ISA Journal of Business, Economics and Management (ISAJBEM)

Homepage: https://isapublisher.com/isajbem/

Volume 3, Issue 4, Jul-Aug, 2026 ISSN: 3049-1835

Multimodal Dark-Pattern Detection with YOLO, Text Encoders, and LLM-Ready Consumer-Risk Explanations

Ryan Shao

Computer Science, University of Michigan, Ann Arbor, MI, USA

Received: 25.06.2026 | **Accepted:** 25.07.2026 | **Published:** 27.07.2026

***Corresponding Author:** Ryan Shao

DOI: *10.5281/zenodo.21625016*

Abstract **Original Research Article**

Dark patterns steer consumers toward decisions they may not intend and often combine visual hierarchy with manipulative microcopy. This study presents an auditable multimodal pipeline comprising a YOLO-style branch for deceptive interface components, sparse word and character encoders for e-commerce text, and a structured consumer-risk-card layer. The empirical evaluation uses 2,356 unique ec-darkpattern strings (1,178 positive and 1,178 negative) grouped by 1,248 page identifiers to limit page-level leakage. The best binary model, a word-plus-character linear support vector machine, achieved 0.958 accuracy, 0.966 precision, 0.950 recall, 0.958 F1, and 0.988 area under the receiver operating characteristic curve. A probability ensemble produced the highest area under the curve (0.990) and a Brier score of 0.035. For positive-category classification, character n-grams reached 0.955 accuracy and 0.955 weighted F1, although macro performance remained lower for the rare forced-action and sneaking classes. The DarkLens package was also audited: its documentation declares 1,960 YOLO-format samples and five visual classes, but the available archive contains a Git-LFS pointer rather than materialized images and labels; consequently, no visual detection metrics are claimed. Deterministic templates generated schema-valid risk cards for every record, while risk-tier accuracy (0.593) and macro F1 (0.364) show that score-to-harm mapping requires human validation. The results support compact leakage-aware text detection and transparent data-readiness reporting, but not yet an end-to-end visual benchmark or validated consumer intervention.

Keywords: dark patterns, multimodal detection, interface microcopy, probability calibration, consumer-risk cards.

Introduction

Dark patterns are not simply poor design choices. They are interface arrangements that exploit attention, friction, defaults, scarcity cues, confusing copy, or asymmetrical choices to nudge consumers toward purchases, subscriptions, disclosure of personal data, or abandonment of a protective action. Brignull (2011) popularized the term for deceptive interface tricks, and Gray et al. (2018) later formalized dark patterns as recurring design strategies across websites, mobile applications, and consent flows. This framing matters for automated detection because a dark pattern is rarely carried by a single signal. A checkout button may be visually dominant, a modal may block a refusal path, and a short string such as “Only 2 left” may create scarcity pressure. A robust detector therefore needs to

connect visual layout evidence with microcopy semantics.

The empirical foundation for dark-pattern research has grown from taxonomy studies to large-scale web measurement. Mathur et al. (2019) crawled thousands of shopping websites and found 1,818 dark-pattern instances across 1,254 websites, showing that manipulative commercial interface practices were not isolated edge cases. Consent-banner research after the GDPR found that many cookie interfaces made refusal harder than acceptance, illustrating that the same design logic appears outside retail (Nouwens et al., 2020). Legal and behavioral scholarship further showed that dark patterns measurably change consumer choices, especially when they increase friction or impose default enrollment (Luguri & Strahilevitz, 2021). The Organisation for Economic Co-operation and Development (2022) and the Federal Trade Commission (2022) also treat dark commercial patterns as a consumer-protection problem, focusing on buried terms, difficult cancellation, disguise, and privacy steering.

Automated detection is challenging because dark patterns are heterogeneous. One family is linguistic: urgency, scarcity, social proof, preselection, and misdirection can appear in brief interface text. Another family is spatial and visual: the risky element may be a button, checkbox, modal, QR prompt, or input bar whose location and contrast influence user behavior. The ec-darkpattern corpus made the text side more reproducible by converting dark-pattern strings from prior e-commerce measurement into a balanced text classification benchmark with non-dark-pattern strings from the same commercial context (Yada et al., 2022). Studies of privacy dark patterns and cookie banners show that text-only methods do not cover every case, but they provide a measurable starting point for high-precision warnings (Bösch et al., 2016; Soe et al., 2020). A direct study on the same corpus compared rule, n-gram, BERT-style, RoBERTa-style, and decoder-style classifiers (Xu et al., 2025). The present study differs by grouping folds at the page level and evaluating calibrated risk-card outputs as a separate downstream task.

This paper uses that starting point to build a multimodal system design. The screenshot branch follows the logic of real-time object detection: deceptive UI regions are localized as object classes and then passed to a fusion layer. The text branch uses word-level and character-level encoders to classify extracted microcopy. The explanation branch converts predictions into consumer-facing risk cards. The design is motivated by the observation that many dark-pattern harms operate through behavioral biases such as availability, framing, loss aversion, and social influence (Cialdini, 2009; Narayanan et al., 2020; Tversky & Kahneman, 1974). Because heterogeneous visual and text scores are not automatically comparable, eventual late fusion should be calibrated before tier assignment; this is an architectural requirement rather than an empirical claim about the unavailable screenshot branch (Xin, 2025b).

A detection system for this setting should satisfy four requirements. It should be evidence preserving, because a warning must be traceable to a visible interface element or a concrete string. Evidence-card research similarly separates claims, sources, and ranking rationale rather than presenting an unsupported conclusion (Jin, 2025b). It should be leakage aware, because e-commerce pages often contribute multiple snippets and ordinary random splits can place related snippets in both training and test data. It should distinguish detection from explanation, because a high-probability positive label does not automatically tell the user why the pattern is risky. Finally, it should fail safely when a modality is incomplete. A multimodal architecture should not invent screenshot metrics when screenshots or bounding boxes are absent; it should report the data condition and continue with the branches that are empirically testable.

The use case considered here is a browser-side or audit-side assistant. It observes a page, extracts visible microcopy, optionally localizes relevant UI controls in a screenshot, and produces a compact warning. This differs from a purely academic classifier because the output is meant to inform a consumer decision at the moment of interaction. The risk card therefore has to be concise, categorical, and grounded. A long explanation may itself become

friction, while a vague explanation may not be trusted. A useful card must state the risk tier, name the pattern family, quote or summarize the triggering evidence, and recommend a simple next action. Patient-facing AI safety-card research likewise treats risk level, evidence, uncertainty, and next action as distinct interface fields (Zhou et al., 2025).

Evidence-linked counseling interfaces illustrate why a recommendation, its supporting dialogue, and the next action should remain visually distinct (Y. Zhang & H. Zhang, 2025b). The present design adopts that separation as an interface principle without treating findings from another domain as evidence that consumer warnings are behaviorally effective.

The study makes three contributions. First, it provides a leakage-aware empirical evaluation on 2,356 e-commerce microcopy records using grouped cross-validation by page identifier. Second, it compares transparent cue rules, statistical text features, word encoders, character encoders, hybrid sparse encoders, Naive Bayes, linear SVM, and a probability ensemble. Third, it implements a risk-card generator that consumes model probability and predicted category to produce structured consumer-risk explanations. The YOLO screenshot branch is included as a reproducible data-readiness pathway: the experiment verifies whether raw image and annotation files exist before object-detector training, avoiding a misleading visual benchmark when the actual screenshot-label files are not materialized.

