DeceptiLens: an Approach supporting Transparency in Deceptive Pattern Detection based on a Multimodal Large Language Model

Emre Kocyigit
University of Luxembourg
Luxembourg, Luxembourg
emre.kocyigit@uni.lu

Arianna Rossi
Scuola Superiore Sant'Anna
Pisa, Italy
arianna.rossi@santannapisa.it

Anastasia Sergeeva
University of Luxembourg
Luxembourg, Luxembourg
anastasia.sergeeva@uni.lu

Claudia Negri Ribalta
University of Luxembourg
Luxembourg, Luxembourg
claudia.negriribalta@uni.lu

Ali Farjami
University of Luxembourg
Luxembourg, Luxembourg
ali.farjami@uni.lu

Gabriele Lenzini
University of Luxembourg
Luxembourg, Luxembourg
gabriele.lenzini@uni.lu

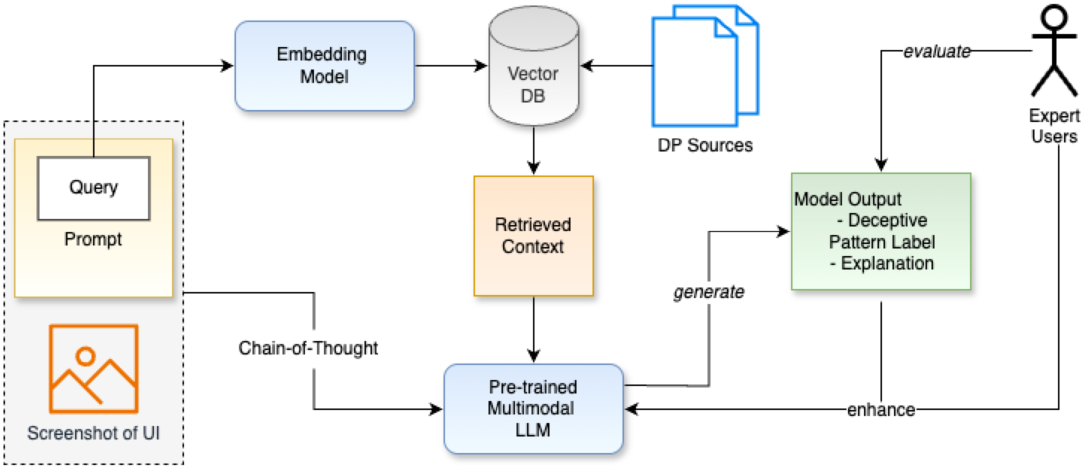


Figure 1: Overview of the deceptive pattern detection and reporting approach design.

Abstract

To detect deceptive design patterns on UIs, traditional artificial intelligence models, such as machine learning, have limited coverage and a lack of multimodality. In contrast, the capabilities of Multimodal Large Language Model (MM-LLM) can achieve wider coverage with superior performance in the detection, while providing reasoning behind each decision. We propose and implement an MM-LLM-based approach (DeceptiLens) that analyzes UIs and assesses the presence of deceptive design patterns. We utilize Retrieval Augmented Generation (RAG) process in our design and task the model with capturing the deceptive patterns, classifying its category, e.g., false hierarchy, confirmshaming, etc., and explaining the reasoning behind the classifications by employing recent prompt engineering techniques, such as Chain-of-Thought (CoT). We first create a dataset by collecting UI screenshots from the literature and web sources and quantify the agreement between the model's outputs and a few experts' opinions. We additionally ask experts to gauge the transparency of the system's explanations for its classifications in terms of recognized metrics of clarity, correctness, completeness, and verifiability. The results indicate that our approach is capable of capturing the deceptive patterns in UIs with high accuracy while providing clear, correct, complete, and verifiable justifications for its decisions. We additionally release two curated datasets, one with expert-labeled UIs with deceptive design patterns, and one with AI-based generated explanations. Lastly, we propose recommendations for future improvement of the approach in various contexts of use.

CCS Concepts

• **Security and privacy** → **Social aspects of security and privacy**; **Social engineering attacks**; • **General and reference** → **Evaluation**; • **Human-centered computing** → *Natural language interfaces*; *HCI design and evaluation methods*.

Keywords

dark patterns, deceptive design patterns, LLMs, multimodal LLMs

---

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.
FAccT '25, Athens, Greece
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1482-5/25/06
https://doi.org/10.1145/3715275.3732129

ACM Reference Format:
Emre Kocyigit, Arianna Rossi, Anastasia Sergeeva, Claudia Negri Ribalta, Ali Farjami, and Gabriele Lenzini. 2025. DeceptiLens: an Approach supporting Transparency in Deceptive Pattern Detection based on a Multimodal Large Language Model. In *The 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25), June 23–26, 2025, Athens, Greece.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3715275.3732129

1 Introduction

Deceptive (design) patterns (or Dark Patterns), which we abbreviate as DPs, are digital practices that manipulate, trick, or coerce users in ways that may harm them but serve the interests of the companies that implement them [61]. DPs are increasingly under scrutiny by the academic community that, among the others, also contributes to determine their harm [37]. Most DPs are recognized as mechanisms to circumvent the law by data protection authorities [7], consumer protection agencies [5, 31, 66] and their use is explicitly prohibited in recent regulatory efforts [19, 24–26]. In light of this, the use of some of these deceptive practices has been sanctioned by relevant watchdogs [13, 51, 76].

Since 2010, when Harry Brignull first named these practices as "dark patterns" and introduced categories such as sneaking, trick questions, and nagging [12], the list of DPs have expanded to more than 70 categories [35, 63], demonstrating that these patterns can manifest in various ways, contexts, and use cases. The usage of this practice is widespread, with the European Commission (EC) reporting in 2022 that 90% of the most popular apps in the EU contain at least one DP [29]. To curb the proliferation of these practices and equip stakeholders with tools that can contrast the existing automation asymmetry (where slow-paced manual detection methods compete with large-scale implementations of DPs), efforts have led to the development of various automated detection tools employing techniques such as machine learning (ML), natural language processing, and computer vision [16, 59, 65]. However, these tools have limited identification coverage and robustness [67], and we claim that this is because they lack multimodality, which makes them type-specific, i.e., they are working on specific aspects, such as text, images.

This study focuses on pre-trained Multimodal Large Language Models (MM-LLMs), which are built using Large Language Models (LLMs) and can simultaneously handle multiple data types, including text and image, performing at the same levels as humans in specific contexts [85]. In principle, MM-LLMs can address the limitations and robustness issues of the existing detection tools, outperform them while being more efficient to use, i.e. less labor intensive. It is not hard-coded like a rule-based model, which could enhance its robustness when addressing previously unseen deceptive design examples. Instead, these advanced language models learn from the data via the neural network components in their architecture [81]. Furthermore, they can provide reasoning behind their decisions [49], which is particularly critical for enhancing the transparency of a model that can be used as a decision-support system. Transparency is a prerequisite for the accountability of the decisions taken by its users. We propose our pre-trained MM-LLM-based approach (which we call DeceptiLens) to detect deceptive designs on the web and mobile user interface (UI), employing one of the state-of-the-art models, GPT-4o [40], along with an advanced prompt engineering technique, the Chain-of-Thought (CoT[84], and a framework that minimizes errors by retrieving relevant information (Retrieval Augmented Generation - RAG) [52].

With this research study, our aim is not only to propose an implementation that identifies a potential DP on a UI, but also to provide relevant information that can enhance user trust in the functioning and the results of the tool. Thus, we guide our model to present explanations of the reasoning behind each decision. Explanations should be clear and objective to prevent misinterpretations and enable users to make informed decisions. Thus, we encourage the model to use the concept of measurable features proposed by [47] and include the information of the reference document that is used to reduce the uncertainty within explanations [43]. We specifically aim to answer the following research questions:

- **RQ1: Is the accuracy of the MM-LLM-based approach in tasks of automated detection of UI-based deceptive design patterns equivalent to that of human experts?**
- **RQ2: How clear, correct, complete, and verifiable is the reasoning generated by our deceptive pattern detection tool for experts?**

To address these questions, we first narrow our focus to 12 DP categories and create a dataset by gathering labeled or reported GUIs with DPs. We then test whether an MM-LLM (GPT-4o) empowered with the CoT technique and the RAG framework, is capable of recognizing these patterns and providing explanations for its decisions. Later, we ask DP experts to assess if the same GUIs contain DPs and we calculate the agreement between the experts' opinion and the model's output. Beyond calculating the accuracy of the performance of the system, we also ask the same experts to evaluate its transparency which we operationalize through clarity, correctness, completeness, and verifiability of the system's explanations. In this light, our contributions are:

- The first approach that proposes, designs and implements an automated deceptive design detection based on Multimodal Large Language Model and RAG process.
- A thorough DP expert evaluation of the accuracy and transparency of the approach based on recognized metrics of clarity, completeness, correctness and verifiability.
- The creation and publication of two datasets: the first one includes screenshots of UIs and deceptive design category annotations, and the second one includes AI-based generated explanations for each label's reasoning.

2 Related Work

2.1 Deceptive Patterns Taxonomies

From the beginning of the DPs scholarship, attempts have been made to classify these various designs into a logical structure that could cover the internal similarities between different groups of patterns. The first classifications were empirical primarily by nature, including Brignull's initial classification, where he described several types of DPs' [12]. Subsequent works were dedicated to specifying the set of DPs' for specific domains, e.g., gaming [88] or e-commerce[60], to specific types of interface, or to proposing higher-level types according to type of harm they can create[45]. As DP's came into the light of public and legislative attention, several bodies presented their own short or extended classifications of DPs' in connection with the type of protection (e.g., data or customer protection) (e.g., [5, 7, 31]) they aimed to establish, creating even additional classifications.

In recent years, more and more efforts have been made to produce a taxonomy that is not just descriptive but also consolidates them into a hierarchical structure. This provides an opportunity to theoretically predict the emergence of new types of low-level patterns (the particular instances of UI-executions) based on higher-level categories. The works of [34, 35] moved towards a standardized set of categories for the DPs, which can be operationalized in terms of logical descriptions to provide a way for algorithmic analysis of DPs [35]. In addition to these studies on DP taxonomies, there are efforts [46, 47] to provide objective descriptions of specific DP types (e.g., through measurable features of the human-computer interaction) to provide actionable definitions while investigating their presence.

2.2 Automated Detection of Deceptive Patterns

As the efforts to recognize and classify instances of DPs with a shared vocabulary progress, advancements in machine learning have enabled researchers to develop AI-based models for automatically detecting DPs. Mathur et al. [60] were the first to experiment with a crawler and a large-scale approach for DP's found on e-commerce websites. Mansur et al.[59] developed a machine learning pipeline that combines computer vision with NLP for text analysis, with color brightness and spatial analysis to examine UI screenshots of mobile and web apps.

Chen et al.[16] proposed UIGuard, a knowledge-driven system using computer vision and NLP matching to automatically detect DPs in mobile UIs. The authors claim that this can be a helpful tool for designers and regulators to check for the presence of DPs in artifacts. Kirkman et al.[44] focused on the automated recognition of DPs in the specific context of cookie consent dialogs. Other approaches have privileged an assessment of the compliance with legal provisions even going beyond user interfaces, such as [11]. The authors developed an automated large-scale analysis of cookie notice compliance with the GDPR by comparing the available consent options and the declared processing purposes with the actual installation of cookies on the user device. Such tools could be categorized as SupTech (Supervisory Technology), assisting human beings in regulatory tasks [74]. Other tools have been developed, however the evaluation performed by [67] highlights that none of them can detect the totality of DPs, collectively achieving a 50% coverage. To the best of our knowledge, to date, there is only one other attempt based on LLMs for the detection of DPs, where [77] leveraged in-context learning capabilities of GPT-3 to perform a new task using inference by conditioning on a few examples.

