myresearchgo Volume 2, Issue 5, May 2026, ISSN: 3107-3816 (Online)

Dark Pattern Sentinel: A Hybrid DOM Heuristics and LLM-Based Real-Time Detection System for Deceptive Web UI Patterns

Mayur Hemnath Karve

Master of Science in Computer Applications (MSc CA)

Department of Computer Science and Information Technology

INSTITUTE OF BUSINESS STUDIES AND RESEARCH

CBD Belapur, Navi Mumbai – 400614

Affiliated to

TILAK MAHARASHTRA VIDYAPEETH

Pune, Maharashtra 411 037

(Deemed to be University u/s 3 of UGC Act, 1956)

PRN: 40724604050

Academic Year: 2025 – 2026ABSTRACT

Dark patterns are deceptive user-interface design choices engineered to manipulate users into actions that serve business interests at the expense of user autonomy, financial well-being, or privacy. As e-commerce and platform economies expand globally, dark patterns have proliferated across websites and mobile applications, drawing regulatory attention from the European Union's Digital Services Act (DSA 2022), India's Consumer Protection Act and CCPA Guidelines (2023), and the United States Federal Trade Commission (FTC Click-to-Cancel Rule, 2024). Despite this regulatory momentum, automated real-time detection of dark patterns at the browser level remains largely unsolved.

This paper presents Dark Pattern Sentinel (DPS), a research-grade Google Chrome extension implementing a novel two-stage hybrid detection pipeline. In the first stage, fourteen purpose-built deterministic heuristic scanners parse the live Document Object Model (DOM) to flag elements matching known dark-pattern families. In the second stage, a Large Language Model (Llama 3.3 70B via Groq) provides semantic verification for detections with intermediate confidence (45–78%), dramatically reducing false positives without incurring unnecessary API cost. Detected patterns are highlighted in real-time using a Shadow DOM overlay, and a logarithmic trust-scoring algorithm aggregates findings into a 0–100 page-level trust score with letter grades. The system covers fourteen dark-pattern categories spanning all major regulatory taxonomies, and exports structured JSON evidence suitable for legal or research documentation.

Preliminary evaluation on a purposely-constructed adversarial demo page containing nine dark-pattern types demonstrates detection within one second of page load, with per-element red highlight boxes and explanatory

tooltips. The system is designed for use in longitudinal e-commerce studies, cross-jurisdictional analysis, and regulatory enforcement support, establishing a foundation for automated dark-pattern auditing at scale.

Keywords: dark patterns, deceptive design, browser extension, DOM heuristics, large language models, Groq, Llama 3.3, trust score, UI manipulation, CCPA, GDPR, DSA, digital consumer protection, Chrome Manifest V3

1. INTRODUCTION

The modern web is saturated with user interfaces designed not to help users accomplish their goals, but to nudge, trick, or coerce them into actions they would not otherwise choose. These deceptive design strategies — collectively termed dark patterns — were first systematically catalogued by Harry Brignull in 2010. Since then, academic research has documented their extraordinary prevalence: a landmark 2019 study by Mathur et al. found dark patterns present in over 11,000 of the largest shopping websites, spanning misleading countdown timers, hidden subscription fees, pre-checked data-sharing consent boxes, and asymmetric cancellation flows.

The economic consequences are significant. Research by the Norwegian Consumer Council (2018) demonstrated that Facebook's dark-pattern-laden privacy settings led the majority of users to accept broader data sharing than they intended. Hidden drip pricing in travel and ticketing has been shown to increase effective prices by 20–30% over advertised rates. Confirmshaming — phrasing decline buttons as self-deprecating statements — increases email opt-in rates by measurable margins. The aggregate consumer harm is substantial.

Regulators have responded. The European Union's Digital Services Act (Regulation EU 2022/2065), effective from August 2023, explicitly prohibits very large online platforms from employing dark patterns under Article 25. India's Central Consumer Protection Authority (CCPA) issued binding Guidelines for Prevention and Regulation of Dark Patterns in November 2023, covering eleven specific pattern types. The United States FTC updated its Negative Option Rule in 2024, establishing a "click-to-cancel" requirement that directly targets the Roach Motel dark pattern. Yet regulatory enforcement requires first identifying violations — a task that today is almost entirely manual, slow, and unscalable.

This paper addresses the detection gap by presenting Dark Pattern Sentinel (DPS): a Manifest V3 Chrome extension that applies fourteen heuristic scanners augmented by Large Language Model (LLM) verification to detect dark patterns in real-time, on any website, without requiring server-side infrastructure beyond an optional Groq API key. The system provides immediate visual feedback, a quantitative trust score, and structured evidence output — enabling both end-user awareness and researcher data collection.