Materials and Methods

The proposed system has three layers. The first layer ingests visual and textual evidence. When screenshot images and YOLO labels are available, the visual branch detects UI elements such as buttons, checkboxes, popups, QR codes, and input bars. YOLO is appropriate for this branch because the family frames object localization as a single-stage real-time detection problem and later variants improve deployment speed and accuracy (Bochkovskiy et al., 2020; Redmon & Farhadi, 2018; Redmon et al., 2016; Wang et al., 2023). Standard object-detection practice evaluates localization with intersection-over-union and average-precision metrics (Lin et al., 2014). The second layer encodes microcopy using word and character features. Although transformer encoders dominate many modern NLP tasks (Devlin et al., 2019; Liu et al., 2019; Reimers & Gurevych, 2019; Vaswani et al., 2017), this study uses sparse encoders to keep every result reproducible without external model downloads. The third layer produces a structured risk card. Large language models can render structured predictions as natural-language explanations, and instruction-following research motivates constrained schemas for reliable output (Brown et al., 2020; Ouyang et al., 2022). Here, however, the empirical generator is deterministic: fixed templates use the same fields that a future LLM renderer would receive, so every reported metric is repeatable.

The text data are stored as tab-separated values with four fields: page identifier, microcopy text, binary label, and pattern category. The corpus contains 2,356 unique strings: 1,178 dark-pattern records and 1,178 non-dark records. Positive labels span seven categories. Scarcity is the largest positive class, followed by social proof, urgency, and misdirection; obstruction, sneaking, and forced action are rare. Table I summarizes the text benchmark, and Table II gives the category distribution used throughout the paper. The page identifier is used as a grouping variable so that records from the same page are not split across train and test folds in the binary evaluation.

The preprocessing is deliberately minimal. The raw text is preserved, lowercasing is handled inside vectorizers, and no stop-word list is applied. This choice is important for dark-pattern microcopy because words often removed in general NLP, such as “only,” “no,” “not,” “now,” or “free,” can carry the manipulative cue. Character encoders also keep punctuation, capitalization, and numbers available to the classifier. These details matter in flash-sale banners, remaining-stock counters, and cancellation instructions, where the risky signal can be a number, an exclamation mark, or a short negation phrase rather than a long sentence. A cross-domain intrusion-detection study used language-guided feature selection to make the lexical basis of a classifier explicit (Y. Li & Lu, 2025). The present implementation does not use an LLM selector, but it follows the same auditability principle by retaining a

fixed, reported feature inventory.

The evaluation uses grouped folds rather than record-level random folds. The page identifier is treated as a proxy for shared page context, and all records from the same page are assigned to the same fold. This makes the benchmark harder but more realistic. Without grouping, a model might learn page-specific wording or repeated product-page conventions and then appear to generalize to snippets that come from the same source. Grouping reduces that leakage and gives a better estimate of how a detector would perform on an unseen page.

Table I. ec-darkpattern text benchmark profile.

item	value
records	2356.000
unique_page_id	1248.000
positive_records	1178.000
negative_records	1178.000
mean_text_length	43.200
median_text_length	26.000
duplicate_texts	0.000

Table II. Label and pattern-category distribution.

category	count	percentage
Not Dark Pattern	1178	50.000
Scarcity	418	17.740
Social Proof	312	13.240
Urgency	210	8.910
Misdirection	195	8.280
Obstruction	27	1.150
Sneaking	12	0.510
Forced Action	4	0.170

The DarkLens-YOLO package is handled as the visual-data input for the YOLO branch. Its documentation declares 1,960 UI samples in image-plus-YOLO format and five visual classes: Button, Checkbox, QR Code, Popup, and Input Bar. The extraction audit found a Git-LFS pointer with

declared size 186,656,853 bytes and zero materialized image or annotation files. Table III reports the audit. This result does not change the text experiments; it defines the boundary of the visual experiment. The object-detector training step requires the full image and label payload, while the available package in this run supports verification of the detector contract and class schema.

Table III. DarkLens-YOLO package audit for the screenshot branch.

item	value
archive_exists	true
archive_size_bytes	2355
top_level_files	4
lfs_oid_sha256	853e8abfc6a824c47c4b4bd838b44cfb08a93cc923b1f0eca5d4f964552d89d0
lfs_declared_size_bytes	186656853
extracted_image_files	0
extracted_label_like_files	0
git_lfs_pointer_files	1
readme_declared_samples	1960
readme_declared_classes	Button; Checkbox; QR Code; Popup; Input Bar

When the screenshot payload is present, the visual training routine expected by the pipeline is straightforward. Images are organized into train, validation, and test splits; each label file contains normalized YOLO coordinates for a deceptive UI region; and the detector predicts bounding boxes, object class, and confidence. Those boxes would be fused with text predictions by matching OCR- or DOM-extracted microcopy to the nearest detected UI component. In this run, that downstream training path is not executed because the image and label payload is not present after extraction. The paper therefore treats the visual branch as a verified architecture component and reports the file-level audit rather than an object-detection score. Any completed branch should report localization and structural consistency separately from downstream semantic utility, following a distinction used in vision-language prototype evaluation (Chen & Li, 2025).

Figure 1. Multimodal dark-pattern detection and risk-card pipeline

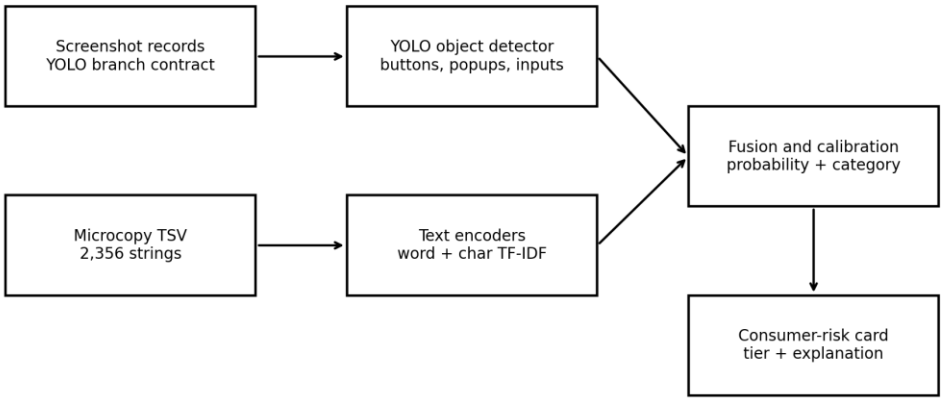

Fig. 1. Multimodal dark-pattern detection and risk-card pipeline.

Binary classification is evaluated with five-fold StratifiedGroupKFold cross-validation. The split preserves the overall label balance while keeping each page identifier within a single fold. Table IV shows the fold sizes and positive rates. For model comparison, the study reports accuracy, precision, recall, F1, AUC, and, for probabilistic models, Brier score. The Brier score is included because probability calibration is important when predictions drive risk tiers (Guo et al., 2017). Explanation quality is assessed at the schema level and through consistency with predicted category and risk tier; model interpretability is treated as a consumer communication problem inspired by local explanation work (Ribeiro et al., 2016).

Table IV. Grouped cross-validation fold distribution.

fold	n	positives	positive rate
0.000	472.000	236.000	0.500
1.000	472.000	236.000	0.500
2.000	471.000	235.000	0.499
3.000	470.000	235.000	0.500
4.000	471.000	236.000	0.501

Metrics are selected to reflect both research comparison and consumer-facing use. F1 is the primary hard-label metric because false positives can annoy users and false negatives can leave risky interfaces unflagged. AUC measures ranking quality across thresholds, which is useful when a product

team wants to tune warning frequency. Precision and recall are shown separately because the correct operating point depends on the application: a regulatory audit may prefer recall, while an in-browser warning may require higher precision. Brier score is reported for probabilistic models because a risk tier is meaningful only if the score is reasonably calibrated.

