Sheep's clothing, wolfish impact: Automated detection and evaluation of problematic 'allowed' advertisements

Ritik Roongta
New York University
ritik.r@nyu.edu

Julia Jose
New York University
julia.jose@nyu.edu

Hussam Habib
New York University
hh3649@nyu.edu

Rachel Greenstadt
New York University
greenstadt@nyu.edu

Abstract

The digital advertising ecosystem sustains the free web and drives global innovation, but often at the cost of user privacy through intrusive tracking and non-compliant ads, especially harmful to under-age users. This has led to widespread adoption of privacy tools like adblockers and anti-trackers, which, while disrupting ad revenues, expose users to alternate forms of tracking and fingerprinting. To address this, many adblockers now allow "non-intrusive" ads by default. In this study, we evaluate Adblock Plus's Acceptable Ads feature and find a 13.6% increase in problematic ads compared to no adblocker use—challenging claims of improved user experience. We also find that ad exchanges on allowlists are more likely to serve problematic content, underscoring the hidden cost privacy-aware users pay when relying on such technologies.

While prior work in the domain has been limited by their practical viability, we further propose a methodology to automate the detection of problematic ads using LLMs with zero-shot prompting, achieving substantial agreement with human annotators (IAA score: 0.79). This establishes the efficacy of LLMs in problematic content detection under well-defined environments. As In-browser LLMs emerge, adversaries may exploit problematic ad content to fingerprint privacy-conscious ABP users. At the same time, these advances present new opportunities for adblockers to develop robust defenses, detect malicious exchanges, and uphold both user privacy and the sustainability of the ad-supported web.

Keywords

Adblockers, Problematic Ads, Large language models, Automation, Advertising, Tracking

1 Introduction

The digital advertising industry was valued at USD 740.3 billion in 2024 and is expected to surpass one trillion US Dollars by 2029 [22]. This revenue is a crucial source of income for tech giants who drive major technological advancements such as artificial intelligence, cloud computing, and search engines. It also supports free access to common services such as reference sites and news platforms. Content creators also rely on advertising to monetize user engagement.

Although online advertising is ubiquitous and financially beneficial, it also introduces harms that erode the browsing experience. Pervasive tracking and data collection [46, 69, 80, 94, 127, 140] raise privacy concerns, while studies report that ads can provoke feelings of creepiness, fear, boredom, and intrusion [148, 159]. Deceptive formats, intrusive placements, and dense ad clutter heighten viewer annoyance [16, 71]. To protect minors, major ad exchanges operate strict content review pipelines [19, 24, 28], and frameworks such as the GDPR and FTC guidelines restrict what can be shown in specific regions [2, 13, 35]. Finally, some ads also serve as vectors for malware, further jeopardizing user security [131, 147].

To protect themselves from these increased tracking and other concerns, many users adopt privacy-preserving tools such as adblockers and anti-trackers. Some examples include uBlock Origin [36], Adblock Plus (ABP) [17], Privacy Badger [32], Ghostery [23], or the DNS-level utility Pi-hole [31]. Datareportal estimates that nearly one billion people—about 40% of internet users—employed an adblocker in 2024 [16, 21].

Paradoxically, stronger privacy measures can worsen outcomes and raise the cost of privacy for users. Studies show that users who reject cookie banners are subsequently subjected to more aggressive tracking [57, 127], and those who decline data sharing often encounter steep paywalls or higher fees [39, 103, 121]. Adblocker users face additional pressures: these privacy-preserving tools remove ads, costing publishers roughly USD 54 billion in lost revenue during 2024 [10]. In response, 30.5% of websites deploy anti-adblock scripts that monitor user behavior [160].

In this study, we investigate a similar widely deployed privacy intervention, Acceptable Ads, which comes as a default setting of the ABP adblocker and ABP browser. With a user base of over 300 million in 2023 [11], it permits non-intrusive advertisement using curated allowlists: exchanges and publishers pay for inclusion and must meet the Acceptable Ads Standard. This helps to avoid blanket ad blocking and web revenue loss, ensuring fair treatment of users.

We focus on ad content as a proxy for measuring user experience, conducting what we believe to be the first large-scale automated analysis of low-quality ads shown to privacy-aware adblocker users. Our findings suggest these interventions impose a hidden cost on privacy-aware users, exposing them to more problematic ad content that degrades their browsing experience. Poor ad experiences often compel users to adopt aggressive adblocking, resulting in increased site breakages [92, 137] or to disable adblockers altogether, thus exposing themselves to increased tracking. Further, by automating the detection of problematic content, we demonstrate that users

Proceedings on Privacy Enhancing Technologies YYYY(X), 1–23

https://doi.org/XXXXXXX.XXXXXXX

can be fingerprinted based on their ABP configuration through ad content analysis. These findings suggest that the Acceptable Ads standard falls short of providing a safe and seamless browsing experience, effectively raising the cost of privacy by penalizing privacy-aware users.

We define **problematic** ads as any ad that is deemed unsuitable for its audience based on its content. We extend this definition to include ads that have annoying mechanisms (e.g. auto-playing video ads) or have deceptive practices (e.g. deceptive/exaggerated claims or fake buttons). Such ads undermine user experience, penalising privacy-aware users for adopting privacy-enhancing technologies.

To this end, we collected ads from two vantage points—US and Germany—each with different regional guidelines, while simulating both under-age and adult audiences. We then instrumented browsers with and without an ABP adblocker that conditionally filters content based on the Acceptable Ads allowlist. Our findings reveal a significant increase of 13.6% in the number of problematic ads shown to ABP's users on average compared to users with no adblocker installed – contradicting its claim of showing non-intrusive and non-annoying advertisements. For US citizens and under-age groups, the increase is much more pronounced at 17.6% and 21.8% respectively. Using ad exchanges as mediator, our analysis reveals that ad exchanges that do not get blocked by the allowlist seem to contribute to this increase.

Prior works highlight the gaps [48, 159] in problematic ad content detection, citing subjectivity in content and a lack of precise definitions. We address this by (i) introducing a detailed taxonomy of problematic categories, each backed by curated keyword pools, and (ii) using that taxonomy to guide large language models. When evaluated against human experts, OpenAI's GPT-4o-mini attains an inter-annotator agreement of 0.74 across all categories and 0.79 for the simpler binary distinction between problematic and non-problematic ads. While LLMs show promise for scalable ad screening, they also expose how fingerprinting can profile ABP users based on the prevalence of problematic ads.

We aim to address the following research questions in this study:

- **RQ1.** Do privacy-conscious ABP users face a higher risk of exposure to problematic advertising?
- **RQ2.** Are specific demographic groups disproportionately exposed to problematic ads, and who are the key drivers behind this?
- **RQ3.** Can LLMs automate the detection of problematic ads, and does this enable fingerprinting of ABP users?

While we recognize that fully resolving the complexities of problematic ad content detection remains challenging, our work represents a significant step toward making the process transparent and viable. Our contributions are as follows:

- We find that ABP users **increasingly encounter** problematic ads, weakening its protection and increasing the privacy trade-off.
- We **quantify the spread** of problematic ads across geographies and age groups, and identify ad exchanges that contribute the most towards it.
- We show that while LLMs can **effectively detect** problematic ads, they also expose privacy-conscious users to fingerprinting risks.

Privacy Implications. Our study presents a large-scale analysis of ad content under the Acceptable Ads program, a widely adopted privacy feature in Adblock Plus. While intended to balance user privacy and web monetization, it inadvertently exposes privacy-aware users to more problematic ads, degrading their experience. This echoes a broader trend in privacy-enhancing technologies, where users often bear hidden costs for asserting control over their data. We take a step toward mitigating these unintended consequences through LLM-assisted detection of problematic ads.

Code and Dataset Availability. We release the complete set of scripts used for ad collection and evaluation, along with our annotated ad dataset. The repository is publicly available at https://github.com/Racro/AcceptableAds_PETS.git.

2 Background

In 2019, GlobalWebIndex reported that 47% of internet users globally employed adblockers. Among these, **ABP** stands out as one of the most widely used, with over 200 million users. It primarily operates through predefined filters, such as EasyList, while also supporting custom filters. Users can optionally enable EasyPrivacy to block trackers. By default, ABP incorporates an exception list under the 'Acceptable Ads' initiative, which permits non-intrusive ads to bypass existing rules, provided they comply with the Acceptable Ads Standard (AAS). Companies generating over 10 million additional monthly ad impressions through Acceptable Ads are required to pay a licensing fee to participate in this initiative.

The Acceptable Ads Standard [15], maintained by the Acceptable Ads Committee, outlines best practices to ensure ads remain non-intrusive. It requires ads to be placed unobtrusively, clearly distinguishable from content, and appropriately sized to minimize distractions. Ad inclusion under this standard occurs either through publishers registering to display ads from specific ad exchanges or ad exchanges registering to display ads on different websites adhering to the guidelines, with compliance assumed for the cosmetic rules established by AAS. ABP charges a license fee to these advertisers and publishers to maintain the allowlist. Similarly, the Coalition for Better Ads [20], comprising major media companies like Google and Meta, proposes guidelines to improve the online advertising experience. Their recommendations, based on consumer research, identify disruptive formats like pop-ups, auto-playing videos, and large sticky ads as unacceptable. While aligned with the Acceptable Ads Standard, the CBA operates as a voluntary body without enforcement mechanisms.