2.3 Multimodal Large Language Models (MM-LLMs)

MM-LLMs overcome the restrictions of conversational LLMs that are capable of processing only text data by integrating multiple data types [86]. They also have the properties of LLMs,such as language generation, due to their design being based on the LLMs [54].

2.3.1 Reasoning capabilities. MM-LLMs have advanced in interpreting visual elements, transitioning from basic text prompts to more sophisticated methods for analyzing visual contexts [73]. In the context of detecting DPs in UIs, MM-LLMs leverage strategies such as encoding textual coordinates, extracting visual features, and employing visual markers to capture specific design elements [60]. These methods are complemented by reasoning frameworks like Chain-of-Thought (CoT) [84] prompting. This decomposition approach has shown improved performance in arithmetic and symbolic reasoning tasks, making it highly applicable to identifying and explaining deceptive design practices. Building on this, [90] introduced Multimodal CoT reasoning, which integrates language and vision inputs to produce intermediate reasoning steps for multimodal tasks. Their approach demonstrated superior performance on benchmarks like ScienceQA.Leveraging these reasoning techniques alongside MM-LLM-specific prompting strategies, such as zero-shot and few-shot learning, offers a unified framework for addressing the challenges of DP detection. This integration enables the model to provide clearer andinterpretable justifications, while enhancing its capacity to identify subtle deceptive elements in both static and dynamic UIs.

2.3.2 Risks and errors. The LLMs' risks of generating incorrect or biased content [69] are also present in MM-LLMs due to the architecture.There are others risks, such as sensitivity to visual noise, which affects the reasoning and overall output of the model [72]. MM-LLMs may fail for different reasons, such as CLIP missing important details [82]. The implementation domain and use cases also affect the performance of MM-LLMs. For example, MM-LLMs can produce infeasible solutions to tasks that require physical reasoning [32]. In light of these previously mentioned risks, MM-LLMs may wrongly reason about the presence of DPs and generate incorrect explanations.

2.3.3 Retrieval Augmented Generation. Retrieval Augmented Generation (RAG) is a technique that combines generative capabilities with a retrieval mechanism to enhance the factual accuracy and relevance of model outputs [52]. By integrating retrieved data into the generation process, RAG reduces hallucinations and improves explainability by providing outputs linked to specific references [41]. This is particularly beneficial in tasks requiring multimodal reasoning, where visual and textual data must be combined for accurate interpretation [90]. While RAG enhances performance by dynamically incorporating up-to-date and context-specific information, its effectiveness depends on the quality of the retrieval system and the relevance of the external data. Despite these challenges, RAG is a preferred approach in scenarios like regulatory compliance, where accuracy, transparency, and adaptability are critical [64].

2.3.4 Quality of Explanation Metrics. The ability to present clear and understandable explanations is the core of explainable AI systems [80], which enhances users' trust [38]. Providing *understandable* and *useful* explanation helps transparency in AI models, including LLMs [55]. Assessing transparency of AI models, can be measured with metrics such as *clarity*, i.e., the explanation is clearly understood, and *informativeness*, i.e., the explanation is correct and relevant [18]. [39] discusses the assessment methods for the goodness and satisfaction of the AI system explanation through *clarity* and *precision*, *completeness, usefulness, accuracy*, and *trustworthiness* metrics. [91] addressed the attributes of AI explanations in high-level categories of interpretability and fidelity through *clarity, completeness*, and *correctness* metrics. [28] have studied the effects of explanations in AI-advised decision-making processes, concluding that explanations allowing verification are considered useful and allow complementary performance; implying that a human performs better via collaboration with the AI system. In this regard, we can include the term of *verifiability*, i.e., ability of determining the correctness of the explanation [28], into the set of metrics. They also highlighted the strong relationship between verifiability and usefulness, which is an important criteria for transparent AI systems [55].

2.3.5 Human-in-the-Loop Systems. Our approach is connected to the concept of human-in-the-loop (HITL) systems, emphasizing the role of human expertise in guiding AI decision-making processes, to ensure accountability and interpretability [23]. By involving experts in evaluating the detection and classification of DPs in user interfaces (UIs), it helps the accuracy and relevance of the model's outputs. The iterative feedback loop in our approach allows human users to provide insights on the model's performance, enabling adjustments to its reasoning and classification mechanisms [2]. Integrating Reinforcement Learning with Human Feedback (RLHF) [17] would further strengthen the connection between the AI model and human evaluators. RLHF provides a framework for leveraging human input to fine-tune the model dynamically, optimizing its performance.

3 Methods

3.1 Research gaps & Research Questions

Our research aims to empirically test our proposal, which aims to support stakeholders detecting DP in a reliable and scalable manner. Our approach can be useful for identification of problematic practices in digital services [10, 62, 88]. Our tool should also be useful for regulatory bodies, seeking to investigate deceptive practices (see sec. 5.4).

Modern AI models face significant challenges, such as the need for huge amounts of structured and well-labelled training datasets and data-type-specific processing techniques. They are also restricted by their singular design, which does not accommodate the multimodal nature (text, graphics, *etc.*) of DPs. This makes ML-based systems insufficient for comprehensive DP detection. We check whether MM-LLMs can assess the presence of DP as well as human experts (**RQ1**) by bypassing the limitations of the aforementioned ML models via their multi-modal design and "emergent abilities" [87].

On the other hand, since transparency helps the auditability of systems [68], we aim for our MM-LLM-based DP detection system to provide meaningful explanations for each detection reasoning to users. Thus, we investigate whether these explanations are transparent, i.e., whether they are clear, correct, complete, and verifiable by human experts (**RQ2**). These concepts are well-defined criteria in the literature, as discussed in section 2.3.4. Moreover, clarity, correctness and completeness are commonly considered fundamental pillars of transparency (e.g., the definition of transparency in Art. 13 GDPR or Art. 13 AI Act). Verifiability of the explanations is also key, as it is linked to their usefulness: "explanations are only useful to the extent that they allow a human decision maker to verify the correctness of the AI's prediction" [28]. None of the existing approaches have ever been evaluated by DP experts in this regard.

Table 1: Target Deceptive Design Pattern Categories

1	False Hierarchy	8	Bad Defaults
2	Disguised Ads	9	Countdown Timer
3	Sneak into Basket	10	Limited Time Message
4	Hidden Costs	11	Confirmshaming
5	High Demand	12	Trick Questions
6	Low Stock	13	Activity Messages
7	Endorsement and Testimonials		

3.2 DP Dataset

We consider the 2024 taxonomy of [35] as a starting point for the creation of our dataset. We apply the granular specification as demonstrated in [47] (i.e., the low-level patterns that exhibit specific, identifiable characteristics) and list 43 categories of DPs. We selected 28 DP categories that are present in more than one taxonomy considered by [35]. Afterwards, we collected screenshots of DPs in web or mobile UIs, from sources given in Table 6. The authors included false positives in the data set. We prefer both UIs without DPs and those with DPs but labeled differently. For example, Figure 2 demonstrates an example of "Countdown Timer" that was presented in [60]. On the other hand, we included false positives, such as Figure 3, which has a similar design. We then exclude categories for which we do not find at least 5 UI examples reported as DPs. Finally, we combine similar and overlapping categories, such as "high demand and low stock" or "countdown timer and limited time messages", and obtain 9 labels covering 13 DPs categories as given in Table 1. We use this dataset as a test set to compare the experts' opinion and our model's evaluation. We present data and code[1].

3.3 Design

In order to answer the **RQ1** and **RQ2**, we design a frame shown in Figure 1 to create a DP detection tool that is capable of explaining the reasoning behind its decisions.

3.3.1 Conceptual motivation. We choose GPT-4o (https://openai.com/index/hello-gpt-4o/) as our MM-LLM due to its state-of-the-art performance in linguistic and vision-based tasks [15] and the ease of use due to its API support. It is one of the leading models in the LLM landscape [57]. In addition, it is multilingual and supports 20 languages, on top of English. Although advanced pre-trained MM-LLMs indicate significant success in various tasks, the well-known issue of generation of factually incorrect responses is still present. To address this problem, we employ Retrieval Augmented Generation (RAG) and Chain-of-Thought (CoT) techniques in our design, which are presented in Section 3.3.2. Furthermore, we force our tool to use a certain structure in its reasoning and in the output explanation, which contains (1) "DP measurable features" [47], (2) a step-by-step analysis, and (3) a reference to the documents used for reaching the conclusion.

3.3.2 Model, Prompt Engineering, RAG. We employ prompt engineering techniques, such as instructive and contextual prompts [33], and apply strategies, such as using clear language and dividing the tasks into subtasks. We construct a prompt given in Appendix C based on the selected DP target category and retrieve a relevant

---

[1] https://github.com/kocyigitemre/deceptilens.git

text, such as a description or an example from a vector database that stores the full text of the ontology of Gray et al. [35] and the seminal paper on DP attributed by Mathur et al. [61], as part of the RAG process. These were selected because they contain DP definitions and examples and systematize existing knowledge. Moreover, they are highly cited, peer-reviewed publications. We did not strive to find the perfect database, but rather to create a knowledge base that would have been familiar to the study participants who are part of the academic community. The model uses this retrieved text while assessing the presence of DP in the given screenshot. An example input-output pair of the model is shown in Table 3.

3.4 Expert Evaluation

We gathered a diverse set of experts composed of 14 academic researchers with multi or interdisciplinary expertise on DPs (i.e., HCI, UI/UX design, consumer protection, data protection, behavioral economics). They all had at least one peer-reviewed publication on the topic of DP. The study was organized in three stages. First, they were asked to assess the presence of a specific DP category on given UIs, identically to the task provided to the system. Secondly, they evaluated the explanations about the DP assessment that were generated by DeceptiLens. Finally, we asked them to comment on their evaluations in an optional interview and asked for their feedback to improve our approach. The details of each stage are given in the following sections.

3.4.1 *Task 1: Classification accuracy.*

Each UI of the dataset was assessed by three different experts who performed a binary classification considering the presence of a specific DP category on randomly distributed instances. An example question is given in the appendix D. Then, we compared our model classification with the classification of experts by taking the majority opinion into account to understand the agreement among them. We calculated the accuracy, precision, and recall scores of the model. We conducted these evaluations for both the majority opinion and the consensus, referring to cases where experts are in complete agreement. Furthermore, we measured the inter-rater agreement among the experts and examined how the examples included in our dataset and reported as DP are evaluated by the experts. We use absolute agreement scores and weighted measures of agreement: Cohen's Kappa [83] and Fleiss' Kappa [27]

3.4.2 *Task 2: Explanation Evaluation.*

Whereas the previous task was geared towards computing the accuracy of the model compared to expert evaluations, the second task was mainly dedicated to assessing the transparency of the explanations. In this task, each expert evaluated five randomly assigned explanations that are generated by DeceptiLens for each DP assessment reasoning in terms of *clarity, correctness, completeness,* and *verifiability* of the explanations. Questions were asked on a 1-5 Likert scale (strongly disagree, disagree, neither agree nor disagree, agree, strongly agree). We provided the UI and the AI explanation, asking the experts to select "strongly agree" if they found the explanation to be clear (i.e., the text is easily understood and unambiguous), correct (i.e., the text accurately reflects the truth), complete (i.e., the text contains all needed information), and verifiable (i.e., the text provides information about the process of determining correctness through concrete and observable elements). An example question of this task is given in Appendix D.