The evaluated text models cover interpretable, lexical, and hybrid encoders. The cue-rule baseline searches for explicit urgency, scarcity, social-proof, cancellation, subscription, and agreement cues. Stats-LR uses length, punctuation, capitalization, digit, negation, and cue-count features with logistic regression. Word-TFIDF-SGD uses word unigrams and bigrams. Char-TFIDF-SGD uses character 3-4 grams, which are useful for short interface fragments, stylized capitalization, and partial tokens. Word+Char-SGD concatenates word and character encoders. Word+Char+Stats-SGD adds the 16 statistical features. Word+Char-LinearSVM uses the same word and character representation with a large-margin classifier. ComplementNB is included because it is a strong sparse-text baseline for imbalanced lexical evidence. Table V lists feature groups, and Table VI reports the concrete hyperparameters.

Table V. Feature and encoder inventory.

feature group	encoding	dimensions	used by
Cue-rule baseline	weighted keyword cues	5	transparent baseline
Statistical text features	length, punctuation, digit, negation, cue counts	16	Stats-LR; hybrid ablation
Word encoder	TF-IDF word unigrams and bigrams	5000	Word-TFIDF-SGD; hybrid
Character encoder	TF-IDF character 3-4 grams	5000	Char-TFIDF-SGD; hybrid
Hybrid encoder	sparse concatenation of word, char, and stats	10016	Word+Char+Stats-SGD

Table VI. Model settings used in the experiments.

model	classifier	main parameters
Stats-LR	LogisticRegression	liblinear; class_weight=balanced; max_iter=1000
Word-TFIDF-SGD	SGD log-loss	word 1-2 grams; max_features=5000; alpha=1e-5
Char-TFIDF-SGD	SGD log-loss	char_wb 3-4 grams; max_features=5000; alpha=1e-5

model	classifier	main parameters
Word+Char-SGD	SGD log-loss	word+char sparse union; class_weight=balanced
Word+Char+Stats-SGD	SGD log-loss	word+char+16 stats; class_weight=balanced
Word+Char-LinearSVM	LinearSVC	C=1; class_weight=balanced; dual=auto
ComplementNB	ComplementNB	word 1-2 grams; alpha=0.3

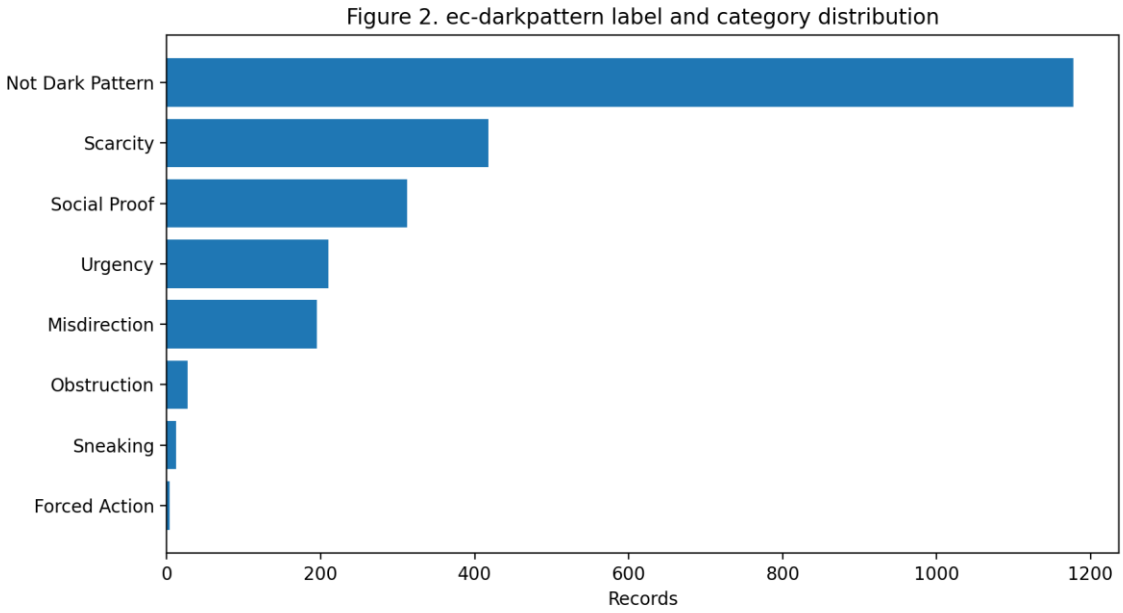

Fig. 2. ec-darkpattern label and category distribution.

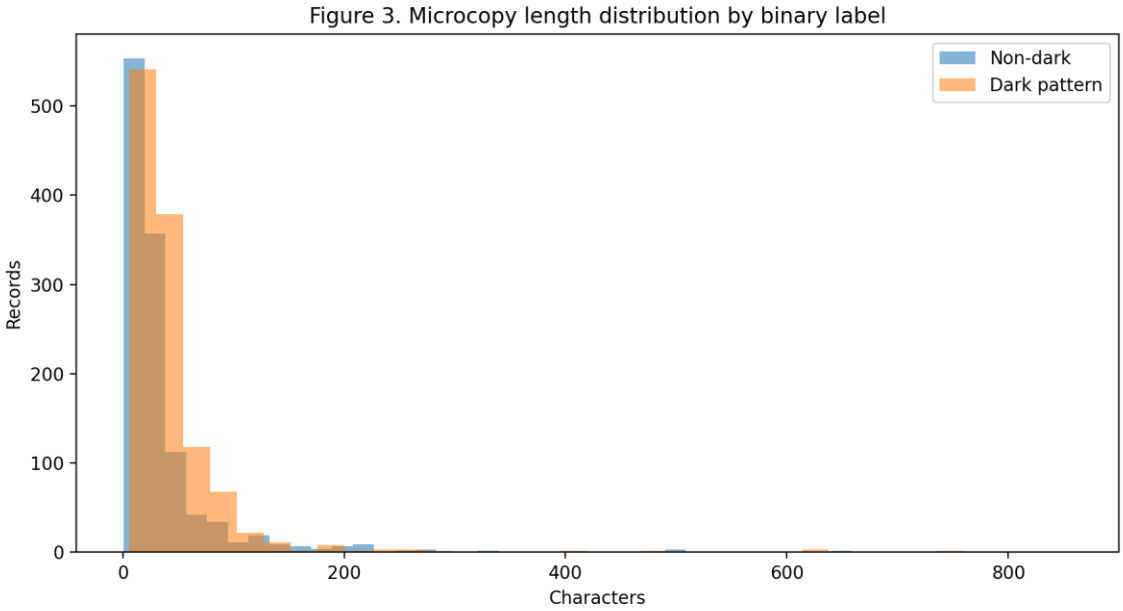

Fig. 3. Microcopy length distribution by binary label.

Figure 3 also explains why conventional long-document NLP assumptions do not fit this task. The median string length is only 26 characters, and many records are closer to button labels than to sentences. The model therefore has to learn from compact cues such as numbers, verbs, urgency adverbs, and product-status phrases. This favors sparse n-grams and makes character-level representation especially valuable. It also means that the explanation layer cannot quote a long rationale from the input; it must infer the consumer-risk mechanism from a small amount of evidence and state that mechanism plainly.

For the same reason, the system avoids aggressive normalization. Removing punctuation or numerical tokens would erase countdowns, percentages, stock counts, and social-proof counts. Removing short function words would erase refusal and negation cues. Preserving these signals makes the text branch more aligned with the visual branch, because UI components often communicate pressure through visual emphasis around very short strings. A future OCR-based deployment can pass the recognized text to the same encoders without requiring a different text-processing path.