Ghostery [23] and Privacy Badger [32] are privacy-preserving browser extensions [135] that block trackers while explicitly permitting non-tracking advertisements. Although sound in principle, this strategy is overly conservative because most ad exchanges employ a similar family of domains for both ad delivery and tracking. Consequently, nearly all advertisements are blocked, making empirical analysis of ads permitted by these extensions exceedingly difficult—our crawlers encounter virtually none. In contrast, Adblock Plus leverages its allowlist to exempt selected exchanges, thereby permitting a limited subset of ads to reach users.

Cho et al. [71] identified three key indicators of ad avoidance in their user study: Perceived goal impediment, Perceived ad clutter, and Prior negative experiences, each with three sub-indicators. Goal impediment arises when ads interfere with users' objectives, with sub-indicators like Search hindrance, Disruption, and Distraction. Ad clutter refers to excessive advertising, reflected in Excessiveness, Exclusiveness, and Irritation. Negative experiences influence attitudes towards ads through Dissatisfaction, Perceived lack of utility, and Perceived lack of incentive. While aesthetic standards like AAS and CBA address some issues such as hindrance and excessiveness, they fall short in reducing overall user annoyance.

Annoyance represents only one facet of advertising's broader impact, as numerous studies have documented the harmful effects of ads on users [156]. These include highlighting problematic ads on children's websites [118, 120, 149], harmful Facebook advertising [48, 49], deceptive news articles [154], parent–child conflicts [62], and the psychological implications of weight loss and alcohol ads [65, 82, 130]. Although the subjective nature of ad content often complicates its evaluation, certain categories, such as profanity, unsubstantiated health claims, and crypto scam ads, are universally considered problematic and have been extensively studied [47, 116, 139, 143, 152]. Additionally, subjective domains like boring, creepy, and scary ads have been explored through various user surveys [148, 159]. These studies highlight how advertisements often exploit cognitive biases, influencing non-rational consumer decisions through emotional and misleading tactics. Harmful ads, like those promoting stereotypes, health claims, or scams, can undermine confidence and increase anxiety [102]. Reducing exposure to such ads is crucial for improving the browsing experience.

2.1 Related Work

Previous efforts [48, 159] derived taxonomies of problematic display ads from user surveys: Zeng et al. [159] examined political advertising during the 2020 U.S. elections, while Ali et al. [48] studied ads within the Facebook platform. We advance this line of work by devising a finer-grained taxonomy that merges user perceptions with ad-exchange policies and regulatory standards. This framework allows us to establish a tighter lower bound on the prevalence of problematic ads and to automate their detection. We also broaden ad collection to cover multiple regions and age-specific scenarios.

This work builds on a body of research that investigates the unintended consequences of privacy-enhancing technologies. Smith et al. [145], analyzed how filterlists used by adblockers can disrupt site functionality and cause web breakages, thereby degrading user experience. Similarly, Demir et al. [84] observed that extensions meant to assist with cookie consent interactions can paradoxically increase tracking activity. Other studies underscore the fragility of regulatory protections: Papadogiannakis et al. [127] reveals how websites circumvent GDPR consent to continue tracking, and Liu et al. [115] shows how CMPs are ineffective in protecting user privacy and rather worsen it by reducing users' anonymity sets, inadvertently enabling fingerprinting. Collectively, these studies highlight a recurring theme: that users adopting privacy measures often bear unintended costs, and current interventions are insufficient to guarantee a secure and private web experience.

Over the past decade, researchers have mapped the rapidly expanding landscape of web-tracking technologies. Large-scale measurement studies reveal the breadth of third-party trackers that follow users across sites and exchange that data with one another [46, 69, 80, 94, 127, 140]. Two complementary strategies dominate. Stateful tracking stores explicit identifiers in the browser—most visibly cookies, but also HTML5 localStorage, ETag cache entries, and similar client-side repositories [53, 70, 119]. Stateless tracking, in contrast, avoids local storage by constructing a unique browser "fingerprint" from configuration cues and hardware idiosyncrasies [67, 88, 110, 123]. Studies show that fonts, clock-skew patterns, GPU pipelines, audio subsystems, installed writing systems, and even active extensions can all serve as distinguishing features [113, 114, 144]. Parallel defensive research has developed tools to detect or disrupt both cookie-based trackers and sophisticated fingerprinting schemes [104, 106, 141, 144].

Advertising has also been examined as a vehicle for user tracking. Several studies quantify these practices by modeling tracker relationships with graph-based frameworks [106, 141]. Marotta et al. [117] estimate the revenue that publisher sites derive from tracking and behavioral targeting, while other work investigates how ad-network practices erode consumer privacy [50, 87].

Research attention has also turned to ad fraud. Advertisers employ cloaking tactics to slip malicious or policy-violating ads past exchange filters. Papadogiannakis et al. [128] document "ad laundering," in which bad actors disguise illicit content within seemingly legitimate sites. Misinformation outlets likewise mix their inventory with unrelated domains to bypass brand-safety checks and subvert transparency standards [126, 150, 151].

Researchers have also scrutinized ad-transparency standards and how users perceive them. Kim et al. [108] examined user attitudes toward advertisers after viewing targeting details provided by transparency dashboards. Zeng et al. [158] analyzed how demographic attributes influence ad-targeting choices and bid values. Investigations into political-ad transparency on major platforms such as Meta and Google [85, 90, 107] reveal how these exchanges address reports of malicious or deceptive campaigns and shed light on their enforcement practices.

Edwards et al. [91] examined how the perceived intrusiveness of pop-up ads leads to irritation and ad avoidance, identifying factors such as ad congruence with user tasks and cognitive intensity at the time of interruption. Cramer [79] found that even high-quality native ads, when closely aligned with surrounding content, can negatively impact perceived site credibility and quality, emphasizing the importance of distinguishing ads from editorial content. Campbell [66] discussed the challenges native advertising poses to consumer protection, suggesting that its seamless integration into content may require novel regulatory approaches to prevent deception. Braun and Eklund [61] explored how programmatic advertising platforms can inadvertently fund fake news publishers, highlighting the need for greater transparency and responsibility within the ad tech industry.

Prior studies have also looked into the use of taxonomies to classify content online. Nickerson et al. [125] demonstrated methods for developing taxonomies from existing literature using deductive, inductive, and intuitive approaches. Morrow et al. [122] investigated content labeling practices in the context of social media moderation, while Singhal et al. [142] conducted a systematic study (SoK) of labeling practices and their enforcement. Banko et al. [54] proposed a unified taxonomy for harmful content.

Table 1: Table summarizing the stakeholders in the ad ecosystem, their contributions to the taxonomy, and the sources used to capture their perspectives.

Stakeholder	Contributions	Sources
Ad Exchanges	Provide insights into Age-based and Geographical Regulations, Restricted topics like Health, Financial commodities, etc. Also highlight policies around fraud and scam content and political content	Meta [28], Google [24], OpenX [29], Taboola [34], OutBrain [30], Pubmatic [33], Bing [19], Amazon [18]
Regulators	Highlight different publisher and advertiser requirements around native advertising, weight loss ads, tracking, etc.	GDPR [13], Federal Trade Commission (FTC) [2, 35, 96]
Ad Consortiums	Establishes aesthetic standards for ads around autoplaying ads, ad dimensions etc. Lays down bottom-line guidelines for disruptive user experience.	Acceptable Ads Standard (AAS) [40–45], Coalition for Better Ads (CBA) [73–76], Interactive Advertising Bureau (IAB) [26]
Industrial Studies & Reports	Yearly reports about user perceptions of ads, popularity of adblockers, and popular reasons for adblocking.	Hubspot [9], CAPV [68], Confiant [77, 78], Op-ed pieces [1, 3, 4, 6, 7]
Academic Studies	Academic journals in Psychology, Privacy, and Security highlighting the intrusiveness of ad content and its effect on user perceptions. Also covers different ways of dark patterns prevalent in online media	Ali [48], Zeng [159], Colin [100, 101], Blase [148], Gomez [99], Bosch [58], Weidelmark [153], Cho [71], Brajnik [60], Gak [97], Burke [63], Cramer [79], Braun [61], Rohrer [136] et al.

3 Why Evaluate Acceptable Ads?

Acceptable Ads is enabled by default in both Adblock Plus and the ABP browser, which together exceed 300 million users. The program currently includes more than 1,000 ad exchanges and display platforms, along with over 48,000 participating publishers. Despite ABP's widespread adoption, the Acceptable Ads Standard (AAS) has faced significant criticism from users. In a study analyzing ABP chrome web store reviews [135], 139 out of 150 (92.7%) reviews expressed dissatisfaction with Acceptable Ads. This dissatisfaction stemmed not only from a general aversion to advertising but also from dissatisfaction with the content of the ads, even when they appeared less intrusive. User feedback highlight this sentiment, with some stating, "Customizable, so I can allow only non-tracking ads, and/or acceptable levels of ads, or none, or only on some websites, as desired". Similarly, another user remarked, "i have been suffering with some websites that have really awful ads but i want to support the other websites that i like, so it would be very appreciated if they had a blacklist feature with acceptable ads on some websites".