Table 2: Demographics of the interviewed participants

No.	Professional domain	Years of exp. in the domain	Years of exp. on DPs
1	UX design	20	4
5	Behav. economics	3	2
6	HCI + law	5-6	5-6
12	Online privacy	7	5
15	Design, Data Prot., AI	10	6

3.4.3 *Task 3: Interviews.*

While inviting the experts to participate in Tasks 1 and 2, we also asked them to join an optional (compensated) supplementary semi-structured interview session. Five participants joined the interviews (see demographics reported in Table 2) that lasted around 45-60 minutes. The objective of these sessions was to use an in-depth qualitative approach to find the features that contributed to the experts' perception of the explanation quality by the parameters described in the definition of Task 2. We applied a deductive-inductive approach to analyze the interviews [48]. To do that, we asked experts to reflect on the ratings they provided and compare the instances with high/low ratings to determine the main reasoning behind the ratings. We also asked experts to provide their general opinion about the characteristics of the system explanations and their perception of the system's utility. The full interview guide can be found in Appendix E. Based on that, Author 2 and Author 3 coded one interview of the dataset and then compared and discussed the coding categories. After the comparison, Authors 2 and 3 coded each 50% of the dataset; after full coding, Authors 2 and 3 discussed the final results. Similarly, Authors 2 and 3 coded the comments to Task 2 using the same code book.

3.5 Ethical and Legal Aspects

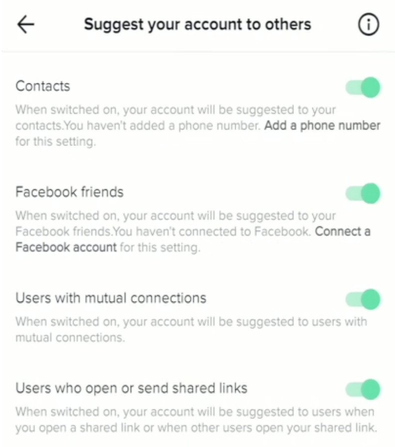
The participants were appropriately informed of the objectives and modalities of the research study and have freely consented to participate. This study has been carried out in accordance with the best practices for research with human subjects [79], the principles established in the EU Code of conduct for research integrity [1] and the recommendations provided by the EC's guidelines on the use of generative AI in research [30]. The study has received ethical approval by the Ethical Review Board of the University of Luxembourg, under authorization 22-005 DPVADM. All participants were compensated for taking part in Task 1 and 2; those that agreed to be interviewed received an extra 15 EUR voucher. Any personal data collection has been carried out in accordance with the GDPR provisions concerning lawfulness, transparency and fairness, including data minimization, purpose limitation, and data protection by design and by default. The data processing activities have been documented in the dedicated institution's register.

4 Results

4.1 Performance of DeceptiLens compared to Experts

We first checked the consensus rate of the experts in the binary classification of the UIs. % 71.62 of the instances were classified in complete agreement. In other words, %28.38 of the instances were classified by

Table 3: Example Input and Output of the DeceptiLens

Input	Output
• {*Query*}: Is there a specific DP (e.g., Bad Default) in the given UI? • Context • Tasks • Chain-of-Thought-based assessment • {*Retrieved Text*} • Screenshot:  Suggest your account to others Contacts When switched on, your account will be suggested to your contacts.You haven't added a phone number. Add a phone number for this setting. Facebook friends When switched on, your account will be suggested to your Facebook friends.You haven't connected to Facebook. Connect a Facebook account for this setting. Users with mutual connections When switched on, your account will be suggested to users with mutual connections. Users who open or send shared links When switched on, your account will be suggested to users when you open a shared link or when other users open your shared link.	**1. Measurable Features:** • Default Settings: All options are toggled on by default. • Visibility: The toggles are clearly visible, but the implications of being switched on are not immediately clear. • Language: The language used is neutral, but it requires users to actively change settings to protect their privacy. **2. Assessment:** • The default settings are privacy-invasive, as they automatically suggest the user's account to various groups without explicit consent. • Users must take action to disable these settings, which can be considered a "Bad Defaults" dark pattern. This subverts the expectation that default settings are in the user's best interest. • The interface does not provide clear, immediate information about the privacy implications of these settings, which can lead to unintentional disclosure of information. **3. Conclusion:** • The design employs a dark pattern by using privacy-invasive defaults that require users to actively opt out to protect their privacy. This aligns with the concept of "Bad Defaults" as described in the literature on dark patterns. • *Quote:* "Subverts the user's expectation that default settings are in their best interest by requiring them to actively change settings that could cause harm or lead to unintentional disclosure of information." (CHI '21, May 8–13, 2021, Yokohama, Japan, Mathur et al.)

majority vote (i.e., two out of three experts agree). We also measured the agreement between the experts by calculating the Fleiss' Kappa, which is 0.5743. This demonstrates that even DP experts had moderate agreement while evaluating whether there is a DP on the UIs. Then, we separately calculated the accuracy, precision, recall and f-1 (harmonic mean of precision and recall) scores by comparing our model's predictions with the expert opinions. This evaluation was carried out for the consensus opinion of the experts and including the majority votes as shown in Table 4. In addition to achieving promising overall scores, our model reached a recall score of %100 based on the consensus of expert opinion. This indicates that our model did not miss any UI classified as DP by the full consensus of the experts. We also calculated the Cohen Kappa score as 68.26% (*consensus*) and 57.14% (*majority*) based on the true positive, true negative, false positive, and false negative numbers for this case. It shows that there is substantial agreement between the model and expert opinion when there is consensus. Therefore, answering our RQ1, the accuracy of the approach can be considered comparable to that of human experts.

4.2 Clarity, Correctness, Completeness and Verifiability of the DeceptiLens' Explanations

Each expert evaluated five different (randomly pulled from the full set) explanations that our model generated about the reasoning behind each DP classification decision. Median values for each criteria, i.e., *clarity, correctness, completeness*, and *verifiability*, are given in Table 5. The results suggest that experts strongly agree that the overall explanations are clear and correct, while they reflect slightly weaker agreement for the completeness and verifiability of the explanations. Although the experts thought that the explanations were correct in general, they expressed lower ratings for the correctness of "Sneak into Basket or Hidden Costs" category assessment and for the verifiability of the "Countdown Timer or Limited Time Messages". However, the sample size is so small that any conclusion about the reasons for these differences would be unreliable, thus we only discuss limitations of specific explanations as raised by the participants. Answering our RQ2, we can conclude that while there is no baseline evaluation of the transparency parameters from previous works, in general, all our results are at least on the "Agree" side of the scale. In the following, we additionally provide the qualitative evaluation of these parameters via semi-structured interviews to grasp the nuances of the experts' perception of the quality of the AI-based explanations.

Table 4: DeceptiLens's Performance Considering Expert Opinions. *Consensus* refers to cases where all experts share the same opinion on the evaluation of DP. *Majority* refers to cases where there is contradictory opinion among experts. *Overall* includes both *consensus* and *majority*.

	Consensus (%)	Majority (%)	Overall (%)
Accuracy	88.67	57.14	**90.54**
Precision	86.36	61.11	89.09
Recall	**100.00**	84.61	84.48
F_1 Score	92.68	70.97	86.72

Table 5: Evaluation Results of the Explanations by Experts on a 1-5 Scale. (1: Strongly Disagree, 5: Strongly Agree)

Category	Clarity	Correctness	Completeness	Verifiability
Overall	5	5	4	4
Bad Defaults	5	4	4	4
Confirmshaming	4.5	4	3.5	4
Countdown Timer or Limited Time Messages	5	4.5	4.5	3
Disguised Ads	5	4.5	5	4
Trick Questions	5	5	5	4
Sneak into Basket or Hidden Costs	4	3	4	4
High Demand or Low Stock	4	5	5	4
Testimonials and Endorsements	5	5	4.5	5
False Hierarchy	5	5	4	5

4.3 Qualitative feedback on the explanations

4.3.1 Perceived quality of the system. Explanations were perceived as "*short, concise and overall well-written*" (P5) and useful as they provide a simple binary answer to the presence of DPs. The structured way of providing a step-by-step explanation was mentioned as one of the elements that supports understandability ("*it's clearly understandable what's the different parts are for. So the first one, it's clearly about what has been detected by the system. Then the second is how the system is kind of analyzing those elements and then [there's] the conclusion*" (P12)). It also enables the comparison between the models' explanation and the given screenshot, which together with the granular assessment of UI elements in turns facilitates the human checking of the correctness and completeness of the answer. Providing the definitions from scientific publications was also positively valued, which contributes to "*a sort of authenticity*" (P6) and helps with the verifiability of the assessment, even though looking up the actual content of the paper "*is a lot of work*" (P2).

There were also some issues in the explanations that were raised by the study participants. Some underscored an occasional lack of coherence between the three parts of the explanations ("*conflicting information between the measured features, the assessment and the conclusion*" (P15)) or the mention of elements that are not related to the identified DP, sometimes raising the fear that the model has "*cherrypicked*" from the DP definitions (P13) or adopted "*a very specific definition*" (P15) that may not be generalizable or accurate.

Participants often lamented that the model would skip certain elements or DPs, without providing a complete assessment of the UI. Others commented that it was difficult to gauge what the model assesses and what it ignores . Some deplored incomplete citations of the references, which negatively impacts the verifiability ("*I am not sure if the description in the cited paper is correct and if this paper exists*" P2). Sometimes the tool is vague about which content it considers of the mentioned publications, using expressions such as "aligning with Mathur et al." or making claims of similarity without explaining the reasons.

The experts underlined a set of issues with the correctness and completeness of the assessment. In a couple of cases,the participants noticed that the elements mentioned in the assessment did not correspond to the screenshots (e.g., "*the checkbox is not preticked*" P1;). Some definitions of DPs were not considered accurate by the participants ("The explanation uses a definition that is not necessarily agreed upon"P2) or they used inaccurate terminology. More importantly, participants observed that the model occasionally applied the DP definitions in an overly comprehensive manner (" *"may make users feel" ... is so carefully phrased that it might always be true*" P2) and that conclusions were seldomly based on an inherent assumption of "darkness" (P6) ("*the fact that the information is false is taken for granted*" P3). Certain instances contained what was perceived as mistaken assessments in the perspective of the experts ("I don't think that the red text is a tactic to draw attention away" P13). The dependence of a correct interpretation on contextual elements that were not provided to the LLM, was also raised, such as the completeness of the UI and the "*user's expertise and interpretation*" (P6).

4.3.2 The issue of detecting DPs on the UI. Many comments maintained that it is difficult, or even impossible, to assess the presence of DPs on a screenshot alone. Sometimes DPs are visible as parts of processes ("*I think my problem here was really about having just this piece of the interface and not having the whole user experience journey*" P15). In other cases, the UI is insufficient ("*It depends on the design's back-end to determine the truth*" P6) and context (or a comparison of the same UI at different times) is needed to avoid making a guess, for example when it comes to countdown timersAs a consequence, P1: "*the conclusion is an overstretch, the AI cannot know if the claim is false or truthful*".

This also raised the question of the objectivity of expert detection ("*I don't think we all agree on what is really a dark pattern [...] there is really quite a huge difference in interpretation.*" P15) and of academic scholarship ("*just because you find a paper that says A, [it] doesn't mean it's the holy truth*" P1).

4.3.3 Trust in the AI model. The issue of trusting the LLM was raised by a few AI-savvy participants in terms of interpretation of the DP publications and fabrication of content ("*you cannot really trust AI to have properly interpreted the paper and is the citation coming from it or is it made-up?*" P1), even though the doubts can be counterbalanced by the confidence in the system designers (i.e., the authors) ("*But because you and your team and you are doing this, I believe in you [...] I trust the provenance of the researcher, not of any researcher*" P12). Even though the impossibility of research objectivity was underscored earlier, trust in research publications seems to be blind for some participants: "*It looks verifiable, and since I know the paper I trust*" P5; "*for me as a researcher, I look at citations and I'm like, yeah, great [...] Trust the definitions!*" P6).