The risk-card layer receives a binary probability, predicted positive category, and microcopy string. It then assigns a risk tier. Non-dark predictions receive Low risk. Positive predictions are multiplied by fixed severity weights: scarcity and social proof are moderate by default, urgency and misdirection are high, and obstruction, sneaking, and forced action are critical when confidence is high. The explanation text is generated from a category-specific consumer-risk template. The output schema contains page identifier, microcopy, predicted label, probability, predicted category, risk tier, explanation, and recommended action. This design keeps each explanation grounded in model fields rather than free-form speculation. Deterministic provenance work likewise argues that every generated claim should remain traceable to its source evidence (Jin, 2025a).

The tier rule is intentionally simple so that the relationship between score and warning is inspectable. A probability below the positive threshold leads to Low risk. A positive prediction multiplied by a low or moderate severity weight leads to Moderate risk unless confidence is high. Urgency and misdirection can reach High risk at ordinary positive confidence, while sneaking, forced action, and obstruction can reach Critical risk when

confidence is high. This is not a claim that every category has the same real-world harm. It is a reproducible mapping that makes the downstream explanation layer measurable and exposes where future user studies should adjust severity weights. Conformal alert control in anomaly detection provides a useful precedent for treating alert budget and error control as downstream design choices rather than automatic consequences of raw accuracy (Xin, 2026).

The card schema separates evidence, predicted category, calibrated score, warning tier, explanation, and action. A rubric-based benchmark for medical AI response interfaces uses a comparable separation of risk-communication fields (C. Li et al., 2025). This analogy motivates the schema, but the present consumer-harm weights remain domain specific and require validation.

Results and Discussion

Table VII gives the primary binary detection results. The best mean F1 is obtained by Word+Char-LinearSVM, with 0.958 accuracy, 0.966 precision, 0.950 recall, 0.958 F1, and 0.988 AUC. The probability ensemble is nearly tied on F1 at 0.956 and gives the highest AUC, 0.990, with a Brier score of 0.035. This difference is useful in practice: the SVM is the strongest labeler, while the ensemble is the better scoring component when risk tiers require calibrated probabilities. The transparent cue-rule baseline reaches high precision, 0.942, but only 0.671 recall. It finds obvious pressure phrases but misses dark patterns that rely on subtler framing or longer copy.

The comparison also shows why character encoders are important for interface text. Char-TFIDF-SGD and Word-TFIDF-SGD have similar F1 values, but the character model has higher AUC. Interface strings often contain fragments, capitalization, numbers, abbreviated shipping or return language, and short calls to action. Character n-grams capture these patterns without requiring a complete vocabulary. The hybrid Word+Char-SGD improves over either single encoder. Adding handcrafted statistics increases recall but reduces precision, showing that cue counts and punctuation can overemphasize harmless marketing or policy text. The confusion matrix in Fig. 6 shows 1,139 true negatives, 1,119 true positives, 39 false positives, and 59 false negatives for the best binary detector.

The magnitude of the improvement over the rule baseline is also informative. The rule baseline is precise because many dark patterns advertise their intent through explicit phrases such as 'limited time', 'only left', or social-proof counts. Yet it misses nearly one third of positives because many examples are not built from a fixed keyword list. Learned sparse encoders recover those cases by combining local lexical context, character fragments, and decision-boundary evidence from many strings. This is a practical advantage for compliance tools: the model can detect risky copy that does not match a hand-written checklist, while the risk-card category still gives the user a recognizable explanation.

The results are consistent with the earlier ec-darkpattern literature, but they are not a direct reproduction of those tables. The direct predecessor compared rules, n-grams, BERT-style encoders, RoBERTa-style encoders, and decoder-style classifiers under a fixed split (Xu et al., 2025). This study uses the same text records but changes the evaluation emphasis: page-grouped folds, risk-card scoring, and a visual-data gate for a multimodal pipeline. The strong performance of sparse encoders is useful because it gives a compact, fast, and inspectable baseline that can be embedded in a browser extension or compliance scanner without relying on large external models.

The grouped evaluation makes these numbers more conservative than a simple random split. The dataset has 1,248 unique page identifiers for 2,356 records, so some pages contribute multiple snippets. If related snippets were split across folds, a model could benefit from page-specific repetition. The grouped folds therefore test whether the model generalizes across page contexts rather than simply memorizing a local vocabulary. The resulting F1 near 0.958 indicates that the strongest text models learn transferable microcopy cues rather than only page-level artifacts.

Table VII. Binary dark-pattern detection results, mean across five grouped folds.

model	accuracy	precision	recall	f1	auc	brier
Word+Char-LinearSVM	0.958	0.966	0.950	0.958	0.988	-
Probability ensemble	0.956	0.958	0.955	0.956	0.990	0.035
Word+Char-SGD	0.955	0.959	0.950	0.954	0.985	0.038
Word-TFIDF-SGD	0.950	0.952	0.947	0.950	0.981	0.043
Char-TFIDF-SGD	0.950	0.956	0.944	0.950	0.985	0.039
Word+Char+Stats-SGD	0.937	0.923	0.960	0.940	0.985	0.059
Complement NB	0.927	0.918	0.939	0.928	0.968	0.083
Stats-LR	0.921	0.955	0.885	0.918	0.959	0.061
Cue-rule baseline	0.815	0.942	0.671	0.783	0.825	0.160

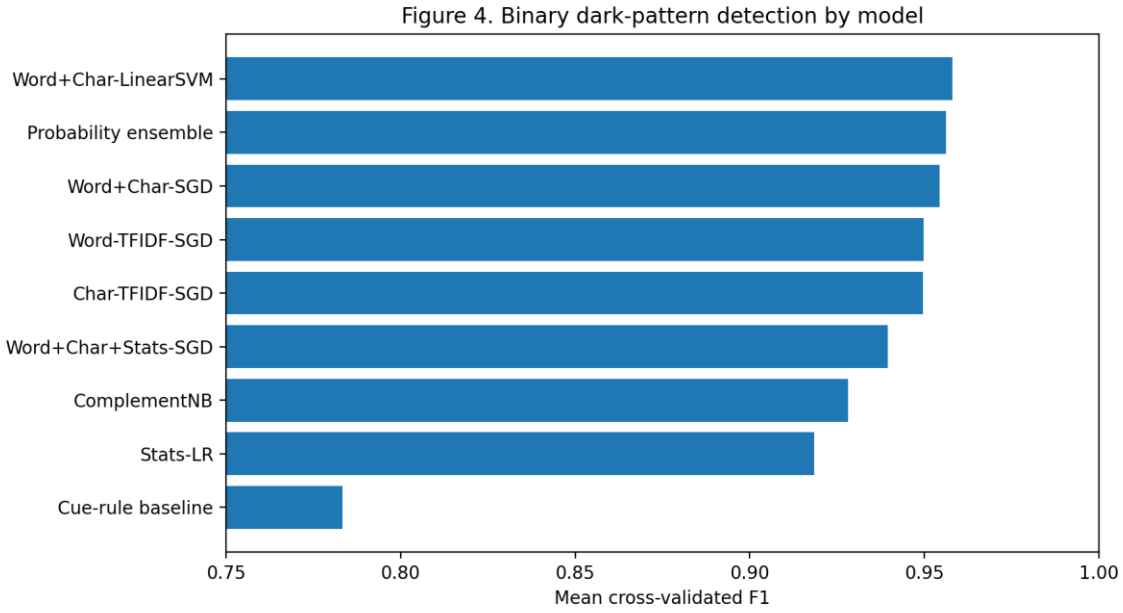

Fig. 4. Binary model comparison by mean cross-validated F1.