While the AAS aims to improve user experience through aesthetic guidelines, its effect on ad content quality remains uncertain. Although aesthetics and content quality are orthogonal, with no inherent correlation, improved aesthetics could potentially attract higher-quality content. However, Edelman's adverse selection theory [89] suggests a counterintuitive outcome: enforcing aesthetics-based allowlists might deter higher-quality ad networks due to increased compliance costs or restrictions, leaving lower-quality networks to dominate. This could degrade overall ad content despite improvements in visual appeal. As a result, this approach risks misleading users, as it permits problematic ad content despite claims of promoting better or acceptable ads. It also makes them vulnerable, as an adversary could now target ABP users differentially or monitor problematic content in ads to fingerprint them. These concerns prompt a critical challenge to the privacy research community: to rigorously evaluate the effectiveness and unintended consequences of the tools they design to safeguard user privacy.

4 Taxonomy Preparation

In this step, we draft a fine-grained taxonomy to define what constitutes problematic ads. The challenge lies in establishing a definition that aligns with the interests of all stakeholders in the ad ecosystem, ensuring wider adoption of our results. Additionally, considerations for using this taxonomy to automate the detection must remain a priority throughout the process. Standards for problematic ads are diverse and subjective. Taboola and Outbrain permit celebrity endorsements if they avoid health-related content (e.g., "See how Tom Hanks Is Recovering from Coronavirus"). In contrast, other ad exchanges focus solely on the advertised product, with no specific policies on celebrity mentions.

We adopted an inductive approach with deductive fine-tuning to construct our taxonomy, ensuring robustness and exhaustiveness. The inductive process followed a grounded-theory-inspired approach [56], involving a comprehensive review of literature on ad content policies and prevalent bad practices in online display advertising. We identified broader categories based on shared characteristics by synthesizing and reorganizing common themes across different sources. While some categories (e.g., clickbait or political ads) may hold greater significance than others (e.g., inappropriate content), all categories found were represented in the reviewed literature and platform policies.

We examined ad policies from major platforms such as Google, Meta, Bing, Amazon, Taboola, and others [18, 19, 24, 28, 30, 34], FTC native advertising guidelines [2, 35], GDPR privacy regulations [13], and independent reports from Hubspot [9] and Confiant [77, 78].

We also reviewed op-eds and independent investigations by publications like the New York Times, Wired, and The Verge [1, 3, 6]. Furthermore, we performed an academic literature review (refer to Table 1) to capture user perceptions and taxonomies developed in prior research. Key insights and keywords from each source were summarized into a document. This summary underwent several language processing tasks: text tokenization using n-grams, stopword removal, TF-IDF-based keyword extraction, and named entity recognition (NER) to identify significant entities. Using TF-IDF vectorization, we converted keywords into vectors and clustered them based on cosine similarity. To visualize and refine these clusters, we applied Principal Component Analysis (PCA), which reduced noise and revealed clear boundaries. Keywords with high cosine similarity (above 0.85) were merged into single representative terms. A threshold of 0.85 marks the "elbow" where intra-cluster precision plateaus—above it we consistently merge true synonyms (e.g., data privacy/privacy protection), while lower cut-offs begin collapsing conceptually distinct terms. Refer to Table 6 in the Appendix for more details on the term and cluster analysis. A manual review by two co-authors ensured contextual accuracy and helped refine broader categories by removing redundant terms.

This step provided us with an initial taxonomy of the major topics in the classification of ad content. In order to refine the taxonomy further and check if new themes originated, we performed a deductive step where two researchers labeled an initial pool of 200 ads, randomly selected from our pool of collected ads (Section 5.1.1), and tried to classify them as problematic/non-problematic. If problematic, we try to classify them into one or more of the categories using the keywords identified in the previous step. If not possible, we try to add keywords. This process was repeated until no new themes emerged. The two researchers have an IAA Krippendorff alpha [109] score of 0.72, which was achieved by iterative annotation of ads and resolving conflicts after each phase. The taxonomy reached a plateau after five iterations. Once the initial taxonomy was complete, we annotated the ads using a pool of ad experts trained on this taxonomy.

4.1 Taxonomy

Our taxonomy is designed to serve three guiding principles: 1. It aims to address user concerns while also **accommodating** the interests of advertisers and publishers. 2. The taxonomy supports **multi-label** classification by drafting mutually exclusive categories, allowing each ad to fall into multiple categories when applicable. For instance, an ad could simultaneously belong to the categories of "User Experience Disruption" (e.g., an SUV promotion with no advertiser information) and "Deceptive Claims and Exaggerated Benefits" (e.g., claiming the same car is priced under $1,000). 3. It maximizes the **analytical ability** of the annotators as well as LLMs by incorporating keywords for each category. This facilitates clear, objective labeling and aligns with our goal of using LLMs for automated ad classification.

The taxonomy is as follows:

1. **Regulations** — Ads targeting or featuring content deemed inappropriate for younger audiences or sensitive groups based on geographic or product-specific restrictions. This includes subcategories like *Age-Based* and *Geographical.* Examples include ads promoting alcohol, gambling, or cosmetic surgery on platforms accessible to minors, as well as prescription drug advertisements in countries where such ads are prohibited. Due to the vast and varying nature of geographical restrictions, this aspect is beyond the scope of our study, and we focus solely on age-based regulations. For instance, Figure 1, images 1a and 1b, depict regulatory violations involving cannabis and gambling advertisements displayed without appropriate age disclosures.

2. **Inappropriate or Offensive Content** — Ads containing language, visuals, or themes that are offensive, graphic, or disrespectful toward specific individuals or communities. Such ads may use racial stereotypes or sexually explicit content to attract attention. An example of this is shown in Figure 1, image 2, which features inappropriate content that could be considered sexually explicit.

3. **Deceptive claims and Exaggerated Benefits** — Ads making unverified or exaggerated claims about a product or service to mislead consumers. This category is further divided into subcategories: *Health Claims* (e.g., promoting "miracle cures"), *Financial Claims* (e.g., "get rich quick" schemes), *Environmental and Ethical Claims* (e.g., misleading environmental benefits), and *Other Impossible Claims* (e.g., "best in the world"). Figure 1, images 3a and 3b, illustrate deceptive claims about health benefits and hair loss treatments, both of which lack appropriate disclosures.

4. **Dark Patterns and Manipulative Design** — Ads employing deceptive design techniques to manipulate user behavior, such as tricking users into clicking, subscribing, or sharing information unintentionally. Examples include ads featuring fake "X" buttons or countdown timers to create artificial urgency. As illustrated in Figure 1, images 4a and 4b represent such dark patterns by hiding essential information and displaying fake buttons, respectively.

5. **User Experience Disruption** — Ads that degrade the browsing experience by being overly intrusive, annoying, or difficult to navigate. Examples include ads that autoplay sound or video, disrupt user activity, or lack clear advertiser information. Figure 1, image 5a, demonstrates an intrusive ad claiming to know the user's location, while image 5b features an ad with no clear advertiser attribution.

6. **Fraud and Scam Content** — Ads promoting fraudulent schemes or products designed to exploit users financially or otherwise. Examples involve fake investment opportunities with promises of high returns but no financial backing or evidence. They focus on making genuine-looking claims specifically intended to bait customers into clicking (click fraud) or advertising counterfeit products (scam content), only to deceive/scam them in the process. Past studies [51, 64, 83, 111, 112, 134, 138, 146] have measured scam and fraud in different contexts but failed to develop an automated setup for its detection, since this requires a thorough knowledge of the past advertiser behavior as well as advanced network analysis (see Section 9). Since our study focuses on automating problematic ad detection using ad creatives, this category is currently beyond our scope.

7. **Political and Socially Sensitive Topics** — Ads that relate to political, social, or controversial issues, particularly during sensitive times like elections. Such ads can polarize public opinion or manipulate perceptions by using misinformation. Figure 1, images 6a and 6b, depict political propaganda designed to create sensationalism without substantial evidence.