1.1 Motivation

Existing dark pattern detection tools are either manual checklists, academic crawlers that operate offline in batch mode, or browser extensions with limited pattern coverage. No publicly available tool combines: (a) comprehensive real-time DOM scanning across all major regulatory pattern categories, (b) LLM-based

disambiguation of genuinely ambiguous cases, (c) a quantitative scoring model aligned with regulatory severity weights, and (d) evidence export for legal documentation. DPS fills this gap.

1.2 Research Objectives

- Design and implement a comprehensive taxonomy of 14 dark patterns aligned with CCPA 2023, EU DSA, GDPR, and FTC regulations.
- Build deterministic heuristic scanners for each pattern category operable entirely within a browser extension content script.
- Integrate LLM-based semantic verification to reduce false positives in ambiguous heuristic detections.
- Develop a logarithmic trust-scoring algorithm that reflects regulatory severity weights.
- Validate the system on a purposely-constructed adversarial test page and assess detection latency.
- Provide structured JSON evidence export suitable for research and legal documentation purposes.

1.3 Contributions

The principal contributions of this work are:

1. A fourteen-category dark pattern taxonomy with per-pattern severity weights, regulatory citations, and user-facing explanations, grounded in Mathur et al. (2019) and current legislation.
2. A hybrid two-stage detection pipeline that combines deterministic DOM heuristics with selective LLM semantic verification, with aggressive caching and a circuit breaker for production reliability.
3. A logarithmic trust scoring model with confidence dampening and multi-instance penalty attenuation.
4. A full Manifest V3 browser extension implementation with Shadow DOM overlay, real-time MutationObserver-driven scanning, and structured evidence export.
5. An open research telemetry framework for longitudinal per-domain dark pattern studies.

2. BACKGROUND AND RELATED WORK

2.1 Dark Patterns: Definition and History

Harry Brignull coined the term "dark patterns" in 2010 to describe user interface design choices that trick users into unintended actions. Unlike design errors, dark patterns are deliberate — they are carefully crafted by UX professionals to exploit cognitive biases including loss aversion, social proof, anchoring, and decision fatigue. Brignull catalogued twelve archetypes including Trick Questions, Roach Motel, Disguised Ads, Bait and

Switch, and Misdirection, which formed the foundational taxonomy that later academic and regulatory work extended.

Gray et al. (2018) extended the taxonomy through UX practitioner interviews and identified five high-level strategies — Nagging, Obstruction, Sneaking, Interface Interference, and Forced Action — that subsume Brignull's archetypes. This stratified model is now widely cited in regulatory frameworks. The FTC's 2022 Bringing Dark Patterns to Light report adopted similar stratification while mapping patterns to specific consumer protection violations.

2.2 Regulatory Landscape

The regulatory environment for dark patterns has matured significantly since 2022. The European Union's Digital Services Act (Regulation EU 2022/2065) — effective for very large platforms from August 2023 — prohibits under Article 25 any interface design that 'distorts or impairs the ability of recipients of the service to make free and informed decisions.' The European Data Protection Board (EDPB) issued Guidelines 03/2022 mapping dark patterns in social media to GDPR violations, providing a direct enforcement linkage.

In India, the Central Consumer Protection Authority promulgated the Guidelines for Prevention and Regulation of Dark Patterns (November 2023) under the Consumer Protection Act 2019. These guidelines are binding and cover false urgency, basket sneaking, subscription traps, bait and switch, drip pricing, disguised advertisements, nagging, trick questions, saas billing, and rogue malwares. The Digital Personal Data Protection Act (DPDP Act 2023) further reinforces this by requiring meaningful, granular consent — directly targeting pre-selected consent and privacy zuckering patterns.

In the United States, the FTC's updated Negative Option Marketing Rule (2024) establishes a Click-to-Cancel requirement mandating that cancellation be as simple as sign-up — a direct response to the Roach Motel dark pattern. Separately, the FTC's Fake Reviews and Testimonials Rule (2024) targets manufactured social proof and fake activity indicators.

2.3 Prior Detection Systems

Mathur et al. (2019) performed the first large-scale automated dark pattern detection study, crawling 11,000 shopping websites using a combination of DOM analysis and trained classifiers. Their work identified 1,818 instances of dark patterns across 1,254 websites and provided the severity taxonomy that this project's scoring model directly adopts. However, their approach operated as a batch offline crawler, not a real-time browser tool.

Nouwens et al. (2020) studied cookie consent interfaces across 10,000 UK websites using automated analysis, finding that 57% failed GDPR standards for valid consent due to asymmetric interface design — a finding directly relevant to the visual interference and privacy zuckering scanners in this work.