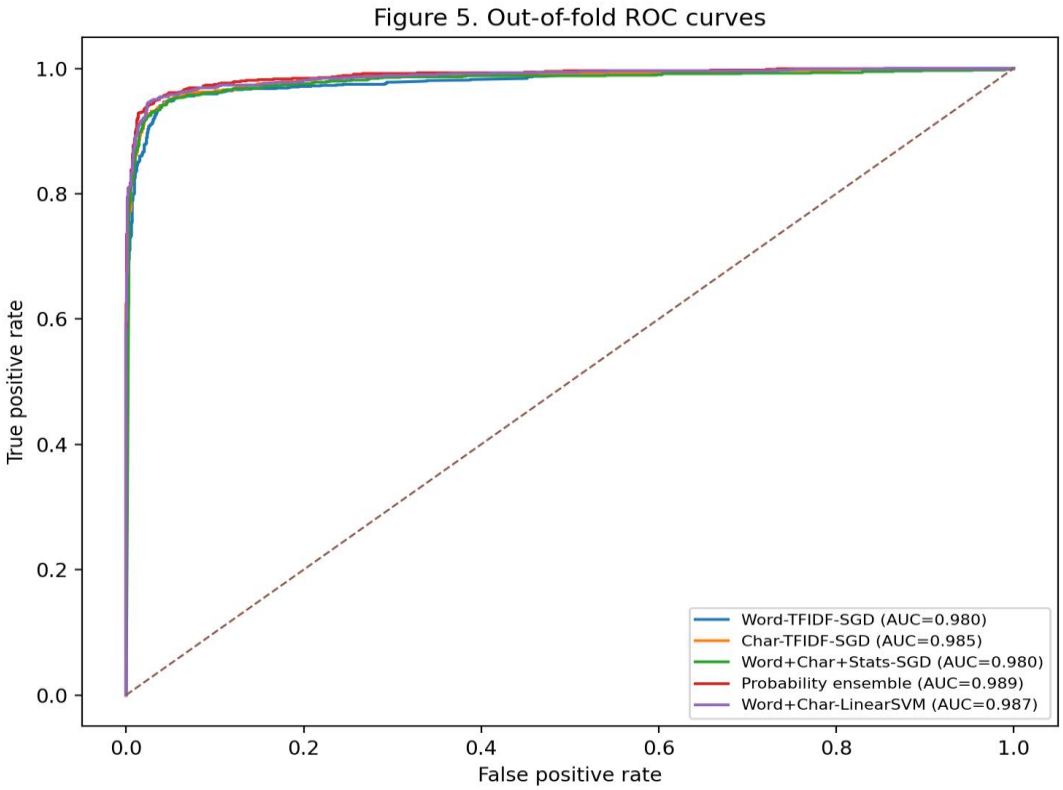

Fig. 5. Out-of-fold ROC curves for the main detectors.

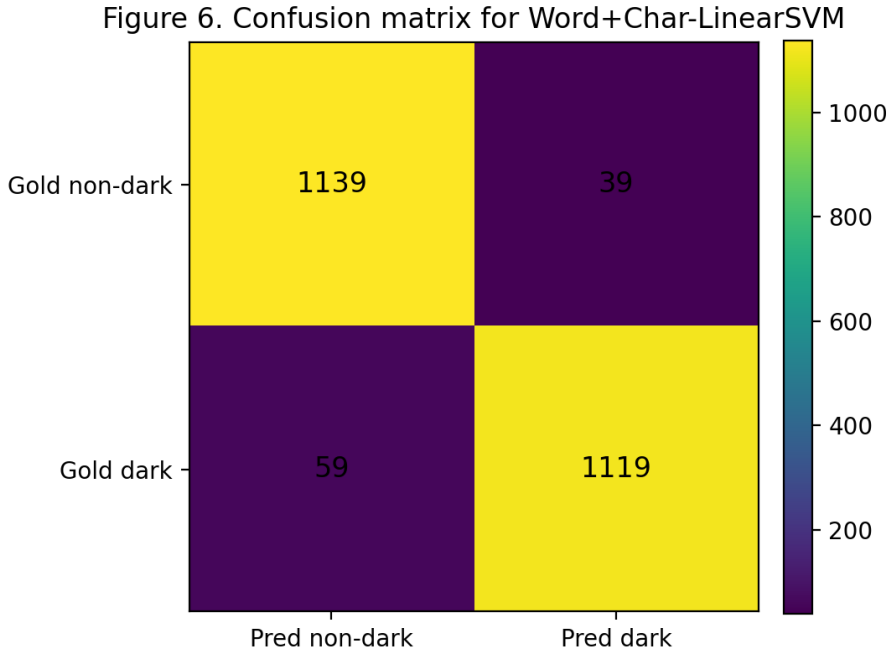

Fig. 6. Binary confusion matrix for the highest-F1 detector.

Table VIII frames the same models as an ablation study. The progression from statistical features to word encoding, character encoding, combined encoding, and probability ensembling demonstrates that the largest gain comes from sparse lexical encoders rather than hand-engineered counts. The ensemble does not beat the SVM on F1, but its AUC and Brier score make it the most suitable input for the risk-card layer. This tradeoff is common in safety-oriented detection: the hard classifier and the calibrated scorer need not be the same component.

Table VIII. Encoder ablation results.

model	accuracy	precision	recall	f1	auc	brier
Probability ensemble	0.956	0.958	0.955	0.956	0.990	0.035
Word+Char-SGD	0.955	0.959	0.950	0.954	0.985	0.038
Word-TFIDF-SGD	0.950	0.952	0.947	0.950	0.981	0.043
Char-TFIDF-SGD	0.950	0.956	0.944	0.950	0.985	0.039
Word+Char+Stats-SGD	0.937	0.923	0.960	0.940	0.985	0.059
Stats-LR	0.921	0.955	0.885	0.918	0.959	0.061

The category classifier is evaluated only on positive examples because non-dark strings do not have a dark-pattern category. Table IX shows that Char-TFIDF-SGD obtains the best weighted F1, 0.955, while Word-TFIDF-SGD obtains the highest macro F1, 0.817. Weighted metrics are high because the largest classes are well separated. Macro metrics are lower because forced action has only four examples, sneaking has twelve, and obstruction has twenty-seven. The category confusion matrix in Fig. 7 confirms that most errors are concentrated in rare or semantically overlapping classes. Misdirection is also harder than scarcity or social proof, because the risky mechanism depends on how an option is framed rather than on a distinctive lexical cue.

Table IX. Positive-category classification results.

model	accuracy	macro precision	macro recall	macro f1	weighted f1
Char-TFIDF-SGD	0.955	0.794	0.831	0.806	0.955

model	accuracy	macro precision	macro recall	macro f1	weighted f1
Word-TFIDF-SGD	0.947	0.805	0.841	0.817	0.947
Word+Char+Stats-SGD	0.924	0.657	0.684	0.666	0.925