4.3.4 Target users and the overall system usefulness. DeceptiLens was perceived as useful for different users, for example for researchers to support the analysis of digital services, to build evidence for the civil rights association doing advocacy work (including supporting the general public who wants to contribute) and to educate students and younger generations. It was not considered helpful for regulators, though, as it uses a research-oriented vocabulary and does not provide legal arguments about the lawfulness of the design practices.

4.3.5 Suggestions for the system's improvement. The experts proposed several recommendations to improve the quality of the system. First, for enhancing the comprehensibility for users and filling the lacuna in the explanations, the experts proposed to reorganise and enhance the content by (a) rearranging the order of information by priority, for example by including a concise conclusion first instead of starting from the measurable elements so the user first understands the system verdict and then moves towards the explanation (which could also be given as a second layer of information on demand that could address the informational needs of different users) or order statements according to what is a more dangerous influence for decision-making; (b) avoiding repetitions since "in general the last assessment is already a conclusion" (P5); (c) separating better the design elements from the judgment on its potential outcomes on the user; (d)standardizing the citations of the references and including a link to support checking the information; (e) referring to only one source , i.e., the ontology, to provide more consistent definitions across DP and resolve the confusions of different interpretations; (f) adding the definition of the DP to the explanation; (g) improving the graphical presentation by e.g., colorcoding the result of the assessment so that the results of the assessment are clear at a glance even without reading the whole text.

In response to the question of verifiability of the AI assessment, experts proposed to show the users a step-by-step analysis or a decision-tree interpretation of analysis, combined with confidence rates of the element being a DPs. Finally, one of the experts proposed that the system should be able to interfere with the proposed interface, showing how the design without the DP could look like.

5 Discussion

5.1 General performance of the system

The overall results about accuracy, clarity, completeness, correctness and verifiability are very remarkable. Completeness ratings were not always very high because the participants expected the assessment to be about all the DPs in the interface, whereas the LLM was instructed to look for the presence of specific DPs. When asked whether any other criteria would have been useful to assess the performance of the system, the experts found that these metrics were sufficient.

5.2 How might the presence of DPs reliably assessed by human beings and by MM-LLMs?

We observed that even experts do not reach full consensus on detection tasks: only 71.62% of the UI examples were classified in complete agreement. When we investigated the reasons through the qualitative analysis, we observed two main causes. First, UI screenshots alone are often insufficient to assess the presence of particular types of DPs: contextual information, user journeys and code is also needed. Other types of data, such as web source files of the related page, or a video record that shows the interactions are needed to be sure about the functionality of the buttons. Second, the lack of objective descriptions of the DPs types causes different interpretations. For example, a clear indication of an advertising objective must be present on the UI, otherwise it can be considered as a "Disguised Ad" according to the definition contained in [35]. However, one of the experts remarked , even when the advertising mark is not clearly visible, contrast is sufficient to distinguish the advertisement from the main content of the web page. As a result, two out of three experts classified an instance that was reported [2] as an example of DP as non-DP. Moreover, some participants explicitly stated that even the expert classification is not objective and pointed to individual characteristics of users that may also influence such classification (e.g., "difficult language" is not an absolute value and rather depends on the actual user).

These results point to structural problems of the current state-of-art DP definitions and the methods for analysis. First, in certain cases, the mechanics of DP necessitates the user-computer interaction beyond static UI representations. Therefore, including statical examples in the datasets of DP creates uncertainty both for humans' and for the model's interpretations. Second, in many cases, the answer about DP presence can be represented as a continuum of options rather than a binary yes/no classification. Generally speaking, there are certain patterns that may be manipulative in one context but not in another (e.g., adding steps as a friction design pattern may be a legitimate strategy to slow down automated behavior and bolster reflective thinking to strengthen security [22], but it is demonstrated that it increases cookie consent acceptance rate when it is associated with consent refusal [6]), while certain DPs (e.g., DPs that omit relevant information) may intrinsically be more difficult to spot than others [9]. The ease of "visual detection" for human beings has also been pointed to as one of the reasons why many legal proceedings against illegal deceptive design practices only address issues at the interface level and disregard the underlying code [51].

An important measurement was that the recall score of our approach achieved 100% when experts unanimously agreed on the presence of DPs without any objection. This indicates that DeceptiLens did not miss any DP compared to the experts. However, when there are contradictory opinions among the experts, it is plausible to observe this situation reflected in the DeceptiLens' behavior. This issue also raises another discussion point: "Can we use the dataset to train an AI-based DP detection model even when the experts do not agree on it?". One solution to this would be to only include data where the label has full consensus. Another observation related to this was the classification of a screenshot, which was presented as an example of DP in a well-known academic paper [60] published in 2019. The experts labeled it as non-DP while the model identified it as an example of "Hidden Costs" consistently with the aforementioned paper. This situation brings up the following question: "Should the previously collected data still preserve its label and can it be reliably used to train or instruct the AI models?". This should be critically considered, otherwise, no matter how reliable the data source, as observed in this case, it may no longer be a DP and mislead the model. A difficulty in this respect is that, at date, there are almost no curated, publicly available datasets of DP instances that have been carefully assessed by DP experts based on the current available knowledge produced by DP scholarship. With our dataset contribution, we seek to bridge this gap in current research, even though we acknowledge that this is just a first step in this direction.

5.3 How might we make the explanations more reliable while avoiding overreliance?

Experts generally agreed that the system provides clear, transparent, and verifiable explanations about its reasoning and results. This is

---

[2] https://darkpatterns.uxp2.com/pattern/relay-for-reddit-ads/

a crucial system design consideration that the other AI-based approaches mentioned in Section 2.2 did not engage with. Yet, study participants still acknowledged some limitations to verifiability. We implemented verifiability through reference to the scientific sources and reference to the measurable elements in the interface, which could be the basis for DP implementations. While the experts mentioned that the latter mostly worked and helped to catch the system's mistakes, the former occasionally became a source of confusion. The inclusion of bibliographic citations could even be perceived as an "anti-transparency" and authority cue [56] that may cause automation bias - an overreliance on decision-support systems' output without looking for or simply ignoring counterfactual information [20]. Experts expressed doubts that DeceptiLens users would check the lengthy, complex publications themselves and would rather blindly trust the "scientific sources" mentioned by the AI system. We add RAG process by including another fully independent component of the DeceptiLens' architecture, responsible for retrieving most relevant scientific sources from our database to minimize risks of made-up scientific references in the explanation of the core MM-LLM.

Since adding explanations seems to increase human overreliance on the system [4], an approach that has been proposed to counter automation bias is the implementation of cognitive forcing functions [50]. Namely interventions that disrupt automatic, heuristic-based reasoning and engage the human being in reflective thinking [14], based on the dual-process theory [42]. For example, asking the user to make a decision before being shown the AI suggestion, delaying the AI answer so that the user can form their own opinion first and allowing the user decide whether they want to look at the AI recommendation, may reduce overreliance at the expense of acceptability [14]. As noted by the authors, beyond improving the intrinsic quality of explanations, efforts should be devoted to designing an interaction between AI systems and humans, to ensure that they make meaningful use of the explanations and make informed decisions. The future versions of the systems should not only perform on a considerable level, but it should also have mechanisms for facilitating user reflection. For example, the system can ask the user about their perception of the interface and facilitate a step-by-step reasoning [21].

5.4 How might we enhance the approach?

5.4.1 Enhancing the Approach for Researchers' use. We believe there are several ways to increase the transparency of the assessment results presentation. One of them includes showing the systems' (un)certainty about the results. Recent studies in this area showed that acknowledging the degree of the system's certainty can raise the trust in the system and help users make better decisions [4, 71, 89]. Furthermore, sharing the test dataset, limitations, and risks while publishing the evaluation results of the LLM-based DP detection system enhances transparency. [55].

Another idea is to implement the ability to switch between different guiding methodologies used for the system's decisions and to show the user the spectrum of outputs proposed based on the difference. It will help to acknowledge methodological differences in the DP scholarship and provide a nuanced understanding of it, leaving to the user the decision.

A more thorough analysis of the various uses-case contexts should also be carried out, in order to address the different informational needs of stakeholders. For instance, some participants suggested to provide a clear, concise, colorcoded conclusion upfront and the details as a follow-up that can be explored on demand. This could be useful especially when researchers intend to inspect digital services at scale: they could first check if the AI system answer corresponds to their expectation and only if it doesn't examine the explanation. In the end, to quote one of the interviewees, "making a better explanation would really depend of the context in which the explanation would be given" (P15).

5.4.2 Enhancing the Approach for Independent Authorities and the General Public. The experts mentioned that the perceived quality of the system's output can be very context-dependent, as the end-users can have different professional backgrounds and knowledge about DPs. Therefore, the following iteration of the tool should be specifically tailored to the context of use. Based on the current evaluation, we identify two additional use cases: a regulatory authority tool for checking compliance; and an educational tool to raise awareness.

Even though we developed DeceptiLens to specifically aid researchers (a decision that is reflected in the design of the system), since DPs often constitute unlawful practices (such as consumer protection), we plan that future versions may be used by regulators to support their enforcement activities related to monitoring and detecting wrongdoing in digital products and services (i.e., EnfTech or Enforcement Technology [74]). In order to do so, there are challenges that need to be addressed. First, the definitions that were selected for the study [35] do not provide legal arguments. Thus there would be a further step that needs to be made: for example, the "Bad default" DP should be mapped to privacy-invasive defaults that are relevant under the GDPR and its national implementations. This information can be enriched with official interpretation guidelines such as the EDPB's on DPs that state that if one option among many is highlighted, then it should be "the most restrictive one regarding personal data" [7, p. 23].

Another database for the RAG should then be implemented, but it would not be a trivial task. The first question would concern selecting relevant documents, spanning across regulations, case law and official guidelines. One of the challenges may be the representativeness of this dataset: for example, [51] show a skew in enforcement decisions towards "visible" UI elements, which neglects other more covert deceptive techniques. Moreover, sometimes guidance on the same design differs across jurisdictions [6], which makes it hard to provide a uniform source of knowledge. Thus a second key aspect would be how to integrate divergent information sources. An idea would be to include multiple vector databases in the RAG process and to utilize LLM-based multi-agent systems, which can provide individual specialized agents that can collaborate effectively [53]. One of the participants with previous experience in a Data Protection Authority (DPA) suggested focusing on specific cases, such as the assessment of the legality of cookie consent banner design [8].

Then, it should be carefully analyzed if enforcement agencies would use such an AI-based tool and if so, according to which safeguards (e.g., only with certain demonstrated performances levels?). Most importantly, it would need to be carefully defined who would use the tool and at which stage of the enforcement process and for which objective: this would also inform the implementation of strategies that lower overreliance on the system, as mentioned earlier.

In contrast, for the general public use, it may be necessary to streamline the explanation and present the decision, even starting from the system verdict and showing the explanations supporting it. Additionally, it can be useful if the system also proposes several ways which can help eliminate the DP from the interface and replace them with a "fair" [70, 75, 78] or "bright" design pattern [36], in line with the interventions proposed by Lu et al.[58]. However, at date, whereas there is an increasing agreement on what a DP is, there is still no common understanding of what a "fair" or "bright" design pattern is. Finally, since our experts have already reported that the system can have an educational effect in helping users learn more about DP, it would be useful if the system could implement more explicit educational strategies and leverage different types of motivation to facilitate learning about DP [3].

6 Compliance with the AI Act

Would our detection tool be subject to the requirements imposed by the AI Act (Regulation 2024/1689)? If this tool was intended for research purposes, it would fall under the research exemption (Art. 2(6)), but would need to respect the key principles for developing trustworthy AI and of research ethics (see Sec. 3.5). However, if the intended use was for enforcement agencies to support their decision-making and investigative process in supervisory tasks, then we would need to determine if the system would be considered a high-risk system under Art. 6(2) (Annex III(6)). This applies only if there is a "significant degree of power imbalance" in the actions of enforcement agencies that have "adverse impacts on fundamental rights" (Rec. 59). Since consumer protection and data protection authorities only investigate administrative offenses by organizations, our analysis suggests that our tool would not be classified as a high-risk system.