Several browser extensions address dark patterns, but none comprehensively. Privacy Badger and uBlock Origin focus on tracking, not UI manipulation. Consent-O-Matic (Nouwens et al.) targets cookie consent

specifically. No existing public extension covers all fourteen pattern types with LLM verification and a quantitative trust score.

2.4 Large Language Models for UI Analysis

Recent work has demonstrated the utility of LLMs for user interface understanding tasks. GPT-4V and similar multimodal models have shown capability in identifying deceptive UI elements from screenshots. However, screenshot-based approaches require either server-side screenshot capture or expensive multimodal inference. The DPS approach employs text-only LLM verification on extracted DOM content — dramatically cheaper, faster, and more privacy-preserving than multimodal alternatives, while remaining sufficient for semantic disambiguation of textual dark patterns.

Groq's inference platform provides Llama 3.3 70B responses typically under 800ms for the short prompts used in this system, making real-time integration feasible. The selective LLM invocation strategy — only for intermediate-confidence detections — ensures that average API cost per page scan remains negligible.

3. SYSTEM DESIGN AND ARCHITECTURE

3.1 Design Principles

The DPS system was designed around six core principles derived from the requirements of a production browser extension:

- Privacy First: All scanning runs locally in the browser. The Groq API receives only short suspicious text snippets (never URLs, screenshots, or full page content). Research telemetry is stored locally in chrome.storage.local and never transmitted.

- Selective LLM Invocation: LLM calls are reserved for detections with intermediate heuristic confidence (45–78%). High-confidence and low-confidence detections are handled entirely by heuristics, minimising API latency and cost.

- Shadow DOM Isolation: Overlay elements are injected inside a Shadow DOM tree, ensuring the page's own CSS cannot interfere with DPS visual indicators — a critical requirement on content-heavy sites with aggressive style resets.

- MutationObserver-Driven Reactivity: Dynamic pages (SPAs, infinite scroll, AJAX-loaded content) are handled by observing DOM mutations and re-running affected scanners on newly inserted subtrees.

- Graceful Degradation: LLM unavailability (no API key, Groq downtime, circuit breaker open) causes no functional degradation — the system continues operating with heuristic-only detections, clearly indicating LLM status in the popup.

• Reproducible Evidence: Every detection carries an element selector, evidence text snippet, heuristic confidence score, and optional LLM reasoning. The full detection set is exportable as structured JSON for research and legal use.

3.2 High-Level Architecture

The system is divided into two runtime contexts: the Content Script, which runs in the page's renderer process and has direct DOM access; and the Service Worker (background script), which runs in the extension's background context and manages the LLM client, per-tab state, and badge updates. The two contexts communicate via the Chrome runtime message passing API.

Component	Technology	Responsibility
Content Script	JavaScript (ES2022)	DOM scanning, overlay rendering, MutationObserver
14 Heuristic Scanners	Regex + DOM API	Pattern-specific detection per taxonomy category
Trust Scorer	Logarithmic algorithm	Aggregates detections into 0-100 trust score
Overlay Engine	Shadow DOM + CSS	Isolated red highlight boxes and tooltips on page
Service Worker	Chrome MV3 SW	LLM proxying, per-tab caching, badge updates
Groq LLM Client	Llama 3.3 70B via API	Semantic verification of ambiguous detections
Popup Dashboard	HTML/CSS/JS (380px)	Per-pattern counts, evidence, scroll-to-element
Options Page	HTML/CSS/JS	Sensitivity config, scanner toggles, API key mgmt

Table 1: System Architecture Components

3.3 Detection Pipeline

The detection pipeline operates in five stages:

6. DOM Traversal: The content script performs an initial full-page scan on DOMContentLoaded. Each of the 14 scanners receives the full document root and returns an array of DetectionResult objects containing: pattern category, affected element reference, evidence text, and heuristic confidence (0–1).

7. Deduplication and Thresholding: Detections with confidence below 0.35 are discarded as noise. Duplicate detections on the same element for the same pattern are merged, retaining the highest confidence instance.

8. LLM Arbitration: Detections with heuristic confidence between 0.45 and 0.78 (the ambiguous zone) are sent to the Service Worker for LLM verification. The Groq client checks an in-memory cache (SHA-1 keyed on pattern plus evidence hash) before making an API call. LLM response confidence replaces heuristic confidence for the final verdict.

9. Overlay Rendering: Confirmed detections are passed to the overlay engine. For each detection, a red-bordered highlight div is injected into the Shadow DOM, absolutely positioned over the affected element using its bounding rectangle. A tooltip displaying the pattern name, explanation, severity, and legal citation is shown on hover.