Figure 7. Positive-category confusion matrix for Char-TFIDF-SGD

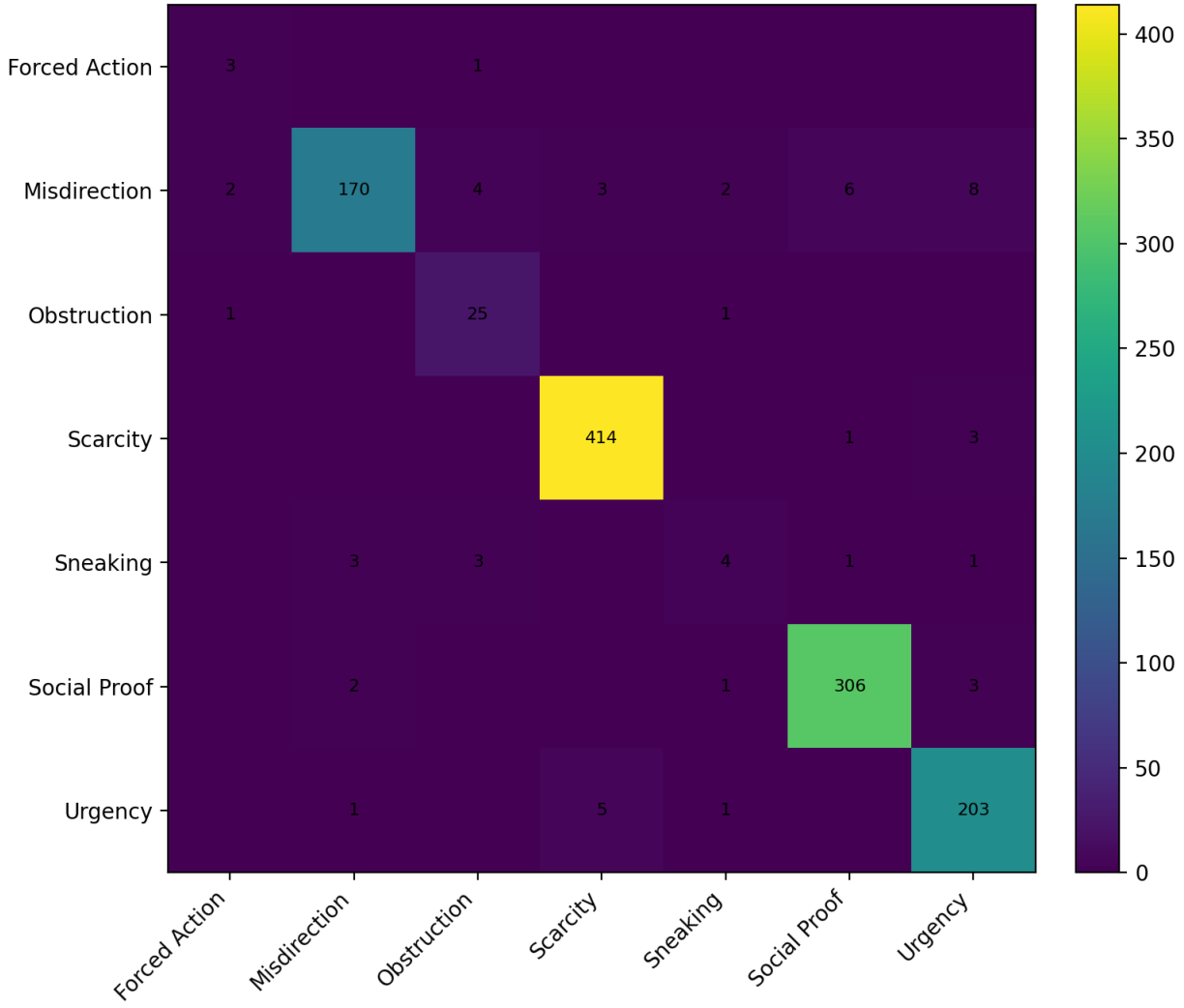

Fig. 7. Positive-category confusion matrix for Char-TFIDF-SGD.

Table X breaks down binary errors by gold category. Scarcity, social proof, urgency, and obstruction all have recall above 0.96 under the best binary detector. Misdirection recall is lower at 0.815, while sneaking and forced action have the weakest recall. These two rare classes should be treated carefully in deployment because a high overall F1 can mask sparse but high-harm categories. The error profile also suggests where additional visual information could help. Sneaking and forced action often depend on hidden costs, mandatory steps, or preselected controls. A YOLO branch that sees checkbox states, modal placement, and dominant call-to-action buttons would provide evidence that the microcopy alone may miss.

False positives have a different interpretation. A false positive on ordinary marketing copy may be acceptable in a back-office audit queue, but it is more problematic in a live browser warning because repeated low-value warnings can reduce trust. The best binary detector has 39 false positives out of 1,178 non-dark records, which is a manageable rate for audit workflows. In a user-facing setting, the probability ensemble could be thresholded above 0.5 to reduce warning frequency. The tradeoff would be fewer false positives at the cost of missing some subtle positives, especially in the harder misdirection and sneaking categories.

Table X. Error analysis by gold category for the best binary detector.

category	n	true positive	true negative	false positive	false negative	category recall
Not Dark Pattern	1178	0	1139	39	0	NA
Scarcity	418	414	0	0	4	0.99
Social Proof	312	305	0	0	7	0.978
Urgency	210	205	0	0	5	0.976
Misdirection	195	159	0	0	36	0.815
Obstruction	27	26	0	0	1	0.963
Sneaking	12	8	0	0	4	0.667
Forced Action	4	2	0	0	2	0.5

The risk-card evaluation is intentionally separated from raw classification accuracy. Table XI reports risk-tier accuracy of 0.593 and macro F1 of 0.364, with 1.000 schema validity. These values show that the card generator reliably emits complete outputs, but mapping classifier probabilities to consumer-harm tiers remains difficult. The low macro F1 is not a formatting failure; it reflects the stricter task of translating a binary/category prediction into Low, Moderate, High, or Critical risk. The tier mapping is conservative and relies on fixed severity weights. In future deployment, those weights should be learned

from consumer studies or regulator-reviewed harm scales rather than chosen from taxonomy severity alone. Any deployed threshold should also be evaluated as a policy. Offline counterfactual research shows why logged-policy overlap and uncertainty diagnostics are prerequisites before claiming that a new intervention will improve outcomes (Mu et al., 2025).

This separation between detection and warning design is essential. A classifier can say that a string resembles scarcity, but a consumer warning must decide whether the evidence is severe enough to interrupt the user. The same phrase can be harmless in an inventory notice and manipulative when paired with a countdown, a preselected add-on, or a hidden subscription term. The risk card therefore should be viewed as a decision-support artifact rather than a final legal conclusion. Cross-domain safety-card design likewise pairs risk tiers with review cues and scope limits rather than presenting the model output as a final decision (Zhou et al., 2025).

Table XI. Risk-card evaluation.

metric	value
risk_tier_accuracy	0.593
risk_tier_macro_f1	0.364
risk_card_schema_validity	1.000
mean_predicted_risk_score	0.295

Table XII and Fig. 8 show how model output is converted into a consumer-facing card. A high-probability urgency or scarcity string produces a direct warning that the copy creates time or availability pressure. A social-proof string produces a different explanation, focused on popularity cues and perceived safety. The cards are short by design: they should alert a user without becoming another source of cognitive load. The recommended action is also standardized. For all non-low tiers, the card advises the consumer to review costs, consent, and alternatives before proceeding.