Nevertheless, enforcement agencies that adopt the detection tool would need to ensure that their staff reaches "a sufficient level of AI literacy[...]" (Art. 4) based on their technical knowledge, experience, education, training and the context of use. For example, the training should encompass the limitations and the risks of overreliance on the tool's outputs. If, instead, the tool was made available to the public to strengthen their ability to recognize DPs and it was not evident that it is an AI, the transparency obligations under Article 50(1) would apply.

7 Limitations

We inherit the cons due to using MM-LLMs, and we relied exclusively on the UIs to detect DPs. As we discussed in Section 8, the evaluation of certain DPs requires additional information, e.g., html of the web page or user journey, which we did not include. We agree that inputs should go beyond the UI. We took the recent ontology [35] as a reference and did not consider the types that do not exist in it. Since we created our test dataset by collecting reported DP examples, we could not include the categories that do not have sufficient reported examples. In parallel with this and considering the human resources, the size of the dataset is kept restricted. Another important limitation is that we did not include all the DP experts. Although the experience and focus area of our experts are diverse, they are researchers or have studied DP problems from a researcher perspective. Moreover, we asked experts whether there is a specific DP instead of asking any DP in the given UI. We are aware that the latter is valid in a real scenario. However, we prefer the former for the simplification of tasks for human experts and measurement of the comparison between the model and experts.

8 Future work

Beyond the recommendations for various users illustrated in 5.4, there are several manners in which we believe the present approach can be enhanced. A future development concerns the implementation of instruction fine-tuning of the core model based on the experts' feedback that we gathered. First, the examples annotated with the consensus of experts could be given to the model as references. Second, rules that are directly relevant to the quality of the explanations could be added within the prompt design. For example, the presentation of the bibliographic citation can be easily standardized, while various layers of explanations could be presented (e.g., a first concise informational layer about the conclusion and other layers that provide details about the reasoning, the definitions and the sources used). Third, the dataset should be enriched with other types of data (e.g., the user journey, metadata, HTML code), as UIs alone in many cases are insufficient to reliably establish the presence of a DP. Fourth, the dataset should be expanded with other types of DPs that are not currently included, in order to enhance the ability of recognizing DPs and the usefulness of the system. Then, it would then be meaningful to carry out a benchmark test that compares the "enhanced" explanations with the explanations produced in the present study with an expert assessment.

9 Conclusion

In this study, we propose an MM-LLM-based DP detection approach, which we call "DeceptiLens". We include recent techniques, such as RAG and CoT, to increase the DP recognition skills on the UIs and quality of the explanations that are generated by the DeceptiLens for explaining the reasoning behind DP assessment. To test our approach, we collected screenshots of UIs, which are reported as DP in the literature and online sources, added false positive examples, and created a dataset that contains 12 types of DPs. We used the DeceptiLens to assess the presence of specific DP types on the instances of our dataset and asked to generate explanations for each of its decisions. The same instances were then evaluated by the DP experts, and we measured an accuracy of 90.54% of our approach compared to the majority opinion of the experts. We achieved the recall score of 100% when there is a consensus of expert opinion. This indicates that our model did not miss any UI classified as DP by the full consensus of the experts. The experts also agreed that, in general, the explanations of the DeceptiLens are clear, correct, complete, and verifiable. We also discussed important observations, such as instances that experts classified as non-DP even though they were reported as DP in the literature and how to enhance the accuracy and reliability of results. In addition, we provided recommendations aimed at enhancing the usefulness of this approach for various users, i.e., researchers, independent authorities and the general public. Finally, we curated a dataset that contains DP examples that were classified in full agreement of the experts and the DeceptiLens-generated explanations. We present the code and dataset as open.

Author Contributions

Emre Kocyigit: conceptualization; formal analysis; investigation; methodology; data curation; software; validation; visualization; writing – original draft,review & editing. **Arianna Rossi:** funding acquisition; conceptualization; investigation; methodology; data curation; validation; writing – original draft, review & editing. **Anastasia Sergeeva:** investigation; methodology; validation; writing – original draft. **Claudia Negri-Ribalta:** methodology; writing – review & editing. **Ali Farjami:** methodology; writing - original draft. **Gabriele Lenzini:** funding acquisition; project administration; supervision; conceptualization; writing – review & editing.

Acknowledgments

This work was funded as part of the Decepticon project Decepticon (grant no. IS/14717072) supported by the Luxembourg National Research Fund (FNR). It was also funded by REMEDIS project "National Fund for Scientific Research (FNRS)" and the PNRR/Next Generation EU project "Biorobotics Research and Innovation Engineering Facilities "IR0000036" – CUP J13C22000400007". We thank the dark pattern experts that participated in the study including Kerstin Bongard-Blanchy, Cristiana Teixeira Santos, Silvia De Conca, Estelle Harry, Gunes Acar, Marie Potel, Lorena Sanchez Chamorro, Thomas Mildner, Joanna Strycharz, and others who contributed their time and insights. A special thanks goes to Irina Carnat for our fruitful discussion on the application of the AI Act to this use case.

References

[1] ALLEA. 2023. *The European Code of Conduct for Research Integrity - Revised Edition 2023.* ALLEA - All European Academies, DE. https://doi.org/10.26356/ECoC

[2] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *AI magazine* 35, 4 (2014), 105–120.

[3] Pooria Babaei and Julita Vassileva. 2024. Drivers and persuasive strategies to influence user intention to learn about manipulative design. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 2421–2431. https://doi.org/10.1145/3630106.3659046

[4] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 81, 16 pages. https://doi.org/10.1145/3411764.3445717

[5] BEUC. 2022. "Dark patterns" and the EU consumer law acquisition. https://www.beuc.eu/sites/default/files/publications/beuc-x-2022-013_dark_patters_paper.pdf Last accessed 9 January 2023.

[6] Nataliia Bielova, Cristiana Santos, and Colin M. Gray. 2024. Two worlds apart! Closing the gap between regulating EU consent and user studies. *Harvard Journal of Law & Technology* 37 (2024), 1295–1333. https://jolt.law.harvard.edu/assets/articlePDFs/v37/Symposium-12-Bielova-Santos-Gray-Two-Worlds-Apart-Closing-the-Gap-Between-Regulating-EU-Consent-and-User-Studies.pdf

[7] European Data Protection Board. 2023. *Guidelines 03/2022 on deceptive design patterns in social media platform interfaces: how to recognise and avoid them. Version 2.0.* European Data Protection Board, Brussels. https://edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-032022-deceptive-design-patterns-social-media_en

[8] European Data Protection Board. 2023. *Report of the work undertaken by the Cookie Banner Taskforce.* European Data Protection Board, Brussels. https://edpb.europa.eu/system/files/2023-01/edpb_20230118_report_cookie_banner_taskforce_en.pdf

[9] Kerstin Bongard-Blanchy, Arianna Rossi, Salvador Rivas, Sophie Doublet, Vincent Koenig, and Gabriele Lenzini. 2021. "I am Definitely Manipulated, Even When I am Aware of it. It's Ridiculous!" - Dark Patterns from the End-User Perspective. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference* (Virtual Event, USA) *(DIS '21)*. Association for Computing Machinery, New York, NY, USA, 763–776. https://doi.org/10.1145/3461778.3462086

[10] Christoph Bösch, Benjamin Erb, Frank Kargl, Henning Kopp, and Stefan Pfattheicher. 2016. Tales from the Dark Side: Privacy Dark Strategies and Privacy Dark Patterns. *Proc. Priv. Enhancing Technol.* 2016, 4 (2016), 237–254.

[11] Ahmed Bouhoula, Karel Kubicek, Amit Zac, Carlos Cotrini, and David Basin. 2024. Automated Large-Scale Analysis of Cookie Notice Compliance. In *Proceedings of the 33rd USENIX Security Symposium.* USENIX Association, Philadelphia, US, 1723–1739. https://www.usenix.org/conference/usenixsecurity24/presentation/bouhoula

[12] Harry Brignull. 2022. Deceptive patterns. https://www.deceptive.design Last accessed 30 October 2022.

[13] Harry Brignull, Mark Leiser, Cristiana Santos, and Kosha Doshi. 2023. Deceptive Patterns Database of Legal cases. https://www.deceptive.design/cases

[14] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 1–21. https://doi.org/10.1145/3449287

[15] Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2025. MLLM-as-a-Judge: assessing multimodal LLM-as-a-Judge with vision-language benchmark. In *Proceedings of the 41st International Conference on Machine Learning* (Vienna, Austria) *(ICML'24)*. JMLR.org, Article 254, 34 pages.

[16] Jieshan Chen, Jiamou Sun, Sidong Feng, Zhenchang Xing, Qinghua Lu, Xiwei Xu, and Chunyang Chen. 2023. Unveiling the Tricks: Automated Detection of Dark Patterns in Mobile Applications. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 114, 20 pages. https://doi.org/10.1145/3586183.3606783

[17] Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 4302–4310.

[18] Miruna-Adriana Clinciu, Arash Eshghi, and Helen Hastie. 2021. A Study of Automatic Metrics for the Evaluation of Natural Language Explanations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (Eds.). Association for Computational Linguistics, Online, 2376–2387. https://doi.org/10.18653/v1/2021.eacl-main.202

[19] CPRA. 2020. The California Privacy Rights Act of 2020. https://vig.cdn.sos.ca.gov/2020/general/pdf/topl-prop24.pdf Last accessed 9 January 2023.

[20] Mary L Cummings. 2017. Automation bias in intelligent time critical decision support systems. In *Decision making in aviation.* Routledge, Abingdon, Oxfordshire, UK, 289–294.

[21] Valdemar Danry, Pat Pataranutaporn, Yaoli Mao, and Pattie Maes. 2023. Don't Just Tell Me, Ask Me: AI Systems that Intelligently Frame Explanations as Questions Improve Human Logical Discernment Accuracy over Causal AI explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 352, 13 pages. https://doi.org/10.1145/3544548.3580672

[22] Verena Distler, Gabriele Lenzini, Carine Lallemand, and Vincent Koenig. 2020. The Framework of Security-Enhancing Friction: How UX Can Help Users Behave More Securely. In *New Security Paradigms Workshop 2020.* ACM, Online USA, 45–58. https://doi.org/10.1145/3442167.3442173

[23] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[24] European Parliament and Council of the European Union. 2022. REGULATION (EU) 2022/1925 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act). https://eur-lex.europa.eu/eli/reg/2022/1925/oj/eng Official Journal of the European Union, L 265/1, 12.10.2022.

[25] European Parliament and Council of the European Union. 2022. REGULATION (EU) 2022/2065 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R2065 Official Journal of the European Union, L 277, 27.10.2022, p. 1-102.

[26] European Parliament and Council of the European Union. 2023. Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act). https://eur-lex.europa.eu/eli/reg/2023/2854/oj/eng Official Journal of the European Union, L 2023/2854, 22.12.2023.

[27] Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement* 33, 3 (1973), 613–619.