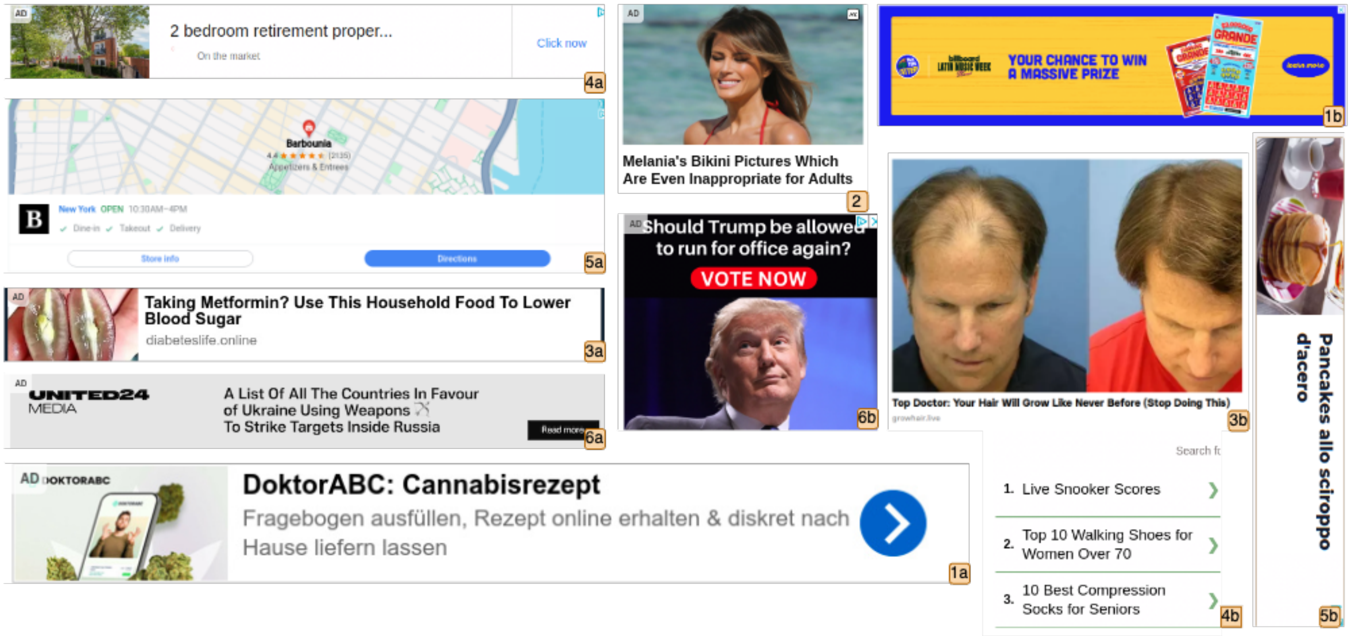
AD
2 bedroom retirement proper...
On the market
Click now
4a
AD
Melania's Bikini Pictures Which Are Even Inappropriate for Adults
2
YOUR CHANCE TO WIN A MASSIVE PRIZE
1b
Barbounia
Appetizers & Entrees
B
New York OPEN 10:30AM–4PM
Dine-in Takeout Delivery
Store info
Directions
5a
AD
Should Trump be allowed to run for office again?
VOTE NOW
6b
Top Doctor: Your Hair Will Grow Like Never Before (Stop Doing This)
3b
Pancakes allo sciroppo d'acero
5b
AD
Taking Metformin? Use This Household Food To Lower Blood Sugar
diabeteslife.online
3a
AD
UNITED24 MEDIA
A List Of All The Countries In Favour of Ukraine Using Weapons To Strike Targets Inside Russia
Read more
6a
AD DOKTORABC
DoktorABC: Cannabisrezept
Fragebogen ausfüllen, Rezept online erhalten & diskret nach Hause liefern lassen
1a
Search fo
1. Live Snooker Scores
2. Top 10 Walking Shoes for Women Over 70
3. 10 Best Compression Socks for Seniors
4b

Figure 1: Different problematic ads identified during the crawl, categorized into six main groups for labeling ad content. Categories include: 1a and 1b (Regulations), 2 (Inappropriate or Offensive Content), 3a and 3b (Deceptive Claims and Exaggerated Benefits), 4a and 4b (Dark Patterns and Manipulative Design), 5a and 5b (User Experience Disruption), and 6a and 6b (Political and Socially Sensitive Topics).

This taxonomy is rooted in the published policies and priorities of advertisers, ad-exchanges, and users. Rather than coining new labels, we unify the language already found in industry guidelines, regulatory documents, and previous user studies, producing a single, cohesive reference that appears in Table 7 (Appendix). This consolidation removes the fragmentation that complicates comparative studies of harmful advertising and supplies a practical foundation for automated analysis. By building on established terminology, the framework keeps well-defined categories—Deceptive, Political, and Manipulative ads highlighted by Zeng et al. and Ali et al.[48, 159]—while excluding vague tags such as boring or irrelevant that obstruct consistent detection. It also introduces classes emphasized in platform rules (Regulatory Violations, Fraud/Scam) and in user-centric studies (User Experience Disruption, Inappropriate Content)[71, 91]. To make these otherwise subjective notions measurable at scale, every category is anchored to an exhaustive, transparent keyword list derived from definitions provided in the literature. No prior work offers a keyword-driven framework of comparable breadth, positioning our taxonomy as a substantive methodological advancement.

While condensing diverse topics into a seven-category taxonomy may risk overlooking certain domains, we mitigate this by iteratively annotating ads and refining the taxonomy until no new categories or keywords emerge. This taxonomy reflects the interest of the industry, advertising bureau, policymakers, and users, accommodating the interests of all stakeholders in the ad ecosystem.

5 Problematic Ad Detection

Having developed the taxonomy, our next task is to collect ads and quantify the proportion of problematic ads users encounter. Ads are collected across four categories: *Unauthenticated_US* (crawls without logging in from a vantage point in the US), *Unauthenticated_Germany* (crawls without logging in from a vantage point in Germany), *Authenticated_US-over-18* (using a Google account with a declared user age over 18 years), and *Authenticated_US-under-18* (using a Google account with a declared user age under 18 years). These categories were chosen to capture two primary factors influencing ad content: **geography** and **age**. Germany was selected due to its strict GDPR regulations, where higher compliance from ad companies is anticipated. The age threshold of 18 years was chosen because most ad exchange policies enforce stringent restrictions for under-age audiences.

Crawling methodology: We prioritize geography and age for two reasons. First, advertising policies and regulations, forming the basis of our taxonomy, vary significantly by region and age group. Second, regional filterlists shape distinct tracker landscapes influencing the ads users receive [59]. A longitudinal study tracking shifts in ad targeting and problematic content would be valuable, especially for real users with curated profiles during events like elections. However, such a study would demand significant user participation and expert annotations, placing it beyond this work's scope.

Additionally, ads from all four scenarios were collected under two categories: without ABP (**Control group**) and with ABP (**AccAds**

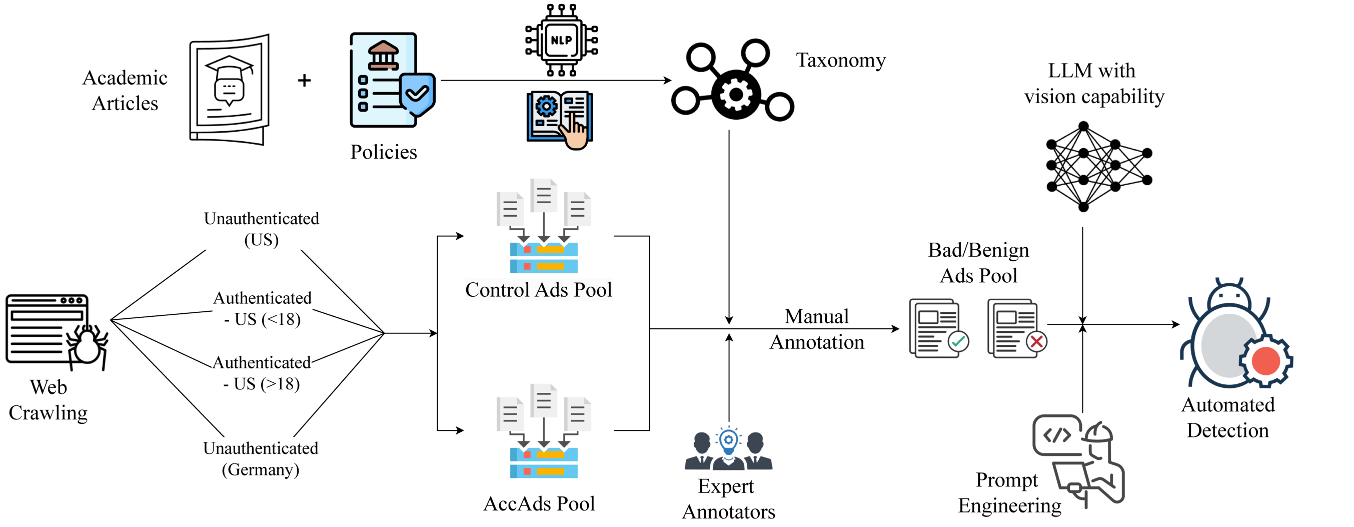

Figure 2: A high-level architecture illustrating the ad classification pipeline. In one half, taxonomy preparation is depicted, leveraging academic articles and exchange policies. In parallel, web crawling is used to collect ads from various scenarios, including unauthenticated users (US and Germany) and authenticated users (under-18 and over-18). Once the ad pool is established, expert annotators provide manual annotations, which are then used to train an LLM model to automate the annotation process.

group). ABP includes the Acceptable Ads allowlist by default, enabling us to enforce cosmetic standards and allowlisted ad exchanges, and analyze the content of ads within this framework. This approach represents a significant contribution of our work, as it examines the impact of the Acceptable Ads program on ad quality. Figure 2 illustrates our methodology. The taxonomy is developed in parallel to a web crawler designed to operate in four distinct scenarios, creating Control and AccAds ad pools. Expert annotators labeled these pools to establish ground truth, and we utilized prompt engineering with a state-of-the-art LLM to evaluate its effectiveness in automated detection.

5.1 Ad Collection

To collect ads, we utilized the DuckDuckGo tracker-radar-collector[1] framework, built on the Puppeteer [12] web crawler. This framework was enhanced with modifications from Moti et al.'s [118] work, including the use of ad detector code based on EasyList[2] to identify ads. Additionally, we incorporated insights from AdFisher [81] to improve the crawling framework by synchronizing multiple crawlers, thereby minimizing temporal-based differences in the collected ad dataset. In each of the four aforementioned scenarios, two separate crawlers were deployed: one for the *Control* group (without ABP) and another for the *AccAds* group (with ABP enabled, leveraging its Acceptable Ads allowlist). Ad screenshots, along with network packet information, were collected from all eight crawls and examined distinctly. The entire crawling process was conducted in October 2024.