Table XII. Sample risk-card outputs from highest-probability records.

text	gold category	pred category	probability	risk tier	consumer risk explanation
Hurry! Only 1 Left In Stock	Scarcity	Scarcity	1.000	High	limited-availability pressure
Hurry, only 2 left in stock	Scarcity	Scarcity	1.000	High	limited-availability pressure

text	gold category	pred category	probability	risk tier	consumer risk explanation
Act now and save $12.99. Limited time only!	Urgency	Urgency	1.000	High	time-pressure cue
Hurry! Only 6 left in stock	Scarcity	Scarcity	1.000	High	limited-availability pressure
Hurry! Only Limited Items Left	Scarcity	Scarcity	1.000	High	limited-availability pressure

Figure 8. Example risk card produced from model output

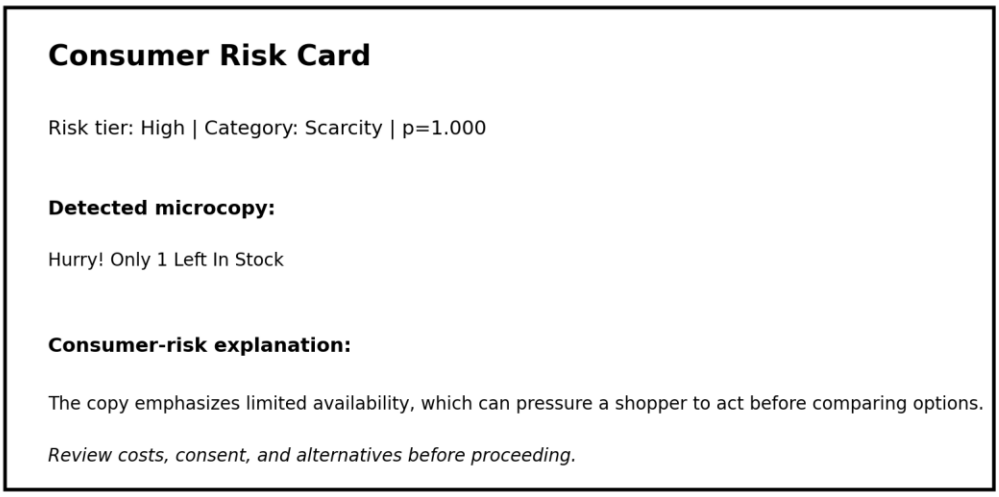

Fig. 8. Example consumer-risk card generated from model output.

The visual-data audit deserves separate discussion. The DarkLens documentation is aligned with a YOLO object-detection task: it describes image-plus-label data, five UI component classes, and a screenshot setting. However, the materialized archive contains a Git-LFS pointer and no image or label files after extraction. That result is important for reproducibility. Object detectors can be trained only when the image pixels and annotation text files are present. Reporting a YOLO mAP score without those records would conflate a system design with a completed detector benchmark. The implemented pipeline therefore stops at the correct boundary: it validates the class contract and file availability, then continues with the text and risk-card experiments that can be computed from the available data.

Taken together, the experiments support a practical

architecture. Text encoders can identify most e-commerce dark-pattern microcopy with strong held-out performance. Grouped evaluation avoids overly optimistic page-level leakage. Probability scores can drive risk cards, but severity tiers need better calibration. A future visual dataset containing actual screenshots and YOLO annotations can be plugged into the first layer without changing the text or risk-card code. The strongest near-term use case is therefore a browser or audit tool that extracts visible microcopy, scores it with a sparse text encoder, and displays a concise risk card while the visual branch is trained when full screenshot labels are available.

The architecture also supports audit transparency. Every warning can be traced to a row identifier, a predicted category, a score, and a deterministic explanation template. Cross-domain prototype-retrieval work shows how an alert can be contrasted with nearby normal cases to make the explanation auditable (Xin, 2025a). This matters in consumer-protection contexts because an organization may need to justify why a page was flagged. A black-box warning without evidence is hard for designers to remediate and hard for consumers to trust. By contrast, a risk card that names the triggering string and predicted mechanism gives both sides a concrete starting point for review.

Limitations and Future Research

The most important limitation is the visual branch. The DarkLens-YOLO package profile declares a screenshot object-detection dataset, but the extracted experimental package contains a Git-LFS pointer rather than the underlying screenshots and YOLO label files. The study therefore does not report visual mAP, detection speed, or bounding-box localization metrics. The paper reports the visual audit because it is an empirical result of the data pipeline and because object-detection claims require raw image-label pairs.

The text experiments are limited to e-commerce microcopy. Dark patterns in social media, games, news, finance, travel, or public-service portals may use different language, consent flows, and harm mechanisms. Trust-calibrated multilingual RAG research likewise treats deployment context and unsupported-evidence handling as distinct evaluation dimensions (Chen & Xu, 2026). The corpus is balanced by construction, while real web traffic is not. A deployment system should therefore recalibrate decision thresholds on the target domain and monitor false positives on ordinary marketing copy.

The category task is affected by severe class imbalance. Forced action has four examples, sneaking has twelve, and obstruction has twenty-seven. Weighted F1 is high, but macro F1 is a better signal for rare classes. Future datasets should intentionally collect more rare high-harm categories and include screenshots that reveal preselection, hidden costs, blocked exits, and visual hierarchy.

The risk-card generator is deterministic. This choice makes the empirical tables reproducible, but it does not measure the full behavior of a free-form LLM. An LLM rendering layer could improve fluency and adapt explanations to context, but it would need guardrails, audit logs, and tests for hallucinated claims. The current risk-card results should therefore be read as structured explanation performance rather than as a user-study validation of warning effectiveness.

The experiment also does not use DOM structure, CSS, color contrast, click paths, or dwell-time signals. Those features can be important in real UI/UX detection. For example, a refusal link may be visible in text but visually de-emphasized, or a cancel action may exist but require multiple screens. Screenshot detection and DOM analysis would help capture these mechanisms once complete data are present.

Finally, the experiment evaluates model outputs against dataset labels rather than direct consumer harm. A dark-pattern warning that is statistically correct may still be unclear, annoying, or ignored. Human-versus-LLM design-critique findings show that automated feedback priorities can differ substantially from human judgment unless region evidence is incorporated (Y. Li et al., 2025). Consumer studies and regulator review are therefore needed to validate severity weights, warning thresholds, and the language of recommended actions.

High-stakes counseling interfaces provide another useful boundary: a therapist-facing copilot can retrieve and rank support while leaving the final judgment with a human professional (Y. Zhang & H. Zhang, 2025a). A consumer-facing warning should likewise preserve user choice, disclose uncertainty, and allow the user to inspect or dismiss the intervention.

Concluding Synthesis

This paper presented a multimodal design for dark-pattern detection that links YOLO-style screenshot localization, microcopy text encoders, and consumer-risk explanations. The empirical text evaluation on 2,356 ec-darkpattern strings shows that sparse word and character encoders are strong baselines under page-grouped cross-validation. The best binary model reaches 0.958 F1 and 0.988 AUC, while a probability ensemble reaches 0.990 AUC and supplies calibrated scores for risk cards. Category classification is reliable for common classes but weaker for rare high-harm categories, especially forced action and sneaking.

The results also clarify the role of visual data. The DarkLens-YOLO package profile is consistent with a screenshot object-detection task, but the raw image-label records are not materialized in the archive used for this run. The correct experimental outcome is therefore a visual data-readiness audit rather than an object-detection benchmark. The code keeps the YOLO branch ready for training once screenshots and labels are present, while the reported detection metrics come from the text and risk-card experiments.

For practical consumer protection, the most promising path is not a single classifier but a pipeline: detect risky microcopy, localize the relevant UI element when screenshots are available, calibrate the probability, and explain the predicted consumer risk in a compact card. Such a system can support browser warnings, compliance audits, and researcher workflows while leaving room for human judgment about context and harm. Before rollout, warning thresholds and display policies should be stress-tested with conservative off-policy evaluation and then confirmed in prospective user studies (Ye et al., 2026).