[28] Raymond Fok and Daniel S. Weld. 2023. In Search of Verifiability: Explanations Rarely Enable Complementary Performance in AI-Advised Decision Making. *AI*

Mag. 45 (2023), 317–332. https://api.semanticscholar.org/CorpusID:258686588

[29] Directorate-General for Justice, Consumers (European Commission), Francisco Lupiáñez-Villanueva, Alba Boluda, Francesco Bogliacino, Giovanni Liva, Lucie Lechardoy, and Teresa Rodríguez de las Heras Ballell. 2022. *Behavioural study on unfair commercial practices in the digital environment: dark patterns and manipulative personalisation: final report.* Publications Office of the European Union, LU. https://data.europa.eu/doi/10.2838/859030

[30] European Commission Directorate General for Research and Innovation. 2024. *Living guidelines on the responsible use of generative AI in research.* European Commission, Brussels. https://research-and-innovation.ec.europa.eu/document/2b6cf7e5-36ac-41cb-aab5-0d32050143dc_en

[31] FTC. 2022. Bringing Dark Patterns to Light. https://www.ftc.gov/reports/bringing-dark-patterns-light

[32] Sadaf Ghaffari and Nikhil Krishnaswamy. 2024. Exploring Failure Cases in Multimodal Reasoning About Physical Dynamics. *Proceedings of the AAAI Symposium Series* 3, 1 (May 2024), 105–114. https://doi.org/10.1609/aaaiss.v3i1.31189

[33] Louie Giray. 2023. Prompt engineering with ChatGPT: a guide for academic writers. *Annals of biomedical engineering* 51, 12 (2023), 2629–2633.

[34] Colin M. Gray, Cristiana Santos, and Nataliia Bielova. 2023. Towards a Preliminary Ontology of Dark Patterns Knowledge. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 286, 9 pages. https://doi.org/10.1145/3544549.3585676

[35] Colin M. Gray, Cristiana Teixeira Santos, Nataliia Bielova, and Thomas Mildner. 2024. An Ontology of Dark Patterns Knowledge: Foundations, Definitions, and a Pathway for Shared Knowledge-Building. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 289, 22 pages. https://doi.org/10.1145/3613904.3642436

[36] Paul Graßl, Hanna Schraffenberger, Frederik Zuiderveen Borgesius, and Moniek Buijzen. 2021. Dark and Bright Patterns in Cookie Consent Requests. *Journal of Digital Social Research* 3, 1 (2021), 1–38.

[37] Johanna Gunawan, Cristiana Santos, and Irene Kamara. 2022. Redress for Dark Patterns Privacy Harms? A Case Study on Consent Interactions. In *Proceedings of the 2022 Symposium on Computer Science and Law*. ACM, Washington DC USA, 181–194. https://doi.org/10.1145/3511265.3550448

[38] Robert R. Hoffman, Gary Klein, and Shane T. Mueller. 2018. Explaining Explanation For "Explainable Ai". *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 62, 1 (2018), 197–201. https://doi.org/10.1177/1541931218621047 arXiv:https://doi.org/10.1177/1541931218621047

[39] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for Explainable AI: Challenges and Prospects. *ArXiv* abs/1812.04608 (2018). https://api.semanticscholar.org/CorpusID:54577009

[40] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276* (2024).

[41] Gautier Izacard and Edouard Grave. 2021. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (Eds.). Association for Computational Linguistics, Online, 874–880.

[42] Daniel Kahneman. 2011. *Thinking, fast and slow.* Macmillan.

[43] Sunnie S. Y. Kim, Q. Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. "I'm Not Sure, But...": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 822–835. https://doi.org/10.1145/3630106.3658941

[44] Daniel Kirkman, Kami Vaniea, and Daniel W. Woods. 2023. DarkDialogs: Automated detection of 10 dark patterns on cookie dialogs. In *2023 IEEE 8th European Symposium on Security and Privacy*. IEEE, Delft, NL, 847–867. https://doi.org/10.1109/EuroSP57164.2023.00055

[45] Agnieszka Kitkowska. 2023. *The Hows and Whys of Dark Patterns: Categorizations and Privacy.* Springer International Publishing, Cham, 173–198. https://doi.org/10.1007/978-3-031-28643-8_9

[46] Emre Kocyigit, Arianna Rossi, and Gabriele Lenzini. 2023. Towards Assessing Features of Dark Patterns in Cookie Consent Processes. In *Privacy and Identity Management (IFIP Advances in Information and Communication Technology)*, Felix Bieker, Joachim Meyer, Sebastian Pape, Ina Schiering, and Andreas Weich (Eds.). Springer Nature Switzerland, Cham, 165–183. https://doi.org/10.1007/978-3-031-31971-6_13

[47] Emre Kocyigit, Arianna Rossi, and Gabriele Lenzini. 2024. A Systematic Approach for A Reliable Detection of Deceptive Design Patterns Through Measurable HCI Features. In *Proceedings of the 2024 European Symposium on Usable Security (EuroUSEC '24)*. Association for Computing Machinery, New York, NY, USA, 290–308. https://doi.org/10.1145/3688459.3688475

[48] Udo Kuckartz and Stefan Rädiker. 2019. *Analyzing qualitative data with MAXQDA.* Springer.

[49] Jinqi Lai, Wensheng Gan, Jiayang Wu, Zhenlian Qi, and Philip S. Yu. 2024. Large language models in law: A survey. *AI Open* 5 (2024), 181–196. https://doi.org/10.1016/j.aiopen.2024.09.002

[50] Kathryn Ann Lambe, Gary O'Reilly, Brendan D. Kelly, and Sarah Curristan. 2016. Dual-process cognitive interventions to enhance diagnostic reasoning: a systematic review. *BMJ quality & safety* 25, 10 (2016), 808–820. https://qualitysafety.bmj.com/content/25/10/808.short

[51] M. R. Leiser and Cristiana Santos. 2024. Dark Patterns, Enforcement, and the emerging Digital Design Acquis: Manipulation beneath the Interface. *European Journal of Law and Technology* 15, 1 (2024). https://ejlt.org/index.php/ejlt/article/view/990

[52] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.

[53] Xinyi Li, Sai Wang, Siqi Zeng, Yu Wu, and Yi Yang. 2024. A survey on LLM-based multi-agent systems: workflow, infrastructure, and challenges. *Vicinagearth* 1, 1 (2024), 9.

[54] Zijing Liang, Yanjie Xu, Yifan Hong, Penghui Shang, Qi Wang, Qiang Fu, and Ke Liu. 2024. A Survey of Multimodel Large Language Models. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering* (Xi' an, China) *(CAICE '24)*. Association for Computing Machinery, New York, NY, USA, 405–409. https://doi.org/10.1145/3672758.3672824

[55] Q. Vera Liao and Jennifer Wortman Vaughan. 2024. AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. *Harvard Data Science Review* Special Issue 5, Special Issue 5 (may 31 2024). https://hdsr.mitpress.mit.edu/pub/aelql9qy.

[56] Xialing Lin, Patric R Spence, and Kenneth A Lachlan. 2016. Social media and credibility indicators: The effect of influence cues. *Computers in human behavior* 63 (2016), 264–271.

[57] LMSYS. 2025. LMSYS Org — lmsys.org. https://lmsys.org/blog/2024-06-27-multimodal/. [Accessed 09-01-2025].

[58] Yuwen Lu, Chao Zhang, Yuewen 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.

[59] S M Hasan Mansur, Sabiha Salma, Damilola Awofisayo, and Kevin Moran. 2023. AidUI: Toward Automated Recognition of Dark Patterns in User Interfaces. In *Proceedings of the 45th International Conference on Software Engineering* (Melbourne, Victoria, Australia) *(ICSE '23)*. IEEE Press, 1958–1970. https://doi.org/10.1109/ICSE48619.2023.00166

[60] 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.

[61] 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* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 360, 18 pages. https://doi.org/10.1145/3411764.3445610

[62] Célestin Matte, Nataliia Bielova, and Cristiana Santos. 2020. Do Cookie Banners Respect my Choice?: Measuring Legal Compliance of Banners from IAB Europe's Transparency and Consent Framework. In *2020 IEEE Symposium on Security and Privacy (SP)*. IEEE, 791–809.

[63] Thomas Mildner, Gian-Luca Savino, Philip R. Doyle, Benjamin R. Cowan, and Rainer Malaka. 2023. About Engaging and Governing Strategies: A Thematic Analysis of Dark Patterns in Social Networking Services. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 192, 15 pages. https://doi.org/10.1145/3544548.3580695

[64] Siddharth Nandagopal. 2025. Securing Retrieval-Augmented Generation Pipelines: A Comprehensive Framework. *Journal of Computer Science and Technology Studies* 7, 1 (2025), 17–29.

[65] Dmitry Nazarov and Yerkebulan Baimukhambetov. 2022. Clustering of Dark Patterns in the User Interfaces of Websites and Online Trading Portals (E-Commerce). *Mathematics* 10, 18 (2022). https://doi.org/10.3390/math10183219

[66] NCC. 2018. Deceived by design, how tech companies use dark patterns to discourage us from exercising our rights to privacy. *Norwegian Consumer Council Report* (2018).

[67] Liming Nie, Yangyang Zhao, Chenglin Li, Xuqiong Luo, and Yang Liu. 2024. Shadows in the Interface: A Comprehensive Study on Dark Patterns. *Proc. ACM Softw. Eng.* 1, FSE, Article 10 (jul 2024), 22 pages. https://doi.org/10.1145/3643736

[68] High-Level Expert Group on Artificial Intelligence. 2019. *Ethics Guidelines for Trustworthy AI.* European Commission, Brussels. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419

[69] Rajvardhan Patil and Venkat Gudivada. 2024. A Review of Current Trends, Techniques, and Challenges in Large Language Models (LLMs). *Applied Sciences* 14, 5 (2024). https://doi.org/10.3390/app14052074

[70] Marie Potel-Saville and Mathilde Da Rocha. 2023. From Dark Patterns to Fair Patterns? Usable Taxonomy to Contribute Solving the Issue with Countermeasures. In *Privacy Technologies and Policy: 11th Annual Privacy Forum, APF 2023, Lyon, France, June 1–2, 2023, Proceedings* (Lyon, France). Springer-Verlag, Berlin, Heidelberg, 145–165. https://doi.org/10.1007/978-3-031-61089-9_7

[71] Snehal Prabhudesai, Leyao Yang, Sumit Asthana, Xun Huan, Q. Vera Liao, and Nikola Banovic. 2023. Understanding Uncertainty: How Lay Decision-makers Perceive and Interpret Uncertainty in Human-AI Decision Making. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 379–396. https://doi.org/10.1145/3581641.3584033

[72] Shuhan Qi, Zhengying Cao, Jun Rao, Lei Wang, Jing Xiao, and Xuan Wang. 2023. What is the limitation of multimodal LLMs? A deeper look into multimodal LLMs through prompt probing. *Information Processing & Management* 60, 6 (2023), 103510. https://doi.org/10.1016/j.ipm.2023.103510

[73] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[74] Professor Christine Riefa and Liz Coll. 2024. *The transformative potential of Enforcement Technology (EnfTech) in Consumer Law*.

[75] Hauke Sandhaus. 2023. Promoting Bright Patterns. In *CHI '23 Workshop: Designing Technology and Policy Simultaneously*. arXiv, Hamburg, DE. https://doi.org/10.48550/arXiv.2304.01157 arXiv:2304.01157 [cs].

[76] Cristiana Santos and Arianna Rossi. 2023. The emergence of dark patterns as a legal concept in case law. *Internet Policy Review* (July 2023). https://policyreview.info/articles/news/emergence-of-dark-patterns-as-a-legal-concept

[77] Yasin Sazid, Mridha Md. Nafis Fuad, and Kazi Sakib. 2023. Automated Detection of Dark Patterns Using In-Context Learning Capabilities of GPT-3. In *2023 30th Asia-Pacific Software Engineering Conference (APSEC)*. 569–573. https://doi.org/10.1109/APSEC60848.2023.00072

[78] René Schäfer, Paul Miles Preuschoff, René Röpke, Sarah Sahabi, and Jan Borchers. 2024. Fighting Malicious Designs: Towards Visual Countermeasures Against Dark Patterns. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 296, 13 pages. https://doi.org/10.1145/3613904.3642661

[79] British Psychological Society and thics Committee of the British Psychological Society. 2018. *Code of ethics and conduct*. The British Psychological Society, Leicester.