Acceptable Ads allowlists are integrated into the ABP adblocker, the ABP browser, and several affiliated platforms. As our Puppeteer based crawlers show best performance with Chrome, we conduct ad collection using the ABP extension on that browser. However, the impact of Acceptable Ads is expected to be consistent across platforms, given that the allowlists and user demographics remain unchanged.

5.1.1 *Website Pools.*

To ensure a diverse spectrum of websites and capture a substantial amount of acceptable ads, we curated a list of 1,500 websites. Of these, 750 were selected from ABP's exception rules list (or allowlist), representing publishers that participate in the Acceptable Ads Program. The remaining 750 sites were randomly chosen from the Tranco 1M top lists, with half originating from the top 10k range and the other half from the 99k-100k range. Duplicates were replaced with randomly selected new websites to maintain diversity. This approach offers two key benefits. First, it enables us to analyze the behavior of acceptable ads across both higher- and lower-ranked websites. Second, it captures two facets of the acceptable ads allowlisting: the first 750 websites represent allowlisted publishers showing ads from any exchange, while the latter 750 reflect allowlisted ad exchanges displaying ads on any publisher.

Additionally, following the methodology used by Roongta et al. [135], for websites with over 10 inner pages, we included three additional inner pages in the pool. This was done because inner pages typically host a greater number of ads compared to landing pages. The final website pool consisted of 4,710 web pages, representing the general web, rather than focusing on misinformation or lower-ranked sites specifically to capture problematic ads. Prior studies [52] indicate that untrustworthy news websites account for only 2.3% of internet traffic, with this figure dropping to 1.4% for actively publishing sites. These statistics suggest that the general web, which attracts the majority of user traffic, provides a more realistic context for analyzing problematic ads.

[1] https://github.com/duckduckgo/tracker-radar-collector
[2] https://easylist.to/easylist/easylist.txt

The website corpus exhibits substantial topical heterogeneity. Using an open-source topic classifier[3], we find that News outlets constitute 20.8% of the sample, whereas Arts & Entertainment and Sports account for 15.7% and 16%, respectively. At a higher level of aggregation, the pool spans nine thematic clusters: Lifestyle & Wellness (11.3%), Arts & Culture (20.9%), Sports & Recreation (23.8%), Current Affairs (22.6%), Digital Engagement (5.7%), Human Connections (4.4%), Travel & Mobility (4.2%), Finance & Business (4%), and a residual category covering Jobs & Education and other topics (3.1%). This distribution yields a sufficiently large and diverse advertisement set, enabling the random selection of approximately 1,000 ads for manual analysis; consequently, no additional websites were incorporated for the present study. Future research that pursues category-specific analyses may expand the corpus as necessary.

5.1.2 Unauthenticated Crawls. For the unauthenticated crawls (*UnAuthenticated_US* and *UnAuthenticated_Germany*), we simultaneously launched two Docker containers, each running a Puppeteer-controlled Chrome browser configured with/without ABP adblocker. Browser state was purged after every crawl. Ad screenshots were collected from the targeted websites using a fresh Chrome profile for each site. These crawls were conducted on an AMD EPYC 7542 128-core machine from vantage points in the USA and Germany.

5.1.3 Authenticated Crawls. The authenticated crawls required a more nuanced approach, leveraging Google account logins as signals for user age to ad exchanges. A manual login step was introduced to generate authenticated profiles, saving the authentication cookie within the browser profile. Prior research [55] has used authenticated profiles for ad collection, and Google support documentation [14] confirms that login information is used for demographic-based ad targeting. To mitigate confounding variables, fresh virtual machines (VMs) and newly created accounts were used to conduct crawls from the same geolocation. Additionally, since many websites in the dataset utilize Google Analytics, demographic data from Google is often shared with non-Google ad exchanges, further influencing ad targeting.

Four Gmail accounts were created: two with a user age of 25 and two with a user age of 16. Within each age group, one account was used for the Control setup, while the other was used for the AccAds setup. These accounts were configured with '*Personalisation off*' and '*Do not store history*' settings to minimize the influence of earlier crawled sites on subsequent ones. Since authenticated crawls cannot be automated, one fresh profile every 50 crawls per user credential is used to prevent staleness. The influence of these random personas is minimized by shuffling and sampling ads randomly.

Given the additional logging-in step, the runtime for authenticated crawls increased. As a result, these experiments were limited to 1,500 webpages of unique websites from the original pool, all accessed from a US vantage point. For these crawls, we used two AMD Ryzen 9 5950X 16-core machines with different IP addresses, running the crawlers for each age group simultaneously. Each website's ads were collected after 180 seconds, following which the browsers were closed, and the next crawls were initiated.

5.1.4 Data cleaning. Across all crawls, we collected over 40,000 ads, but the dynamic nature of ads often resulted in blank screenshots or misaligned content. To clean the dataset, we used Meta's Faiss library [5] to deduplicate ads within each category by generating perceptual hashes for images and grouping those with close to 0 distance. The Google Vision API[4] was then employed to detect and filter out blank ads with no text, removing approximately 8,000 entries. Misaligned ads with unclear content were manually filtered during the annotation phase. This process ensured a high-quality dataset of 18,000 ads, free from duplicates, blank images, and unusable content.

5.1.5 Ethical Considerations. We acknowledge that the ads we scraped could have been shown to real users, potentially resulting in minor revenue losses to advertisers. However, given the vast size of the industry (valued at USD 740.3 billion) and the significance of this study in identifying problematic ads and advancing ad content moderation, the impact is negligible. To further minimize potential losses, we adopted a cautious crawling methodology, limiting our scope to 38,000 webpages across geographies and age groups, with only three inner pages per website included in the ad pool. Also, the expert annotators deployed were part of the same organization and often collaborate with us on different privacy studies.

5.2 Ad Annotation

5.2.1 Recruitment and Annotator Agreement. We recruited a pool of seven student researchers with prior experience in working with ads to assist with the annotation task. Annotators were familiarized with the taxonomy and trained on the definitions and nuances of its terms. To evaluate inter-annotator agreement (IAA), we used Krippendorff's Alpha [109] with Jaccard distance [27], a method well-suited for multi-label classifications.

During onboarding, all annotators were asked to annotate a pool of 50 ads, with any discrepancies or misclassifications reviewed and resolved collaboratively. Once the annotator pool achieved an IAA score above 0.7, they were divided into groups of four to ensure each ad image received four independent annotations. Given that our ad pool included German ads, the annotator pool included both native German and English-speaking experts to account for language-specific nuances and ensure consistent labeling across the dataset.

5.2.2 Manual Annotation. We randomly selected 600 ads (150 ads from each of the four categories) within each of the Control and Acceptable Ads groups for manual annotation. Each ad was independently labeled by four annotators using Label-studio[5]. Annotators were instructed to scrutinize obscure or lesser-known advertisers to evaluate the legitimacy of their claims. When definitions overlapped, for example, between User Experience Disruption and Dark Patterns, annotators were asked to follow the keyword list strictly. Requiring agreement from at least two experts per ad yielded a reliable ground-truth dataset.

The inter-annotator agreement (IAA) score among annotators was calculated as 0.84, indicating strong agreement. For binary

[3] https://github.com/yohhaan/topics_classifier

[4] https://cloud.google.com/vision

[5] https://labelstud.io/guide

Table 2: Krippendorf Alpha agreement scores for each human annotated problematic label. Asterisk (*) represents substantial agreement (>0.61), and meaningful inference can be drawn.

Label	Krippendorff Alpha
Political and Socially Sensitive Topics	0.91*
Deceptive Claims and Exaggerated Benefits	0.75*
Regulations	0.91*
Inappropriate or Offensive Content	0.71*
Dark Patterns and Manipulative Design	0.81*
User Experience Disruption	0.85*
Binary Classification	0.89*
Overall Agreement	0.84*

classification—determining whether an ad is problematic or non-problematic, regardless of specific labels—the IAA score increased to 0.89. Additionally, we calculated IAA scores for agreement within problematic categories, assessing the consistency of annotations for each specific label as well as the overall label pool. As shown in Table 2, all labels achieved an IAA score above 0.7, highlighting strong consistency and agreement across all categories.

6 LLM-assisted Automation Analysis

Identifying problematic ad content involves subjective judgment, as definitions are often nuanced. This highlights the challenges of ad moderation, further complicated by the absence of scalable automated solutions, leading to substantial leakage of problematic ads. We address this by assigning keywords to each problematic category, with strong inter annotator agreement supporting their utility. We leverage modern Large Language Models (LLMs) via prompt engineering to automate detection.

Unlike simple pattern recognition, detecting problematic content requires context aware reasoning, making LLMs better suited than image classifiers. LLMs adapt to evolving norms through text instructions, combine visual and textual reasoning to capture subtle context, and provide human readable explanations. We compare state of the art LLMs to CLIP [133], a visual language model, to demonstrate these advantages. Recent efforts [98, 132] also explore LLMs for content moderation; for instance, Claire et al.[155] show GPT 4 turbo outperforming crowdworkers in classifying harmful videos. Sekharan et al.[124] fine tuned GPT to improve clickbait detection in ad titles, underscoring LLMs' superior reasoning.

