the promo offer: the call-to-action ("GET $25 OFF") is styled as a prominent button, while the option to decline is a less prominent hyperlink,
4. The subtype of interface interference present in the promo offer module is aesthetic manipulation, as visual styling makes acceptance appear more salient and preferable compared to
DARKPattern
The primary shopping interface is transparent regarding product availability; however, the promo offer at the bottom uses aesthetic manipulation to privilege the acceptance action, qua
Dark pattern: 1
Processing image 3/200: screenshot_003.png
Prompted result: 1. **User's likely goal:** The user is likely trying to configure or refine their preferences for who they want to interact with on a social or dating platform–includi
2. **Visual structure of choices and information:** Options are organized in a simple vertical list grouped by questions (e.g., "Who do you want to see?", "Who may contact you?"). Each



Acknowledgements The authors thank HCI colleagues and experts who participated in the experiment. They wish to thank reviewers and the editor for their constructive criticisms throughout the review process.

Author contributions JO contributed to the conceptual development, methodological framing, writing, critical analysis, and visual design elements. TI contributed to methodological framing, writing, and critical analysis. MN contributed to methodological framing, literature review, writing, and critical analysis. MY contributed to coding, writing and visual design elements. All the authors reviewed and approved the final manuscript.

Funding This study is partly supported by XXX Endowment provided by XXX from XXX University, XXX (anonymized for peer-review). Indiana Wesleyan University.

Data availability Data generated from this study were deposited into the GitHub database and are available at the following URL: https://github.com/AI-Dark-Pattern-Project/human_vs_llm_project.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethics approval and consent to participate Not applicable. This study did not involve human participants, animals, or sensitive personal data.

AI use We used LLM to generate UI images of dark patterns, reformat and debug Python codes, and rewrite our prompts for clarity. Also, we used ChatGPT and Gemini as tools to correct typos, grammatical errors, and get suggestions to improve our paper.