10. Trust Scoring and Badge Update: The trust scorer computes the page-level score from all confirmed detections (see Section 4.2). The score and grade are sent to the Service Worker, which updates the browser action badge colour (green, amber, or red) and count.

3.4 Manifest V3 Constraints and Solutions

Chrome Manifest V3 imposes several architectural constraints relevant to the design. Service workers replace persistent background pages, requiring all extension state to be serialised to chrome.storage rather than held in long-lived memory. The DPS LLM client uses a dual-layer cache: an in-memory Map for the lifetime of the service worker process, and chrome.storage.local for persistence across worker restarts. The circuit breaker state (failure count and circuit-open timestamp) is likewise persisted to storage to survive service worker termination.

MV3 also prohibits remotely hosted code, requiring all detection logic — including the full taxonomy and scorer — to be bundled with the extension. This was achieved by structuring the codebase as native ES2022 modules declared in the manifest content_scripts configuration with module type.

4. IMPLEMENTATION

4.1 Dark Pattern Taxonomy

The DPS taxonomy covers fourteen dark pattern categories, selected to achieve full coverage of the CCPA 2023 guidelines, EU DSA Article 25, GDPR cookie consent rulings, and FTC enforcement actions. Each pattern is implemented as a discrete module in content/scanners/ and registered in lib/taxonomy.js with its severity rating, icon, short description, full explanation, and legal citation.

Dark Pattern	Severity	Scanner Module	Legal Anchor
False Urgency	HIGH (3)	urgency.js	EU DSA Art. 25; CCPA 2023
False Scarcity	HIGH (3)	scarcity.js	FTC; CCPA Guidelines 2023

Dark Pattern	Severity	Scanner Module	Legal Anchor
Manufactured Social Proof	MEDIUM (2)	social-proof.js	FTC Fake Reviews Rule
Confirmshaming	HIGH (3)	confirmshaming.js	EU DSA Art. 25
Hidden Costs / Drip Pricing	CRITICAL (4)	hidden-costs.js	CCPA §3; EU Omnibus
Sneak Into Basket	CRITICAL (4)	sneak-basket.js	EU CRD Art. 22
Pre-selected Consent	MEDIUM (2)	preselected.js	GDPR Art. 7(2); DPDP 2023
Roach Motel	HIGH (3)	roach-motel.js	FTC Click-to-Cancel 2024
Forced Continuity	HIGH (3)	forced-continuity.js	FTC Negative Option
Disguised Ads	MEDIUM (2)	disguised-ads.js	FTC .com Disclosures; ASCI
Trick Questions	MEDIUM (2)	trick-questions.js	EU DSA Art. 25
Visual Interference	MEDIUM (2)	visual-interference.js	EU DSA; GDPR rulings
Fake Countdown Timer	HIGH (3)	countdown.js	CCPA Guidelines 2023
Privacy Zuckering	HIGH (3)	privacy-zuckering.js	GDPR Art. 25; DPDP 2023

Table 2: Dark Pattern Taxonomy with Severity, Scanner Module, and Legal Anchor

4.2 Heuristic Scanner Design

Each scanner is a pure function that accepts a DOM root element and an options object (including sensitivity level and any user-configured exclusions) and returns an array of DetectionResult objects. Scanners do not mutate the DOM and are stateless, enabling safe concurrent execution.

4.2.1 Text-Based Scanners

Urgency, Scarcity, Confirmshaming, Forced Continuity, and Fake Countdown Timer scanners operate primarily on text content. They combine two detection strategies: a vocabulary-based approach using curated keyword lists (e.g., 'hurry', 'limited time', 'only X left', 'no thanks I hate saving money') with regex patterns for numeric context (e.g., countdown timer syntax), and a structural approach that checks element types and positions (e.g., a countdown timer must be in a visible, non-hidden element with appropriate role or class names).

Heuristic confidence is computed as a weighted sum of keyword match strength, keyword density relative to element text length, proximity to purchase-flow elements (Add to Cart buttons, checkout forms), and optional negative signals (elements inside admin panels or development tools are down-weighted). The final confidence value is clamped to [0.35, 0.95].

4.2.2 DOM-Structural Scanners

Preselected Options, Sneak Into Basket, Roach Motel, and Trick Questions scanners rely more heavily on DOM structure than text content. The Preselected Options scanner queries for all checked checkbox and radio inputs, then heuristically classifies each as marketing consent (via associated label text analysis), data sharing, or paid upgrade. The Sneak Into Basket scanner inspects form elements and hidden inputs on product and checkout pages for pre-checked items not explicitly placed by the user. The Roach Motel scanner compares the ease of subscription-initiation UI versus cancellation-navigation UI by counting steps to visible cancellation affordances.