We use OpenAI's multimodal GPT 4o-mini [25] for classification. Using expert labeled ground truth, we first conduct binary classification (problematic or non problematic) followed by multi label category identification to identify the specific categories within the taxonomy. Iterative prompt refinement guided by model explanations produced an optimized prompt (Figure 6 in Appendix). OCR data from the Google Vision API was also supplied to assist the model. To ensure consistency, we set temperature to 0 and instructed the model to explain decisions strictly using the taxonomy, removing the need for multiple runs.

For evaluation, we computed Krippendorff's Alpha using Jaccard distance. GPT 4o-mini achieved an IAA of 0.74 (Table 3), a substantial increase from 0.39 without keyword information. Binary classification alone reached 0.79, showing strong agreement across all categories and validating our taxonomy's effectiveness.

We use three different scenarios to argue about the efficacy of our prompt and the taxonomy. First, we use CLIP model as a baseline, to show that traditional CV and NLP models with limited contextual understanding are inefficient for this task. We computed cosine similarity scores between ad (image) and category (text) embeddings corresponding to each of our ad categories. We applied the sigmoid function to these scores and performed threshold calibration on a validation set to determine optimal thresholds per category. We find GPT 4o-mini outperforms CLIP on IAA and classification metrics (Table 3, 10 in Appendix).

Second, we assess the role of taxonomy keywords by prompting GPT 4o-mini with and without them. The keyword based prompt yields substantially higher agreement with human labels (Table 3). Third, we compare GPT 4o mini with GPT 4o. Using identical prompts, both show similar agreement, with mini slightly outperforming. This suggests model size has limited impact in zero or few shot settings, and GPT 4o mini suffices for automated ad classification.

Table 3: Krippendorff's Alpha agreement scores between human and LLM labels for each problematic category in manually annotated ad pool. Asterisk (*) represents substantial agreement (>0.61) and meaningful inference can be drawn.

Label	CLIP	GPT-4o-mini		GPT-4o
		No Keywords	Keywords	
Political and Socially Sensitive Topics	0.09	0.61*	**0.78***	**0.78***
Deceptive Claims and Exaggerated Benefits	-0.01	0.51	0.65*	**0.69***
Regulations	0.02	0.45	**0.79***	0.72*
Inappropriate or Offensive Content	0.14	0.39	0.63*	**0.70***
Dark Patterns and Manipulative Design	0.12	0.54	**0.72***	0.70*
User Experience Disruption	-0.03	0.17	**0.80***	0.68*
Binary Classification	0.17	0.44	**0.79***	0.75*
Overall Agreement	0.09	0.39	**0.74***	0.69*

The GPT-4o-mini model achieved a precision of 0.88, indicating it accurately identifies relevant ads with relatively few false positives, and a recall of 0.84, suggesting it misses some true positives. The resulting F1 score of 0.86 reflects a balance between these metrics, highlighting fair performance overall (see Table 9 in Appendix). These results suggest that the LLM is capable of providing reasonable annotations. Krippendorff's alpha score of 0.74 corroborates these findings, demonstrating substantial agreement with the ground truth. This also establishes the efficacy of our keyword-based taxonomy that helps us to achieve such significant agreement between human and LLM labels.

Misclassifications stem from several factors. LLMs often over flag ads as inappropriate, for example open bars as alcohol related, fashion ads with women as sexually suggestive, or horror movie posters as offensive. These patterns suggest a lack of common sense reasoning, where human annotators outperform LLMs. For instance, a 4.5% CD interest rate ad from American Express would be seen as normal by a human but flagged as deceptive by the model. LLMs also face challenges in multi-label classification, sometimes identifying one correct category but missing others. To address these shortcomings, integrating web search for contextual grounding and fine-tuning with high-quality datasets are promising solutions.

Takeaway 1: GPT-4o-mini, when prompted with keyword based taxonomy, is effective at identifying problematic advertisements. Traditional CV/NLP approaches are ineffective, and superior models (GPT-4o) achieve similar performance for the task of zero shot ad classification. With the advent of In-browser LLMs, this finding increases the risk of profiling ABP users by monitoring ad content.

7 Evaluation

7.1 Prevalence of Problematic ads

In this section, we analyze the results obtained from the crawls, examining the prevalence of problematic ads between AccAds and Control groups. The two groups are internally compared as well to understand the effect of demographics on the problematic content under each configuration. To assess the statistical significance of our findings, we employ Z-test statistics for proportions [38], since it is particularly suited for data with binary outcomes (problematic/non-problematic). This method allows us to compare the proportions of problematic labels present in the ad content.

Figure 3 illustrates the number of problematic ads for each label, derived from the **ground truth** generated during the annotation phase. The ground truth for each ad is constructed by including all labels that appear at least twice in the annotations provided by the expert annotators for that ad. Given the high IAA scores, agreement by two annotators on a single label is deemed sufficient to include it in the ground truth. This ensures a reliable and robust representation of problematic labels within the dataset.

7.1.1 Control vs AccAds Case. We compare Control and AccAds groups across all four scenarios to study the impact of the Acceptable Ads Standard (AAS) and its allowlist on the ad content. We find a significant increase in problematic ads in the AccAds group across all scenarios except for German ads, which show a non-significant increase of 5.3%. The increase in the German ads pool can be attributed to a higher number of Deceptive claims and Political ads.

Table 4: Statistical difference in the prevalence of problematic ads between different scenarios within each group. The ratio of problematic ads present is represented in column two against each scenario. Symbols: +(-) indicates increase(decrease) in problematic ads from Control to AccAds group. Asterisks (*) indicate statistical significance (p<0.05)

Scenarios	Problematic Ads Ratio		% difference
	Control	AccAds	
US	0.23	0.40	+17.61*
Germany	0.34	0.39	+5.30
over-18	0.35	0.44	+9.62*
under-18	0.26	0.47	+21.84*

For US ads, we observe a steep increase of 17.61% in problematic content from the Control to the AccAds group. This is primarily driven by Dark patterns and Regulatory violations. In the under-18 ad pool, the increase is even higher at 21.84%, largely due to Political ads and Dark patterns. For the over-18 ad pool, we find an 9.62% increase in problematic content, driven by a high prevalence of Political ads in the AccAds group, despite a lower level of User experience disruption. Overall, there is a 13.6% increase in the number of problematic ads for the AccAds group. This confirms Edelman's adverse selection theory [89], that Acceptable Ads standard is not analogous to good ad content and rather promotes more problematic content. It also contradicts ABP's claim to support non-intrusive and non-annoying advertising.

Takeaway 2: Acceptable Ads are significantly more problematic than Control group ads in all scenarios, except in Germany, where the increase was 5.30% (p>0.05). On average, there is a significant 13.6% rise in problematic ads, with every category contributing to the increase-—challenging ABP's claim of non-intrusive advertising and highlighting the heightened privacy cost for its users.

7.1.2 Control Case. In the Control group, we examine significant changes in problematic content across two scenarios: *UnAuth-US vs. UnAuth-Germany* and *over-18 vs. under-18*. These comparisons help us understand the influence of geography and age on problematic content under non-ABP conditions.

We observe a statistically significant 10.96% increase in problematic ads for Germany compared to the US (refer to Table 5). This rise is primarily driven by a higher prevalence of Dark patterns in German ads, despite their comparatively lower levels of user experience disruption. Additionally, problematic ads shown to users over 18 years of age exhibit a statistically significant 18.19% increase compared to those shown to younger audiences (excluding regulatory violations). Regulatory violations are excluded in this comparison as they apply only to under-18 ads and could overshadow other categories. The increase in problematic content for over-18 ads is mainly attributed to dark patterns and user experience disruptions, while the under-18 ad pool has more deceptive claims. Furthermore, 8.5% of under-18 ads constitute regulatory violations, including ads for dating sites and cannabis products, indicating inefficient content moderation for this demography.

7.1.3 AccAds Case. For the AccAds group, we perform the same comparisons: *UnAuth-US vs. UnAuth-Germany* and *over-18 vs. under-18*. For both demographics, the change in the proportion of problematic ads is insignificant. We observe a counterbalancing of higher Deceptive claims and Dark patterns in German ads with higher Regulatory violations and User experience disruptions in the US ads. Notably, 10.67% of under-18 ads in this scenario constitute regulatory violations. For age groups, the increase in User experience disruptions is counterbalanced by a decline in Deceptive claims for adult users. Similar to the Control group, regulatory violations are excluded to avoid overshadowing other categories.