[80] European Data Protection Supervisor. 2023. TechDispatch #2/2023: Explainable Artificial Intelligence. https://www.edps.europa.eu/data-protection/our-work/publications/techdispatch/2023-11-16-techdispatch-22023-explainable-artificial-intelligence_en Accessed: 2024-12-23.

[81] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine* 29, 8 (2023), 1930–1940.

[82] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 9568–9578.

[83] Matthijs J Warrens. 2015. Five ways to look at Cohen's kappa. *Journal of Psychology & Psychotherapy* 5 (2015).

[84] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 24824–24837. https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf

[85] J. Wu, W. Gan, Z. Chen, S. Wan, and P. S. Yu. 2023. Multimodal Large Language Models: A Survey. In *2023 IEEE International Conference on Big Data (BigData)*. IEEE Computer Society, Los Alamitos, CA, USA, 2247–2256. https://doi.org/10.1109/BigData59044.2023.10386743

[86] Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S. Yu. 2023. Multimodal Large Language Models: A Survey. In *2023 IEEE International Conference on Big Data (BigData)*. 2247–2256. https://doi.org/10.1109/BigData59044.2023.10386743

[87] Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024. Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond. *ACM Trans. Knowl. Discov. Data* 18, 6, Article 160 (April 2024), 32 pages. https://doi.org/10.1145/3649506

[88] José P. Zagal, Staffan Björk, and Chris Lewis. 2013. Dark Patterns in the Design of Games. In *Foundations of Digital Games 2013*. RISE, Swedish ICT, RISE, Swedish ICT, Interactive Institute. GAME., 7. http://urn.kb.se/resolve?urn=urn:nbn:se:ri:diva-24252 Conference paper, Refereed. Available from: 2016-10-31.

[89] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 295–305. https://doi.org/10.1145/3351095.3372852

[90] Zhuosheng Zhang, Aston Zhang, Mu Li, hai zhao, George Karypis, and Alex Smola. 2024. Multimodal Chain-of-Thought Reasoning in Language Models. *Transactions on Machine Learning Research* (2024). https://openreview.net/forum?id=y1pPWFVfvR

[91] Jianlong Zhou, Amir H. Gandomi, Fang Chen, and Andreas Holzinger. 2021. Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics. *Electronics* 10, 5 (2021). https://doi.org/10.3390/electronics10050593

A Data sources

Table 6: Sources of the Dataset

Source	URL
Web	https://www.deceptive.design/types/
Web	https://hallofshame.design/
Web	https://darkpatterns.uxp2.com/patterns/
Web	https://tuta.com/blog/outlook...
Publication	https://dl.acm.org/doi/10.1145/3359183
Report	https://www.edpb.europa.eu/system/files
Web	https://thomasmildner.me/darkpatterns.html
Web	https://givewp.com/addons/recurring-donations/
Web	https://paymentsplugin.com/blog/donation-page/
Web	https://uxplaybook.org/articles/ux-dark...
Report	https://www.ftc.gov/system/files...
Web	https://www.osano.com/articles/dark-pattern...
Web	https://think.design/blog/responsible-design...
Web	https://blog.crobox.com/article/dark-patterns
Web	https://things.qz.com/2019/dark-patterns...
Web	https://blog.mobiversal.com/dark-patterns...

B True and false positives

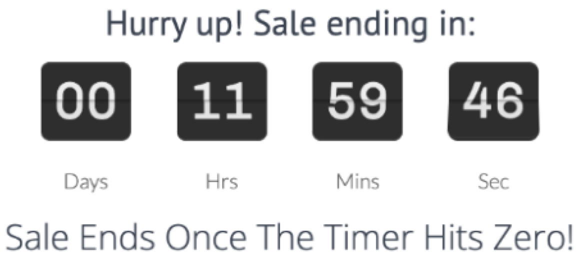
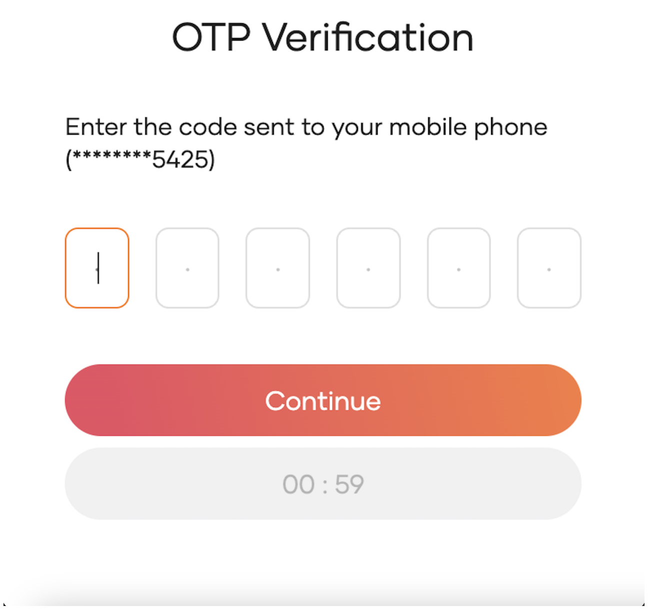

Hurry up! Sale ending in:
00
11
59
46
Days
Hrs
Mins
Sec
Sale Ends Once The Timer Hits Zero!

Figure 2: A True Positive Example of *Countdown Timer DP*

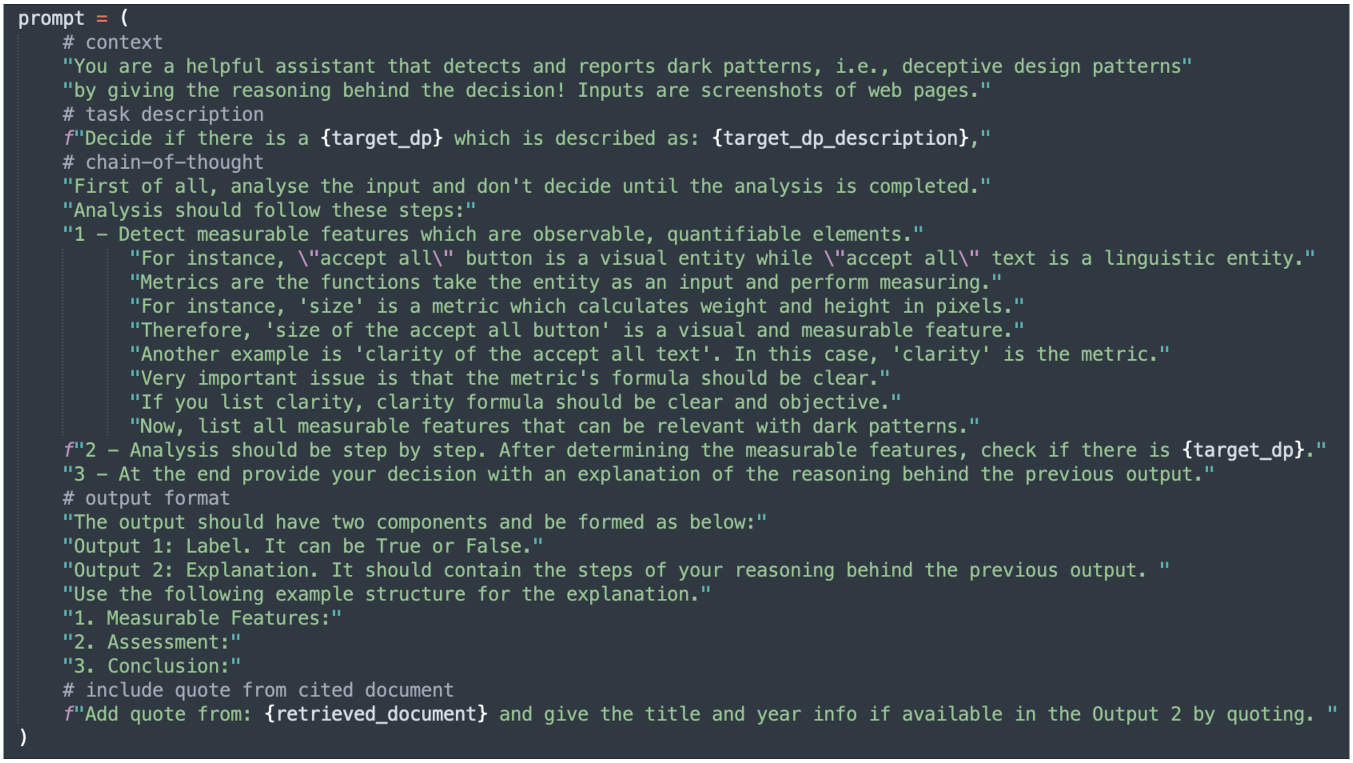


OTP Verification

Enter the code sent to your mobile phone (********5425)