4.2.3 Visual-Layer Scanners

Visual Interference, Disguised Ads, and Social Proof scanners combine text and CSS analysis. The Visual Interference scanner computes contrast ratios between CTA buttons (Accept, Subscribe, Buy) and their adjacent decline options (Reject, No thanks, Cancel) using computed CSS colors, flagging pages where the accept option has substantially higher visual prominence. The Disguised Ads scanner identifies elements with sponsored content labels that fail to meet minimum contrast and size requirements for disclosure legibility. The Hidden Costs scanner inspects checkout flow pages for fee line items that appear only after an initial product price has been displayed.

4.3 LLM Semantic Verification

The Groq LLM client in background/groq-client.js manages all interaction with the Groq inference API. The client uses Llama 3.3 70B Versatile with JSON response mode enabled, temperature 0.1, and a maximum of 220 output tokens per call. The system prompt instructs the model to act as a UX research expert and return structured JSON with four fields: is_dark_pattern (boolean), confidence (0–1), category_match (boolean), and reason (plain-English explanation, maximum 25 words).

The calibration guidance embedded in the system prompt establishes that confidence 0.85–1.00 denotes a textbook manipulative example, 0.60–0.84 denotes probable manipulation, 0.40–0.59 denotes ambiguous cases, and 0.00–0.39 denotes likely legitimate content. The model is instructed to be skeptical and to penalise false positives — a calibration that reflects the asymmetric costs of false alarms in a consumer-facing tool.

The normalisation function post-processes LLM verdicts: if is_dark_pattern or category_match is false, the final confidence is capped at 0.40 to ensure the detection does not reach the overlay threshold; if both are true, confidence is clamped to [0.40, 0.98]. This ensures that LLM output cannot directly override the fundamental is_dark_pattern classification from the heuristic layer.

The circuit breaker opens after three consecutive LLM call failures, pausing LLM verification for 30 seconds. During an open circuit, all ambiguous detections fall back to their heuristic confidence scores. This ensures that Groq API outages do not degrade the user experience beyond the false-positive rate increase inherent to heuristic-only operation.

4.4 Trust Scoring Algorithm

The trust score is computed by the logarithmic penalty function in lib/scorer.js. The base score is 100. For each confirmed detection above the confidence floor (0.45), a penalty is applied:

$$\text{penalty} \mathrel{+}= \text{base\_severity} \times \text{confidence} \times \left(1 / \sqrt{\text{count\_of\_same\_pattern}}\right)$$

where base_severity is {LOW: 2, MEDIUM: 5, HIGH: 9, CRITICAL: 14}. The 1/√count dampening factor models diminishing marginal harm — a site with ten instances of the same medium-severity pattern is worse than one with one instance, but not ten times worse. The final score is clamped to [0, 100] and mapped to a letter grade and verdict string:

Grade	Score Range	Severity	Verdict
A	90 – 100	Clean	No significant dark patterns detected
B	75 – 89	Mostly Fair	Minor issues flagged
C	60 – 74	Caution	Manipulation tactics present
D	40 – 59	Deceptive	Multiple dark patterns in use
F	0 – 39	Hostile	Proceed with extreme caution

Table 3: Trust Score Grade Scale and Verdict

myresearchgo Volume 2, Issue 5, May 2026, ISSN: 3107-3816 (Online)

4.5 Overlay and User Interface

The overlay engine in content/overlay.js manages all visual output. On initialisation, a single div is injected into the page as the Shadow DOM host; the Shadow DOM root contains a style element (loaded from styles/overlay.css) and the overlay container. This approach ensures complete CSS isolation from the host page.

For each confirmed detection, an absolutely-positioned highlight div is created and positioned using the getBoundingClientRect() of the affected element, with a fixed-position scroll-tracking update on window scroll events. The highlight has a 2px red border, a semi-transparent red background, and a z-index of 2147483647. On hover, a tooltip panel is displayed showing the pattern name with its icon, the severity badge, the short and long descriptions, the evidence text snippet, the LLM reasoning (if available), and the legal citation.

The popup dashboard (popup/popup.html and popup.js) communicates with the content script via chrome.runtime.sendMessage to retrieve the current detection set and trust score. It renders a per-pattern breakdown with instance counts and evidence snippets. Clicking any pattern entry triggers a scroll-and-pulse animation on the corresponding overlay element, allowing users to locate the offending UI element in context.

4.6 Configuration and Extensibility

The options page exposes four sensitivity presets — Lenient, Balanced, Strict, and Paranoid — which adjust per-scanner confidence thresholds. Individual scanners can be toggled on or off. Specific hostnames can be excluded from scanning (for development environments or trusted internal tools). The Groq API key is stored in chrome.storage.sync, which Chrome encrypts if the user has a sync passphrase configured.

