Graduate School of Science and Technology Master's Thesis Abstract

Laboratory name (Supervisor)	Ubiquitous Computing Systems (Keiichi Yasumoto (Professor))		
Student ID	2411431	Submission date	2026 / 7 / 22
Name	YAO KOUAME JEAN FLORENTIN		
Thesis title	An Empirical Study of the Effect of Screen Representation, Interaction Context, and Model Scale on Dark Pattern Detection Across Mobile Interaction Flows Using Large and Small Language Models		

Abstract

Dark patterns are interface designs crafted not in the interest of users, but to steer them toward decisions they would likely not have made otherwise.

The growing prevalence of such manipulative designs has motivated the development of automated dark pattern detection systems. A first approach trains, on a fixed set of annotated patterns, systems that combine a recognition model, such as a computer-vision detector or a contrastive-learning classifier, with predefined rules to recognize these manipulative techniques. While this allows detection to take place, such systems remain unable to recognize any technique that falls outside the taxonomy they were trained on. More recently, the emergence of large language models (LLMs) has enabled a second approach, in which the model's general knowledge is leveraged for detection instead, removing the need to fine-tune a system to a fixed taxonomy. Implementations of this approach have so far only operated on inputs taken in isolation, which precludes effective detection of dark patterns that only become apparent once the interaction flow itself is considered, as is the case with the Roach Motel pattern.

The need for such detection systems is further underscored by the emergence of computer-use agents, which have themselves been shown to be susceptible to dark patterns. Existing work in this space, however, measures the vulnerability of the agent to manipulation, not the capacity of a separate system to detect and flag a dark pattern present in the interface, independently of whether the agent encountering it would be susceptible to it. Guardrail-style systems have been proposed for computer-use agents, in which a second model is invoked at a given step of the interaction to determine whether a dark pattern is present, but these remain designed specifically for commercial LLMs operating in a web context, are limited to a single step at a time, and do not systematically examine what is gained or lost by varying the amount of context given to the model. This design choice also raises a privacy concern, since it requires sending a screen, potentially containing sensitive information, to the cloud.

This thesis studies the capacity of language models to detect dark patterns in mobile interaction flows along three axes, each varied independently of the others. The first is the representation of the screens themselves: the raw screenshot, a textual description of it generated by an LLM, and two screen parsers, OmniParser, used off-the-shelf, and Qwen-Parser, fine-tuned specifically for this study, with the aim of observing the influence of input format on detection capability. The second axis is the amount of context provided to the model, increased gradually from a single representation of one screen of a flow at a time (isolated), to all representations of a flow at once (flow), to the flow together with the stated goal of the task (goal). The third axis is a variation in model scale, ranging from commercial LLMs to SLMs, with the aim of assessing the feasibility of on-device deployment. Combining all three axes yields a total of 144 configurations, each evaluated both on its capacity to detect the presence of a dark pattern and on its capacity to identify the category of dark pattern present, using both the F1 score and the Matthews Correlation Coefficient (MCC), the latter required because the dataset used in this study is imbalanced.

A textual description of the screen is the strongest representation for binary detection across every model (F1 of 0.801, against 0.685 for the weakest representation), and the largest locally runnable models match, and on single-screen judgments even marginally exceed, the commercial LLMs' MCC (0.630 against 0.629), a parity that gives way to a gap of 0.188 once the model must integrate evidence across a full flow annotated with the task's goal. Neither additional flow context nor stating the task's goal reliably improves detection. Identifying the category of dark pattern present proves markedly harder: exact matching of category names caps the best configuration at a macro F1 of 0.165, partly because only 10 of the 25 established categories appear in the evaluation data. Relaxing this matching using a table of 20 category pairs justified by prior literature raises macro F1 from 0.091 to 0.138 on average, confirmed by a randomization test not to be a mere artifact of relaxed matching in general ($p=0.022$). F1 and MCC disagree often enough on which model performs best that reporting both, rather than F1 alone, is necessary to avoid mistaking over-prediction for genuine discrimination.