Figure 3: A False Positive Example of *Countdown Timer DP*

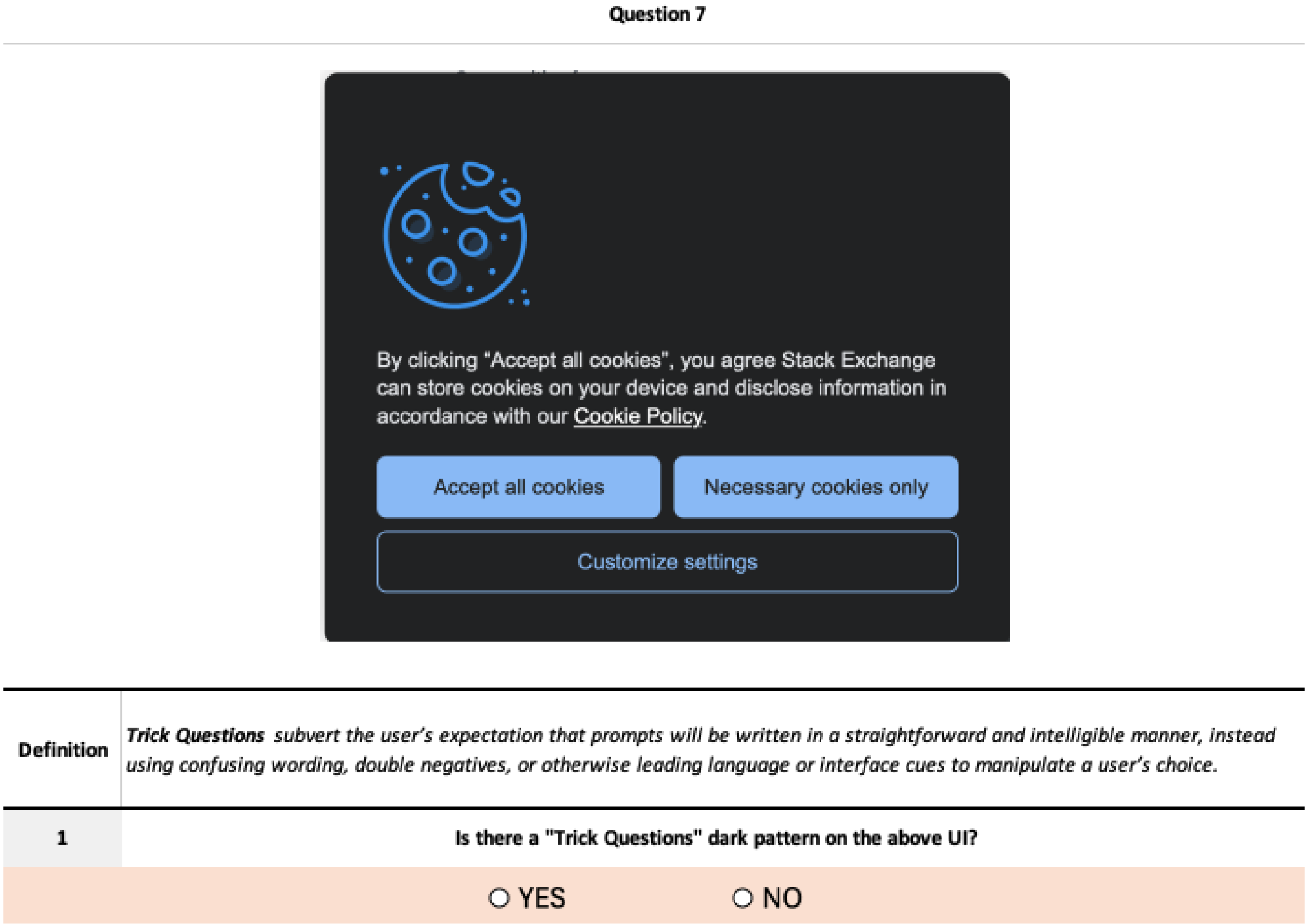
C Entire prompt



```
prompt = (
    # context
    "You are a helpful assistant that detects and reports dark patterns, i.e., deceptive design patterns"
    "by giving the reasoning behind the decision! Inputs are screenshots of web pages."
    # task description
    f"Decide if there is a {target_dp} which is described as: {target_dp_description},"
    # chain-of-thought
    "First of all, analyse the input and don't decide until the analysis is completed."
    "Analysis should follow these steps:"
    "1 - Detect measurable features which are observable, quantifiable elements."
          "For instance, \"accept all\" button is a visual entity while \"accept all\" text is a linguistic entity."
          "Metrics are the functions take the entity as an input and perform measuring."
          "For instance, 'size' is a metric which calculates weight and height in pixels."
          "Therefore, 'size of the accept all button' is a visual and measurable feature."
          "Another example is 'clarity of the accept all text'. In this case, 'clarity' is the metric."
          "Very important issue is that the metric's formula should be clear."
          "If you list clarity, clarity formula should be clear and objective."
          "Now, list all measurable features that can be relevant with dark patterns."
    f"2 - Analysis should be step by step. After determining the measurable features, check if there is {target_dp}."
    "3 - At the end provide your decision with an explanation of the reasoning behind the previous output."
    # output format
    "The output should have two components and be formed as below:"
    "Output 1: Label. It can be True or False."
    "Output 2: Explanation. It should contain the steps of your reasoning behind the previous output. "
    "Use the following example structure for the explanation."
    "1. Measurable Features:"
    "2. Assessment:"
    "3. Conclusion:"
    # include quote from cited document
    f"Add quote from: {retrieved_document} and give the title and year info if available in the Output 2 by quoting. "
)
```

Figure 4: A prompt for the MM-LLM and `target-dp` refers to the targeted DP in the UI, and its description is stored in `target-dp-description` variable.

D Overview of the tasks

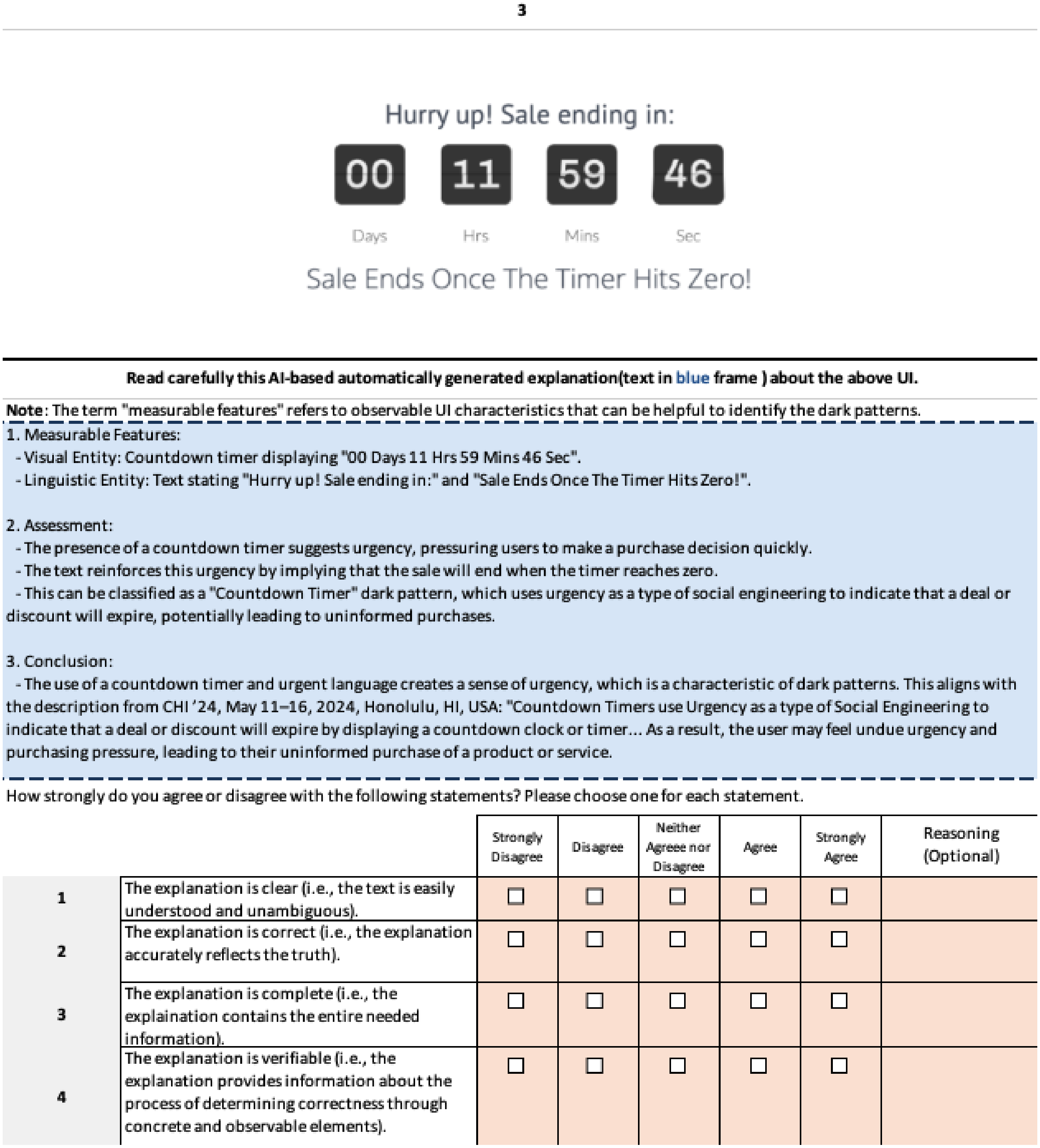


Question 7

By clicking "Accept all cookies", you agree Stack Exchange can store cookies on your device and disclose information in accordance with our Cookie Policy.

Definition	***Trick Questions*** *subvert the user's expectation that prompts will be written in a straightforward and intelligible manner, instead using confusing wording, double negatives, or otherwise leading language or interface cues to manipulate a user's choice.*
1	**Is there a "Trick Questions" dark pattern on the above UI?**
	○ YES ○ NO

Figure 5: An example question for the experts that was asked in the Task 1.In this task(more details are given in Section 3.4.1), the experts were asked to evaluate whether there is a specific deceptive pattern in the given UI. For each question, we also provide the related definition of the DPs that were taken from the recent ontology study [35] below the image.


3
Hurry up! Sale ending in:
00
11
59
46
Days
Hrs
Mins
Sec
Sale Ends Once The Timer Hits Zero!

Read carefully this AI-based automatically generated explanation(text in blue frame) about the above UI.

Note: The term "measurable features" refers to observable UI characteristics that can be helpful to identify the dark patterns.

1. Measurable Features:
  - Visual Entity: Countdown timer displaying "00 Days 11 Hrs 59 Mins 46 Sec".
  - Linguistic Entity: Text stating "Hurry up! Sale ending in:" and "Sale Ends Once The Timer Hits Zero!".

2. Assessment:
  - The presence of a countdown timer suggests urgency, pressuring users to make a purchase decision quickly.
  - The text reinforces this urgency by implying that the sale will end when the timer reaches zero.
  - This can be classified as a "Countdown Timer" dark pattern, which uses urgency as a type of social engineering to indicate that a deal or discount will expire, potentially leading to uninformed purchases.

3. Conclusion:
  - The use of a countdown timer and urgent language creates a sense of urgency, which is a characteristic of dark patterns. This aligns with the description from CHI '24, May 11–16, 2024, Honolulu, HI, USA: "Countdown Timers use Urgency as a type of Social Engineering to indicate that a deal or discount will expire by displaying a countdown clock or timer... As a result, the user may feel undue urgency and purchasing pressure, leading to their uninformed purchase of a product or service.

How strongly do you agree or disagree with the following statements? Please choose one for each statement.

		Strongly Disagree	Disagree	Neither Agreee nor Disagree	Agree	Strongly Agree	Reasoning (Optional)
1	The explanation is clear (i.e., the text is easily understood and unambiguous).	☐	☐	☐	☐	☐	
2	The explanation is correct (i.e., the explanation accurately reflects the truth).	☐	☐	☐	☐	☐	
3	The explanation is complete (i.e., the explaination contains the entire needed information).	☐	☐	☐	☐	☐	
4	The explanation is verifiable (i.e., the explanation provides information about the process of determining correctness through concrete and observable elements).	☐	☐	☐	☐	☐	

Figure 6: An example question for the experts that was asked in the Task 2. In this task(more details are given in Section 3.4.2), the experts were asked to evaluate *clarity, correctness, completeness* and *verifiability* of the explanations that were generated by our system.

E Interview questions

After a short introduction by the researchers on the research goal, the interviews were organized in three main parts.

E.1 Part 1: Demographics

(1) What is your professional area of expertise and how long have you been working in the area?
(2) How long have you been working on Dark Patterns problems?

E.2 Part 2: Explanation of assessment

Now I would like you to reflect on the scores you provided to the explanations.

(1) Can you tell us your opinion in general about the [clarity, completeness, correctness, verifiability] of the explanations you assessed?
(2) (if there is one with extremely bad rating): You marked one of the explanations you received as particularly bad, comparing with the rest. Can you explain why? What was the difference between this and others explanations?
(3) (if some of the ratings are much higher, compared to the rest): You marked one of the explanations you received as particularly good, comparing with the rest. Can you explain why? What was the difference between this and others explanations?
(4) (if there are groups of different ratings, e.g 4-5 and 1-2, or a totally diverse sample): In your sample you marked some of the explanations as much better than others. Can you reflect on these two groups and explain the difference between them
(5) Beyond these four criteria, would you like to add something more about your perception of quality of explanations?
(6) Is there any other criteria that you believe would be useful to evaluate the quality of explanations?

E.3 Part 3: Usefulness of the AI-based explanations

(Explanation about the tool) We have created a detection tool based on a multi-model large language model that, based on screenshots and prompts of the researcher, assesses the presence of dark patterns and provides explanations for the decision it takes. To increase its accuracy and reliability, instead of relying on publicly available data to answer the prompts, (the design contains Retrieval Augmented Generation techniques) our tool considers a few peer-reviewed academic articles on dark patterns, which contain definitions, attributes and examples.

(1) In your domain of practice how useful a tool like this can be?
(2) Can you envision possible applications?
(3) Would you personally use this tool?
(4) (if yes) For which purposes would it be suitable to you
(5) (if no) Why? Is there any modification that it could make it useful to you?

E.4 Conclusions

(1) Is there anything else that you would like to add?
(2) Is there anything you would like to ask us?