Adding a new pattern category requires: (1) creating a new scanner module in content/scanners/, (2) adding a taxonomy entry in lib/taxonomy.js with name, severity, descriptions, and legal citation, and (3) registering the scanner in content/main.js. No changes to the scoring, overlay, or popup infrastructure are required.

5. EVALUATION AND RESULTS

5.1 Demo Page Evaluation

A purposely-constructed adversarial demo page (demo/demo.html) was created containing nine confirmed dark pattern instances spanning urgency, false scarcity, confirmshaming, pre-selected consent, hidden costs (simulated), sneak-into-basket, visual interference, social proof, and fake countdown timer. The page was designed to represent a realistic e-commerce checkout flow.

On loading the demo page with DPS installed and Groq API enabled, the extension detected and highlighted all nine target patterns within 800ms of DOMContentLoaded. No false positives were observed on elements not intended as dark patterns. The trust score computed was 31/100 (Grade F, Hostile UX), consistent with the density and severity of intentionally embedded patterns. All highlighted elements displayed correct tooltips with appropriate regulatory citations.

5.2 Real-World Website Observations

Informal testing across representative e-commerce and subscription platforms yielded qualitatively consistent results. High-density dark pattern sites (flash-sale pages, subscription SaaS landing pages) typically scored in the D–F range (25–55), exhibiting clusters of urgency, scarcity, and confirmshaming. News and informational sites typically scored B–A (75–95) with occasional visual interference detections on cookie consent banners. The pattern co-occurrence hypothesis from the research roadmap — that scarcity, urgency, and confirmshaming cluster together on flash-sale pages — was consistently observed informally.

5.3 LLM Impact on Precision

On pages with ambiguous urgency and scarcity language (e.g., legitimate limited-time sale events with real inventory constraints), the heuristic layer generates intermediate-confidence detections that the LLM correctly downgrades in a majority of observed cases. Phrases such as 'Sale ends Sunday' are rated 0.30–0.38 by the LLM (below the overlay threshold) while 'HURRY! Only 2 left! 87 people viewing now!' receives 0.88+ confidence. This qualitative outcome supports the hypothesis from the research roadmap that LLM verification reduces false positives on novel phrasings, though rigorous precision/recall measurement on a labelled benchmark dataset remains as future work.

5.4 Performance Characteristics

The content script's initial scan of a moderately complex e-commerce page (approximately 2,000 DOM nodes) completes in under 120ms on a mid-range laptop, measured using performance.now() instrumentation. MutationObserver callbacks for incremental DOM changes (e.g., infinite scroll additions) complete in under 20ms for typical content additions. Groq LLM calls return in 400–900ms under normal network conditions. The overlay rendering for ten simultaneous detections adds approximately 15ms. Overall, the user-perceptible latency from page load to highlighted overlays is under one second in the common case.

6. RESEARCH ROADMAP AND FUTURE WORK

The current implementation establishes a solid detection foundation. Several research directions of significant academic and regulatory value are identified for future development:

6.1 Longitudinal Per-Domain Study

The local research telemetry system stores per-host detection counts in chrome.storage.local. With user consent, this data could be aggregated across a panel of users to produce a longitudinal ranking of e-commerce sites by dark-pattern density. Comparing pre- and post-CCPA 2023 enforcement detection rates would provide direct empirical evidence of regulatory effectiveness — a study design currently absent from the literature.

6.2 Precision/Recall Benchmarking

A rigorous evaluation requires a labelled benchmark dataset of web pages with ground-truth dark pattern annotations. Construction of such a dataset — analogous to the Mathur et al. (2019) corpus but updated for

current regulatory definitions — would enable formal precision/recall reporting for both the heuristic-only and hybrid pipelines, and meaningful ablation studies on the LLM contribution.

6.3 Cross-Jurisdictional Analysis

A systematic study comparing dark pattern density on the same websites when accessed from Indian versus EU versus US IP addresses would provide the first direct empirical evidence of jurisdiction-specific site behaviour — testing whether sites comply with stricter regulatory regimes only when they detect users from those jurisdictions.

6.4 Multimodal Detection

Visual Interference detection is currently approximated through CSS property analysis. A multimodal LLM (e.g., GPT-4o or Llama Vision) could analyse actual page screenshots to detect contrast asymmetry, typography manipulation, and layout-level interference that CSS analysis cannot capture. Integration would require screenshot capture via the chrome.tabs.captureVisibleTab API in the service worker.

6.5 Pattern Evolution Tracking

Dark pattern tactics evolve in response to regulatory pressure — regulators banned cookie walls, so platforms introduced "legitimate interest" pre-ticked options. A detection system that logs pattern evolution over time, keyed to regulatory events, would produce novel data on regulatory effectiveness and the innovation dynamics of deceptive design.