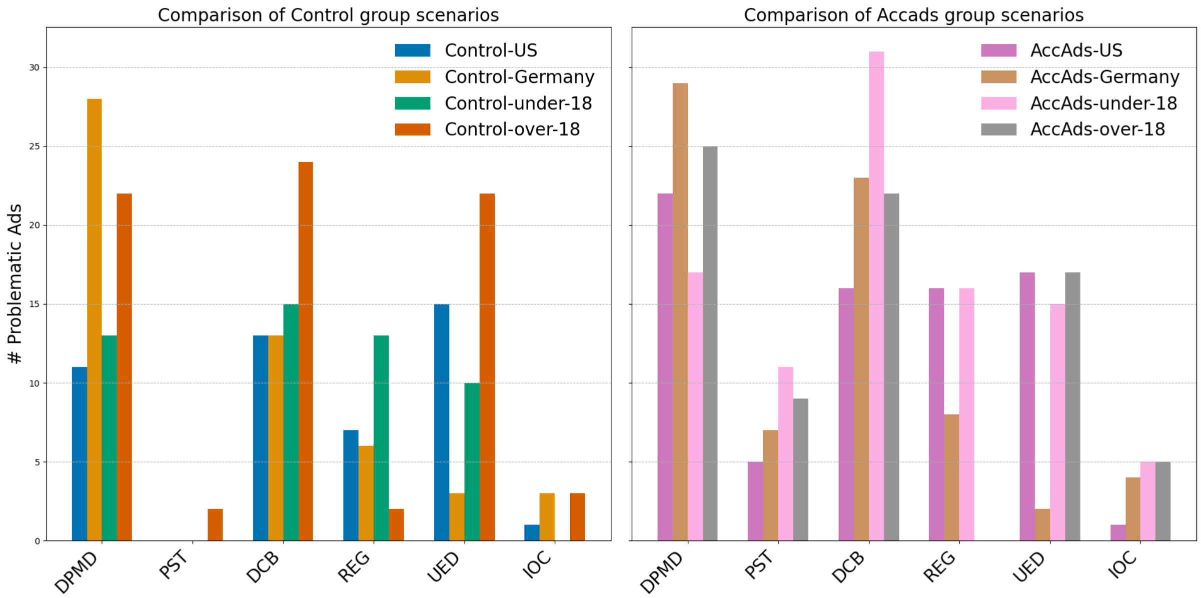

Figure 3: Prevalence of problematic ad labels across four scenarios for both Control and AccAds groups. The analysis highlights a clear increase in problematic content in the AccAds group compared to the Control group. Certain labels have zero frequency due to the limited size of the ad pool. Notably, the over-18 scenario shows zero frequency for Regulations, as advertisers are permitted to display such ads to this demographic. Abbreviations: DPMD - Dark Patterns and Manipulative Design, PST - Political and Socially Sensitive Topics, DCB - Deceptive Claims and Exaggerated Benefits, REG - Regulations, IOC - Inappropriate or Offensive Content, UED - User Experience Disruption.

Takeaway 3: Control group ads are significantly problematic for German(/adult) users compared to US(/under-age), while no significant demographic trend can be observed for Acceptable Ads. Since Acceptable Ads are proportionally more problematic, this observation suggests that other demographic vectors cease to make an observable impact under the presence of allowlisted ad exchanges.

Table 5: Statistical difference in the prevalence of problematic ads between different scenarios within each group. The ratio of problematic ads present is represented in column two against each scenario. Symbols: # - Including Regulations[6]. It has been compared separately. Plus (+) - indicates an increase in problematic ads from US/under-18 to Germany/over-18 scenarios. Asterisks (*) - indicate statistical significance (p<0.05)

Group	Problematic Ads Ratio	% difference
Control	US: 0.23 Germany: 0.34	+10.96*
Control	under-18: 0.26 (0.24#) over-18: 0.44	+18.19*
AccAds	US: 0.40 Germany: 0.39	-1.35
AccAds	under-18: 0.50 (0.47#) over-18: 0.50	0.00

7.1.4 Ad Clutter. As discussed in Section 1, ad clutter is a major source of user dissatisfaction. The Acceptable Ads Standard specifies that ads visible in the browser window upon page load must not collectively occupy more than 15% of the visible portion of the webpage. For ads placed lower on the page, this limit increases to 25%. To expand this analysis, we also calculate the number of ads rendered per page, as shown in Figure 4. Our findings indicate that

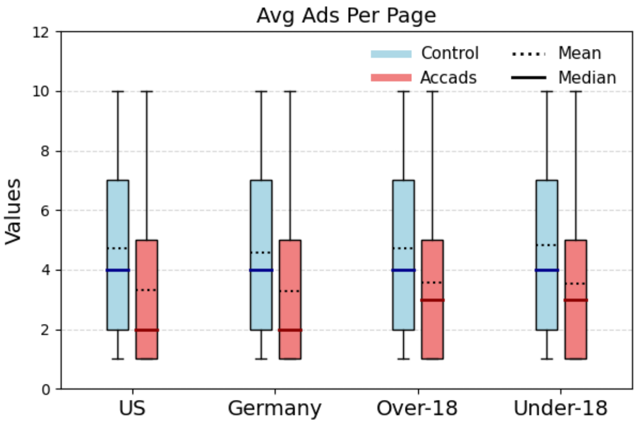

Figure 4: Plot showing ad clutter in Control and AccAds groups within the UnAuthenticated and Authenticated scenarios. Both mean and median are consistently lower for the AccAds group compared to the Control group suggesting lower ad clutter.

[6] Regulations were removed while comparing over-18 and under-18 as over-18 scenario doesn't contain that label and it would skew the overall results for under-18 if included.

the median and mean for average ads shown are consistently less for the AccAds group compared to the Control group (for each scenario), suggesting better compliance with the Acceptable Ads aesthetic guidelines. Additionally, within each group, the average ad clutter remains largely consistent across different scenarios, including the US, Germany, under-18, and over-18 categories. This consistency highlights the minimal impact of crawling location or configuration on average ad clutter levels.

7.2 Exchange Analysis

In the presence of ABP, we observe three effects on websites: new ad exchanges appear (*added*), some are removed (*blocked*), and others persist (*remaining*). The rise in problematic ads under Acceptable Ads, contrary to ABP's claim of showing only non-intrusive and non-annoying ads, raises several concerns. Do the *added* exchanges contribute to this increase, suggesting a biased selection by the AccAds allowlist maintainers? Or does the allowlist fail to include benign exchanges that served non-problematic ads? For the *remaining* exchanges, do they unintentionally degrade their ads under AccAds, potentially enabling fingerprinting of ABP users?

To address these questions, we require deeper insights into exchange behavior at the network level. We conduct mediation analysis followed by a per-exchange study to examine the effects of ad exchanges in the AccAds versus Control cases. For clarity, we use the term ad exchanges to refer to both real-time bidding endpoints and ad display networks.

7.2.1 Mediation Analysis. We investigate whether the treatment condition (AccAds) influences the problematic ads ratio through its effect on the number of ad exchanges that are *added*, *blocked*, or *remaining*. To do this, we apply mediation analysis, which tests whether the direct impact of an independent variable (the treatment condition) on a dependent variable (problematic ads ratio) diminishes when a mediator (ad exchange churn) is included. A reduction would imply that the treatment affects the ratio indirectly via its influence on churn. To gather problematic ad ratios, we visit 1,000 random sites from our website pool under both treatment (AccAds with ABP) and control (no ABP) conditions, labeling ads as problematic or not using our validated GPT-4o-mini model. To measure churn, we first establish a baseline set of exchanges per website by crawling with a fresh control browser and extracting all fetch request URLs—likely indicators of exchange communication. We then repeat this process for both conditions and compute churn via simple set operations. Finally, we perform the mediation analysis and present the results.

Our findings show significant effects of treatment condition on the number of *added*, *blocked*, and *remaining* exchanges. Specifically, in the presence of ABP, there was a significant increase of 2.74 *added* exchanges, significant increase in the number of exchanges *blocked* (8.71), and a significant average decrease of 14.48 in exchanges *remaining*. This confirms that new ad exchanges get allowlisted in the presence of ABP through Acceptable Ads allowlist, while a significant number of ad exchanges get blocked, thus reducing the number of remaining ad exchanges. Although change in the number of exchanges *blocked* or *remaining* had no significant effect on the ratio of problematic ads, surprisingly, we find that an increase in the number of *added* exchanges significantly reduces the ratio of problematic ads. In other words, the increase in *added* exchanges influenced by the presence of ABP, in fact, slightly, yet significantly, reduces the ratio of problematic ads by a log odds ratio of 0.015. Crucially, we find, regardless of mediation, the treatment condition exerts a direct significant and positive effect on the problematic ad ratio—echoing our results from Section 7.1 and also demonstrating the effectiveness of the GPT-4o-mini model in identifying problematic ads. To understand this phenomenon further, we explain a pathway of this direct effect in the next section.

7.2.2 Per Exchange Analysis. Findings from our mediation analysis imply that the churn of ad exchanges is not responsible for the increase in the ratio of problematic ads; rather, they highlight the possibility of another pathway through which the presence of ABP influences the ratio of problematic ads. In this section, we analyze whether the *remaining* allowlisted ad exchanges have a differential effect, in regard to serving problematic ads, in the presence of ABP. We operationalize this analysis by testing whether the ratio of problematic ads served by each ad exchange across all websites vary significantly in different treatment conditions. To this end, we first identify the 12 allowlisted ad exchanges that most frequently appear in our prior analysis and compute the ratio of problematic ads each served in both treatment and control conditions. We compare the ratios of problematic ads in the control condition versus the ABP condition using a t-test. This approach is designed to determine whether the presence of ABP significantly affects the proportion of problematic ads served by an exchange.