References

1. Alberts, L. et al.: Computers as bad social actors: dark patterns and anti-patterns in interfaces that act socially. In: Proceedings of the ACM on Human-Computer Interaction. 8, CSCW1 (2024). https://doi.org/10.1145/3653693.
2. Alharbi, J.A., et al.: An empirical analysis of E-Governments' cookie interfaces in 50 countries. Sustainability **15**(2), 1231 (2023). https://doi.org/10.3390/SU15021231
3. Amershi, S. et al.: Guidelines for human-AI interaction. In: Proceedings of the 2019 chi conference on human factors in computing systems (2019), 1–13.
4. Atreja, S. et al.: What's in a Prompt?: A large-scale experiment to assess the impact of prompt design on the compliance and accuracy of LLM-generated text annotations. In: Proceedings of the International AAAI Conference on Web and Social Media (2025), 122–145.
5. Becher, S. and Alarie, B.: LexOptima: the promise of AI-enabled legal systems. University of Toronto Law Journal. 75, 1 (2025), 73–121.
6. Bender, E.M. et al.: On the dangers of stochastic parrots: can language models be too big??. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (2021), 610–623.
7. Besta, M. et al.: Graph of thoughts: solving elaborate problems with large language models. In: Proceedings of the AAAI conference on artificial intelligence (2024), 17682–17690.
8. Bishop, C.M.C.C.M.: Pattern Recognition and Machine Learning. Springer. (2006)
9. Bommasani, R. et al.: On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258. (2021).
10. Brignull, H.: Dark patterns: Deception vs. honesty in UI design. *Interaction Design, Usability*. 338, (2011), 2–4.
11. Brignull, H.: Deceptive patterns (aka Dark Patterns)-spreading awareness since 2010. (2024).
12. Chen, B. et al.: Unleashing the potential of prompt engineering for large language models. *Patterns*. (2025).
13. Chen, Z. et al.: Hidden darkness in LLM-generated designs: exploring dark patterns in ecommerce web components generated by LLMs. arXiv preprint arXiv:2502.13499. (2025).
14. Conti, G. and Sobiesk, E.: Malicious interface design: exploiting the user. In: Proceedings of the 19th International Conference on World Wide Web, WWW '10. (2010), 271–280. https://doi.org/10.1145/1772690.1772719;PAGE:STRING:ARTICLE/CHAPTER.
15. Esteva, A. et al. A guide to deep learning in healthcare. Nature Medicine. 25, 1 (2019), 24–29. (2019).
16. Ghazali, A.B.E. et al.: Lightweight multi-stage holistic attention-based network for image super-resolution. IET Image Processing. 19, 1 (2025), e70013.
17. Gray, C.M. et al.: The dark (patterns) side of UX design. In: Proceedings of the 2018 CHI conference on human factors in computing systems (2018), 1–14.
18. H., B. et al.: Deceptive patterns - legal cases. https://www.deceptive.design/.
19. Han, Y. et al.: When teams embrace AI: human collaboration strategies in generative prompting in a creative design task. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (2024), 1–14.
20. Hastie, T., et al.: The elements of statistical learning: data mining, inference, and prediction. Springer (2009)
21. Herman, J.: Dark patterns: EU's regulatory efforts. Security and Privacy **7**(6), e441 (2024). https://doi.org/10.1002/SPY2.441
22. Hettler, F.M., Schumacher, J.P., Anton, E., Eybey, B., Teuteberg, F.: Understanding the user perception of digital nudging in platform interface design. Electron. Comm. Res. **25**(3), 2097–2134 (2025). https://doi.org/10.1007/S10660-024-09825-6/METRICS
23. Islam, M. et al.: Multimodal generative AI with autoregressive LLMs for human motion Understanding and generation: A way forward. arXiv preprint arXiv:2506.03191. (2025).
24. Islam, M. et al.: Unified large language models for misinformation detection in low-resource linguistic settings. arXiv preprint arXiv:2506.01587. (2025).
25. Islam, M. and Azhar, M.: Sarcasm detection in multilingual text through embedding-enhanced language models: bert variants. In: 2024 26th International Multi-Topic Conference (INMIC) (2024), 1–6.
26. Kelly, C.J., Karthikesalingam, A., Suleyman, M., Corrado, G., King, D.: Key challenges for delivering clinical impact with artificial intelligence. BMC Med. **17**(1), 195 (2019)
27. Kitkowska, A.: The hows and whys of dark patterns: categorizations and privacy. *Human factors in privacy research*. (2023), 173–198.
28. Kojima, T., et al.: Large language models are zero-shot reasoners. Adv. Neural. Inf. Process. Syst. **35**, 22199–22213 (2022)
29. Kollmer, T., Eckhardt, A.: Dark patterns. Bus. Inform. Syst. Eng. **65**(2), 201–208 (2023)
30. Kran, E. et al.: DarkBench: benchmarking dark patterns in large language models. In: 13th International Conference on Learning Representations, ICLR 2025. (2025), 64016–64036.
31. Landis, J.R. and Koch, G.G.: The measurement of observer agreement for categorical data. *biometrics*. (1977), 159–174.
32. Li, L., et al.: Exploring the impact of temperature on large language models: hot or cold? Procedia Comput. Sci. **264**, 242–251 (2025)
33. Li, S. et al.: A review of LLM-assisted ideation. arXiv preprint arXiv:2503.00946. (2025).
34. Liao, Q.V. and Vaughan, J.W. 2023. Ai transparency in the age of llms: A human-centered research roadmap. arXiv preprint arXiv:2306.01941. 10, (2023).