7. PRIVACY AND ETHICAL CONSIDERATIONS

The development of an automated dark pattern detection system raises several ethical questions that this section addresses explicitly.

User Privacy: DPS is designed with privacy-by-design principles. No page URL, screenshot, or full page content is ever transmitted to external servers. The Groq API receives only short text snippets (maximum 400 characters of evidence plus 600 characters of context) stripped of all personally identifiable information. The extension does not record browsing history. Research telemetry is strictly local.

False Positives and Harm: Incorrect dark pattern identification could unfairly stigmatise legitimate design choices. The graduated confidence threshold and LLM verification layer are specifically designed to minimise false positives. The overlay is informational, not blocking — it does not prevent users from interacting with flagged elements, preserving user autonomy.

Adversarial Robustness: Sophisticated actors aware of the extension's detection methods could potentially design dark patterns to evade the specific heuristics. This is a fundamental limitation of any published detection system. However, the combination of multiple independent detection layers (text, structure, CSS, and semantic LLM) makes comprehensive evasion substantially more difficult than defeating any single heuristic.

Informed Consent for Research Use: Deployment of DPS for data collection in academic studies requires appropriate ethical approval and participant informed consent under applicable research ethics frameworks. The telemetry system is disabled by default and requires explicit user activation.

8. CONCLUSION

This paper presented Dark Pattern Sentinel (DPS), a research-grade Chrome browser extension implementing a novel hybrid detection pipeline for real-time identification of deceptive web UI patterns. The system combines fourteen deterministic DOM heuristic scanners with selective LLM-based semantic verification via Groq's Llama 3.3 70B, a logarithmic trust-scoring algorithm aligned with regulatory severity frameworks, and a Shadow DOM overlay providing immediate visual feedback to end users.

The design is directly motivated by the global regulatory momentum against dark patterns — India's CCPA Guidelines 2023, the EU's Digital Services Act 2022, and the FTC's Click-to-Cancel Rule 2024 — and the persistent gap between regulatory intent and practical enforcement capability. DPS addresses this gap by providing both a consumer awareness tool and a research data collection platform for the academic study of deceptive design at scale.

Preliminary evaluation on a purposely-constructed adversarial demo page demonstrated detection of nine dark pattern types within one second, with no false positives on non-targeted elements. Qualitative real-world testing confirmed score distributions consistent with expected site behaviour. The LLM verification layer demonstrated qualitatively correct disambiguation of genuinely ambiguous language cases.

Future work will pursue labelled benchmark construction for rigorous precision/recall evaluation, longitudinal site-ranking studies, cross-jurisdictional comparative analysis, and multimodal visual interference detection. The system represents a foundation for automated, scalable, evidence-based dark pattern auditing — a capability increasingly demanded by regulators, researchers, and consumers alike.

REFERENCES

[1] Mathur, A., Acar, G., Friedman, M., 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), 1–32. https://doi.org/10.1145/3359183

[2] Brignull, H. (2010+). Deceptive Design. Retrieved from https://www.deceptive.design/

[3] Gray, C. M., Kou, Y., Battles, B., Hoggatt, J., & Toombs, A. L. (2018). The Dark (Patterns) Side of UX Design. Proceedings of CHI 2018. ACM. https://doi.org/10.1145/3173574.3174108

[4] Nouwens, M., Liccardi, I., Veale, M., Karger, D., & Kagal, L. (2020). Dark Patterns after the GDPR: Scraping Consent Pop-ups and Demonstrating Their Influence. CHI 2020. https://doi.org/10.1145/3313831.3376321

[5] Government of India, Central Consumer Protection Authority. (2023). Guidelines for Prevention and Regulation of Dark Patterns 2023. Ministry of Consumer Affairs.

[6] European Union. (2022). Regulation (EU) 2022/2065 of the European Parliament and of the Council (Digital Services Act), Article 25. Official Journal of the European Union.

[7] European Data Protection Board. (2022). EDPB Guidelines 03/2022 on Dark Patterns in Social Media Platform Interfaces.

[8] Federal Trade Commission. (2024). Rule Concerning Negative Option Marketing (Click-to-Cancel Rule). 16 CFR Part 425.

[9] Federal Trade Commission. (2024). Trade Regulation Rule on the Use of Consumer Reviews and Testimonials (Fake Reviews Rule). 16 CFR Part 465.

[10] Norwegian Consumer Council. (2018). Deceived by Design: How Tech Companies Use Dark Patterns to Discourage Us from Exercising Our Rights to Privacy.

[11] Government of India. (2023). The Digital Personal Data Protection Act, 2023 (DPDP Act). Ministry of Electronics and Information Technology.