References

Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y. M. (2020). YOLOv4: Optimal speed and accuracy of object detection. *arXiv*. https://doi.org/10.48550/arXiv.2004.10934

Bösch, C., Erb, B., Kargl, F., Kopp, H., & Pfattheicher, S. (2016). Tales from the dark side: Privacy dark strategies and privacy dark patterns. *Proceedings on Privacy Enhancing Technologies, 2016*(4), 237–254. https://doi.org/10.1515/popets-2016-0038

Brignull, H. (2011, November 1). *Dark patterns: Deception vs. honesty in UI design*. A List Apart. https://alistapart.com/article/dark-patterns-deception-vs-honesty-in-ui-design/

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901. https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

Chen, Y., & Li, M. (2025). From hand-drawn sketches to interactive web prototypes: A reproducible vision-language approach with structural and visual consistency evaluation. *Journal of Technology Informatics and Engineering, 4*(2), 364–384. https://doi.org/10.51903/jtie.v4i2.490

Chen, Y., & Xu, H. (2026). Trust-calibrated multilingual RAG for humanitarian information platforms: Empirical evaluation on OMoS-QA for migration information access. *International Journal of Graphic Design, 4*(1), 141–164. https://doi.org/10.51903/ijgd.v4i1.3552

Cialdini, R. B. (2009). *Influence: Science and practice* (5th ed.). Pearson.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Federal Trade Commission. (2022). *Bringing dark patterns to light*. https://www.ftc.gov/reports/bringing-dark-patterns-light

Gray, C. M., Kou, Y., Battles, B., Hoggatt, J., & Toombs, A. L. (2018). The dark (patterns) side of UX design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Article 534, pp. 1–14). Association for Computing Machinery. https://doi.org/10.1145/3173574.3174108

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning* (Vol. 70, pp. 1321–1330). PMLR. https://proceedings.mlr.press/v70/guo17a.html

Jin, J. (2025a). Evidence-chain reliable RAG: Hallucination detection, source attribution, and deterministic provenance explanations. *Journal of Technology Informatics and Engineering, 4*(2), 520–533. https://doi.org/10.51903/jtie.v4i2.535

Jin, J. (2025b). LLM-style evidence cards for scientific search interfaces: A UI/UX design framework for retrieval transparency, ranking trust, and visual evidence hierarchy. *International Journal of Graphic Design, 3*(2), 397–414. https://doi.org/10.51903/ijgd.v3i2.3698

Li, C., Zhou, B., & Gao, K. (2025). Risk-calibrated patient-facing AI safety cards: A UI/UX benchmark for explainable medical AI response interfaces. *International Journal of Graphic Design, 3*(2), 381–394. https://doi.org/10.51903/ijgd.v3i2.3709

Li, Y., & Lu, S. (2025). Language-guided feature selection for DDoS and intrusion detection on CICIDS2017. *Journal of Technology Informatics and Engineering, 4*(1), 284–305. https://doi.org/10.51903/jtie.v4i1.531

Li, Y., Lu, S., & Zhao, L. (2025). LLM-as-design-critic: Aligning AI-generated UI feedback with human graphic design judgment. *International Journal of Graphic Design, 3*(1), 196–215. https://doi.org/10.51903/ijgd.v3i1.3661

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In *Computer Vision–ECCV 2014* (pp. 740–755). Springer. https://doi.org/10.1007/978-3-319-10602-1_48

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv*. https://doi.org/10.48550/arXiv.1907.11692

Luguri, J., & Strahilevitz, L. J. (2021). Shining a light on dark patterns. *Journal of Legal Analysis, 13*(1), 43–109. https://doi.org/10.1093/jla/laaa006

Mathur, A., Acar, G., Friedman, M. J., Lucherini, E., Mayer, J., Chetty, M., & Narayanan, A. (2019). Dark patterns at scale: Findings from a crawl of 11K shopping websites. *Proceedings of the ACM on Human-Computer Interaction, 3*(CSCW), Article 81, 1–32. https://doi.org/10.1145/3359183

Mu, J., Ye, T., & Patel, P. (2025). Offline counterfactual evaluation for advertising and recommendation slot policies: A reproducible study on the Open Bandit Dataset (Small). *Journal of Technology*

Informatics and Engineering, 4(3), 521–543. https://doi.org/10.51903/jtie.v4i3.500

Narayanan, A., Mathur, A., Chetty, M., & Kshirsagar, M. (2020). Dark patterns: Past, present, and future. *Communications of the ACM, 63*(9), 42–47. https://doi.org/10.1145/3397884

Nouwens, M., Liccardi, I., Veale, M., Karger, D., & Kagal, L. (2020). Dark patterns after the GDPR: Scraping consent pop-ups and demonstrating their influence. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). Association for Computing Machinery. https://doi.org/10.1145/3313831.3376321

Organisation for Economic Co-operation and Development. (2022). *Dark commercial patterns* (OECD Digital Economy Papers No. 336). OECD Publishing. https://doi.org/10.1787/44f5e846-en

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., … Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems, 35*, 27730–27744. https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html

Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. *arXiv*. https://doi.org/10.48550/arXiv.1804.02767

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 779–788). https://doi.org/10.1109/CVPR.2016.91

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (pp. 3982–3992). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778

Soe, T. H., Nordberg, O. E., Guribye, F., & Slavkovik, M. (2020). Circumvention by design: Dark patterns in cookie consent for online news outlets. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction* (Article 19, pp. 1–12). Association for Computing Machinery. https://doi.org/10.1145/3419249.3420132

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998–6008. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Wang, C.-Y., Bochkovskiy, A., & Liao, H.-Y. M. (2023). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 7464–7475). https://doi.org/10.1109/CVPR52729.2023.00721

Xin, Q. (2025a). Explaining OpenStack failure-injection log anomalies with retrieved normal prototypes. *Emerging Information Science and Technology, 6*(2), 125–146. https://doi.org/10.18196/eist.v6i2.31232

Xin, Q. (2025b). Uncertainty-aware late fusion for 3D perception (confidence calibration + fusion rule learning). *Journal of Technology Informatics and Engineering, 4*(1), 215–238. https://doi.org/10.51903/jtie.v4i1.485

Xin, Q. (2026). Log anomaly detection with conformal alert control and evidence-grounded incident ticket generation. *Aviation Electronics, Information Technology, Telecommunications, Electricals, and Controls (AVITEC), 8*(2), 247–264. https://doi.org/10.28989/avitec.v8i2.3974

Xu, H., Chen, Y., & Med, A. (2025). Automatic detection and explanation of dark patterns from interface microcopy: Empirical comparison of BERT-style encoders, RoBERTa-style encoders, and LLM-style decoders on the ec-darkpattern dataset. *Journal of Technology Informatics and Engineering, 4*(3), 590–612. https://doi.org/10.51903/jtie.v4i3.491

Yada, Y., Feng, J., Matsumoto, T., Fukushima, N., Kido, F., & Yamana, H. (2022). Dark patterns in e-commerce: A dataset and its baseline evaluations. In *2022 IEEE International Conference on Big Data* (pp. 3015–3022). IEEE. https://doi.org/10.1109/BigData55660.2022.10020800

Ye, T., Mu, J., & Hunter, J. (2026). Off-policy evaluation and conservative policy selection for slot-level dynamic bidding and ranking on the Open Bandit Dataset (Small). *Journal of Technology Informatics and Engineering, 5*(1), 178–199. https://doi.org/10.51903/jtie.v5i1.503

Zhang, Y., & Zhang, H. (2025a). A therapist-facing session copilot for live counseling support: Reasoning-guided retrieval and ranking from multi-turn counseling dialogues. *Journal of Technology Informatics and Engineering, 4*(2), 464–486. https://doi.org/10.51903/jtie.v4i2.547

Zhang, Y., & Zhang, H. (2025b). Visualizing the right counseling support: Evidence-linked recommendation cards for explainable mental health intake interfaces. *International Journal of Graphic Design, 3*(1), 214–229. https://doi.org/10.51903/ijgd.v3i1.3722

Zhou, B., Li, C., & Liu, L. (2025). Risk-calibrated patient-facing AI safety cards: A UI/UX design framework for rubric-based medical risk communication. *International Journal of Graphic Design, 3*(2), 365–380. https://doi.org/10.51903/ijgd.v3i2.3696