35. Ling, G., et al.: A study on the impact of fatigue on human raters when scoring speaking responses. Lang. Test. **31**(4), 479–499 (2014)
36. Luguri, J., Strahilevitz, L.J.: Shining a light on dark patterns. J. Leg. Anal. **13**(1), 43–109 (2021). https://doi.org/10.1093/jla/laaa006
37. Lukoff, K. et al.: What can CHI do about dark patterns?. In: Extended abstracts of the 2021 chi conference on human factors in computing systems (2021), 1–6.
38. Maier, M.: *Dark patterns-An end user perspective*. UMEA University. (2019).
39. Mansur, S.M.H. et al.: AidUI: Toward automated recognition of dark patterns in user interfaces. In: Proceedings - International Conference on Software Engineering. (2023), 1958–1970.https://doi.org/10.1109/ICSE48619.2023.00166
40. Mansur, S.M.H. and Moran, K.: Toward automated tools to support ethical GUI design. In: Proceedings - International Conference on Software Engineering. (2023), 294–298.https://doi.org/10.1109/ICSE-COMPANION58688.2023.00079
41. Mathur, A. et al.: Dark patterns at scale: Findings from a crawl of 11K shopping websites. In: Proceedings of the ACM on human-computer interaction. 3, CSCW (2019), 1–32.
42. Mathur, A. and Mayer, J.: What makes a dark pattern... dark? design attributes, normative considerations, and measurement methods. In: Conference on Human Factors in Computing Systems - Proceedings. (2021). https://doi.org/10.1145/3411764.3445610
43. McTear, M. and Ashurkina, M.: A new era in conversational AI. transforming conversational AI: Exploring the power of large language models in interactive conversational agents. Springer. 1–16.
44. Mildner, T. et al.: Defending against the dark arts: recognising dark patterns in social media. In: Proceedings of the 2023 ACM Designing Interactive Systems Conference (2023), 2362–2374.
45. Mills, S. and Whittle, R.: Detecting dark patterns using generative AI: some preliminary results. SSRN Electron. J. (2023). https://doi.org/10.2139/SSRN.4614907.
46. Narayanan, A., et al.: Dark patterns: past, present, and future. Queue **18**(2), 67–92 (2020). https://doi.org/10.1145/3400899.3400901
47. Naser, M.Z.: A review of prompt engineering techniques for large language models. Int. J. Hum. Comput. Interact. **2025**, 1–35 (2025)
48. Neumann, A. et al.: Position is power: system prompts as a mechanism of bias in large language models (llms). In: Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (2025), 573–598.
49. Nielsen, J.: Usability inspection methods. In: Conference companion on Human factors in computing systems (1994), 413–414.
50. Nwokeji, J.C. et al.: AI-Dark-Pattern-Project/human_vs_llm_project. (2025).
51. Oh, S. et al.: Uncovering the dark patterns of phishing emails: an eye-tracking analysis. IEEE Access. (2025).
52. Park, P.S. et al.: AI deception: a survey of examples, risks, and potential solutions. *Patterns*. 5, 5 (2024).
53. Quijada, O. and Serezlic, T.: Can AI avoid deceptive design?: investigating the effectiveness of prompting AI to prevent dark patterns in web design. (2025).
54. Radford, A., et al.: Learning transferable visual models from natural language supervision. Int. Conf. Mach. Learn. **2021**, 8748–8763 (2021)
55. Raji, I.D. et al.: Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In: Proceedings of the 2020 conference on fairness, accountability, and transparency (2020), 33–44.
56. Ramesh, A. et al.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125. 1, 2 (2022), 3.
57. Sahoo, P. et al.: A systematic survey of prompt engineering in large language models: Techniques and applications. arXiv preprint arXiv:2402.07927. (2024).
58. Schäfer, R. et al.: Don't Detect, Just Correct: Can LLMs Defuse Deceptive Patterns Directly?. In: Conference on Human Factors in Computing Systems - Proceedings. 25, (2025). https://doi.org/10.1145/3706599.3719683;SUBPAGE:STRING:FULL.
59. Shneiderman, B.: Human-centered artificial intelligence: reliable, safe & trustworthy. Int. J. Human-Comput. Interact. **36**(6), 495–504 (2020)
60. Soe, T.H. et al.: Automated detection of dark patterns in cookie banners: how to do it poorly and why it is hard to do it any other way. arXiv preprint arXiv:2204.11836. (2022).
61. Umar, A. et al.: Detecting dark patterns in user interfaces using logistic regression and bag-of-words representation. arXiv preprint arXiv:2412.14187. (2024).
62. Utz, C. et al. (Un) informed consent: Studying GDPR consent notices in the field. In: Proceedings of the 2019 acm sigsac conference on computer and communications security (2019), 973–990.
63. Wang, S., et al.: LLM can achieve self-regulation via hyperparameter aware generation. Find Associat Comput Linguist.: ACL **2024**, 6632–6646 (2024)
64. Wang, X. et al.: Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171. (2022).
65. Wei, J., et al.: Chain-of-thought prompting elicits reasoning in large language models. Adv. Neural. Inf. Process. Syst. **35**, 24824–24837 (2022)
66. Zhang, Z. et al.: Personalization of large language models: A survey. arXiv preprint arXiv:2411.00027. (2024).