[12] Luguri, J., & Strahilevitz, L. J. (2021). Shining a Light on Dark Patterns. Journal of Legal Analysis, 13(1), 43–109.

[13] 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.

[14] European Commission. (2019). Behavioral Study on Dark Patterns in Online Interfaces. Publications Office of the EU.

L. C. Kasireddy, L. Popuri, G. Karunanithi, A. Varghese, S. Ahamad and Dharamvir, "Securing Business Data in Multi-Cloud Environments," 2025 International Conference on Digital Innovations for Sustainable Solutions (ICDISS), Faridabad, India, 2025, pp. 1-6, doi: 10.1109/ICDISS68238.2025.11320589.

L. C. Kasireddy, S. Paruchuri, C. Janakamma, A. Sarawat, K. C. Ravi and R. Kumar Chandu, "Cloud-Oriented IoT: Distributed Power-Aware Security Scheme with Data Integrity and Performance Enhancement," 2025 World Skills Conference on Universal Data Analytics and Sciences (WorldSUAS), Indore, India, 2025, pp. 1-6, doi: 10.1109/WorldSUAS66815.2025.11199185.

L. C. Kasireddy, A. Jeraldine Viji, P. K. Sholapurapu, D. Sowjanya Kolluru, D. U. Vishweshwar and P. Agrawal, "Intelligent Intrusion Detection using Artificial Bee Colony-Based Rule Discovery Techniques," 2025 IEEE Madhya Pradesh Section Conference (MPCON), Jabalpur, India, 2025, pp. 691-696, doi: 10.1109/MPCON66082.2025.11256592.

L. C. Kasireddy, S. Paruchuri, C. Janakamma, A. Sarawat, K. C. Ravi and R. Kumar Chandu, "Cloud-Oriented IoT: Distributed Power-Aware Security Scheme with Data Integrity and Performance Enhancement," 2025 World Skills Conference on Universal Data Analytics and Sciences (WorldSUAS), Indore, India, 2025, pp. 1-6, doi: 10.1109/WorldSUAS66815.2025.11199185.

J. L., L. Chandrakanth Kasireddy, R. V. Palanivel, G. Sushma, K. Bhimaavarapu and P. V. Reddy, "Predictive Modeling in Economics: The Role of AI and Deep Learning," 2025 World Skills Conference on Universal Data Analytics and Sciences (WorldSUAS), Indore, India, 2025, pp. 1-7, doi: 10.1109/WorldSUAS66815.2025.11199198.

N. Soni, L. C. Kasireddy, T. S., C. Sinhgadiya, S. Kumar and A. T. S., "A Recurrent Neural Network Framework for Effective DDoS Attack Detection in Cloud Computing," 2025 2nd International Conference on Multidisciplinary Research and Innovations in Engineering (MRIE), Gurugram, India, 2025, pp. 594-598, doi: 10.1109/MRIE66930.2025.11156616.

Jadhav, D., & Shinde, C. (2026). Sakhi: Stay safe stay fashionable. myresearchgo, 2(1), 1. https://doi.org/10.64448/myresearchgo.vol2.issue1.01.

Jadhav, A. (2026). AI-enhanced employee management system. myresearchgo, 2(1), 8. https://doi.org/10.64448/myresearchgo.vol2.issue1.02.

Rane, G., & Matteti, V. (2026). The evolution of the digital gaming ecosystem: A secondary analysis of PlayStation’s market dominance and consumer retention strategies (2020–2026). Myresearchgo, 2(3), 1. https://doi.org/10.64448/myresearchgo.vol2.issue3.01.

Ansari, N., Sharma, A., & Yadav, S. (2026). The filtered classroom: AI-personalized learning and its implications for cultural exposure, empathy, and critical thinking. Myresearchgo, 2(3), 12. https://doi.org/10.64448/myresearchgo.vol2.issue3.02.

Junghare, P., Chheniya, J., Behare, M., Kashte, P., Belekar, S., Dhoble, V., & Kumari, S. (2026). Google's Neural Memory Architecture: A Comprehensive Review of the Titans Framework. Myresearchgo, 2(4), 75. https://doi.org/10.64448/myresearchgo.vol2.issue4.12.

DECLARATION

I, Mayur Hemnath Karve, hereby declare that this research paper entitled "Dark Pattern Sentinel: A Hybrid DOM Heuristics and LLM-Based Real-Time Detection System for Deceptive Web UI Patterns" submitted to Tilak Maharashtra Vidyapeeth, Pune, is a record of original research work carried out by me under my own supervision. This work has not been submitted elsewhere for the award of any degree or diploma. All sources of information used in this research have been appropriately acknowledged and cited.