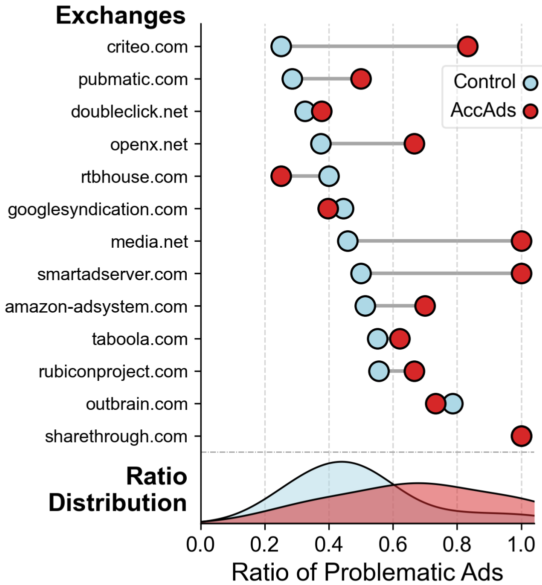

Figure 5: Problematic ads ratios for each exchange showing a general trend of increase in the AccAds case. The distribution curve is also shown at the bottom, indicating a significant shift in the problematic ads for AccAds case across all exchanges on average.

We find that most of the ad exchanges show a significant increase in their problematic ratio in the presence of ABP (see Figure 5). Specifically, we see a significant 34% increase in the problematic ad ratios from control (0.49) to treatment (0.67). This finding suggests a differential treatment towards ABP users by ad exchanges.

This is specifically alarming since these ad exchanges are allowlisted to serve non-intrusive ads, yet they undermine the Acceptable Ads program by showing a higher ratio of problematic ads under conditions where only Acceptable Ads are shown, negatively affecting privacy-aware users. This behavior also exposes these users to an increased risk of fingerprinting and targeted advertising.

> Takeaway 4: While new ad exchanges added by the Acceptable Ads allowlist decrease problematic ad ratios on websites, ad exchanges that do not get blocked increase it. This shows that ad exchanges inadvertently differentially treat ABP users to show them problematic ads–nullifying the claim for non-intrusive and non-annoying advertisements and making ABP users susceptible to fingerprinting, tracking and poor web experience.

8 Summary

In this study, we developed a keyword-based taxonomy for identifying problematic ads accounting for the perspectives of all stakeholders in the ad ecosystem such as advertisers, publishers, users, etc. Using this taxonomy as prompt, we addressed the challenge of automated detection of problematic ads by leveraging OpenAI's GPT-4o-mini model and achieving a Krippendorff alpha agreement score of 0.79 for binary classification and 0.74 for multi-label classification, indicating strong alignment with expert human annotators.

We manually annotated 1,200 ads collected from user profiles in the US and Germany, covering both under-18 and over-18 age groups. Overall, 35.87% of the ads were found to be problematic. Notably, 9.57% of ads shown to underage users violated regulations, highlighting the failure of ad exchanges to safeguard younger audiences. Alarmingly, Acceptable Ads were 13.6% more problematic than regular ones. We also examined how ad exchanges' behavior changes under Acceptable Ads and found that newly added exchanges or networks reduced the prevalence of problematic ads, while existing ones increased it, revealing uneven compliance across the ecosystem.

9 Conclusion and Discussion

Conclusion. By establishing the effectiveness of LLMs in detecting problematic content, we highlight the risks privacy-aware users face from adversaries who can profile them using in-browser LLMs. The higher prevalence of problematic ads in Acceptable Ads underscores a critical issue: While programs like Acceptable Ads aim to balance user and advertiser interests by permitting less disruptive ads, their standards often fall short of addressing user concerns comprehensively. The misleading nomenclature of terms like "acceptable" or "better" ads creates a perception of enhanced user experience, which is not fully realized.

Moreover, the differential treatment of adblock users by ad exchanges or display networks highlight another important issue - privacy aware users are worse off when they try to enhance their privacy, often leading them to be more exposed towards targeting and fingerprinting. These concerns raise an important question for the privacy research community: are the tools designed to protect user privacy truly effective, or do they introduce new vulnerabilities in the process, essentially raising the cost of privacy for privacy-aware users?

Limitations. Measuring fraud and scam content poses significant challenges in automating its detection using ad creatives and landing pages. While different studies [51, 64, 111, 138, 146] have tried to measure fraud and scams on the network, they often rely on signature-based or anomaly-based detection that involves using pre-existing lists of malicious advertisers and associating them with fraudulent behaviors in the past [83, 157]. Investigating the legitimacy of the products advertised makes this category hard for any ML classifier to automate, as it is usually carried out by analyzing public forums and user reviews where real users verify their legitimacy [72, 86, 95, 129]. Additionally, our crawling framework's focus on affluent Western democracies like the US and Germany may not capture the global scope of problematic advertising. Less affluent regions, with varying regulations and economic incentives, could face a higher prevalence of problematic ads [8], suggesting our findings may underestimate issues at the global level.

Differential Treatment. Despite the possibility of various confounding factors within the ad ecosystem such as the value of the user profile, tracked information of the user, etc., individual exchange analysis revealing a consistent pattern where most ad exchanges were found to be increasing the prevalence of problematic ads on websites raises serious questions. Do ad exchanges detect the presence of these privacy-preserving extensions and intentionally target their users with problematic content? Or does blocking trackers degrade the value of the user profile, resulting in problematic ads with cheaper bids making their way to the user? The former signifies the differential treatment of adblocker users at the exchange level, which aligns with similar observations on Youtube and Twitch, suggesting the degradation of user experience for adblocker users to nudge them towards disabling adblocker. The latter is a more innocuous effect of blocking trackers, resulting in less information about the user and leading to poor ad quality.

While it's hard to comment concretely on either of them, this additional information opens up new avenues for fingerprinting privacy-aware users, compromising their privacy, and making them vulnerable to attack. This is also corroborated by previous findings [93, 104, 105] that show the adverse effect of fingerprinting filter lists on privacy-aware users.

Content Moderation. The possibility of putting privacy-aware users at risk makes it more eminent for modern-day adblockers to innovate their filtering process in order to improve user experience and privacy, along with allowing non-intrusive advertising. LLM-assisted detection of problematic content provides them with a unique way to monitor ad exchanges that repeatedly render problematic ads at scale, block them, and replace them with more benign ad exchanges in the allowlist. With the advent of In-browser LLMs [37] and efficient smaller models, our study puts privacy-preserving browsers like Adblock Browser, Brave, Mozilla, etc. in a unique position where they can innovate filtering and show actual "Acceptable" ads to users. This would also motivate ad exchanges to increase content moderation on their respective platforms to remain profitable and contribute positively to the ad ecosystem.

10 Acknowledgement

We thank the anonymous PETs reviewers for their valuable feedback and insights. We are also grateful to Cat Mai, Meghna Manoj Nair, Prajata Roy, and Tobias Lauinger for their support with the labeling tasks, which were critical to this study. Finally, we thank Prof. Zubair Shafiq for his helpful feedback and perspectives.

References

[1] 2014. You might also like this story about weaponized clickbait - The Verge. Online. https://www.theverge.com/2014/4/22/5639892/how-weaponized-clickbait-took-over-the-web

[2] 2015. Native Advertising: A Guide for Businesses | Federal Trade Commission. Online. https://www.ftc.gov/business-guidance/resources/native-advertising-guide-businesses

[3] 2016. Publishers Are Rethinking Those 'Around the Web' Ads - The New York Times. https://www.nytimes.com/2016/10/31/business/media/publishers-rethink-outbrain-taboola-ads.html

[4] 2017. Examples of Russian facebook ads - Washington Post. Online. https://www.washingtonpost.com/graphics/2017/business/russian-ads-facebook-targeting/

[5] 2017. Faiss: A library for efficient similarity search. Online. https://engineering.fb.com/2017/03/29/data-infrastructure/faiss-a-library-for-efficient-similarity-search/

[6] 2017. Fake news is making headlines. But what about fake ads? | WIRED. Online. https://www.wired.com/story/fake-news-outbrain-taboola-hillary-clinton/

[7] 2019. What Are Bad Ads? Why Do They Matter? And How Can You Block Them? Online. https://blog.adlightning.com/what-are-bad-ads-why-do-they-matter-and-how-can-you-block-them

[8] 2020. How outdoor advertising can deepen inequality. Online. https://www.bbc.com/worklife/article/20200817-the-inequality-of-outdoor-advertising-exposure

[9] 2021. 20 Display Advertising Stats That Demonstrate Digital Advertising's Evolution. Online. https://blog.hubspot.com/marketing/horrifying-display-advertising-stats

[10] 2023. Ad Blocking Will Be a $54B Publisher Problem in 2024. Online. https://www.admonsters.com/ad-blocking-a-54b-problem-for-publishers-in-2024/

[11] 2023. eyeo Ad-Filtering Report. Online. https://26760064.fs1.hubspotusercontent-eu1.net/hubfs/26760064/2023%20eyeo%20Ad-Filtering%20Report.pdf?utm_campaign=2023%20eyeo%20Ad-Filtering%20Report&utm_medium=email&_hsmi=79020731&utm_content=79020731&utm_source=hs_email

[12] 2023. Puppeteer. Online. https://pptr.dev/

[13] 2023. What is GDPR, the EU's new data protection law? Online. https://gdpr.eu/what-is-gdpr/

[14] 2024. About demographic targeting. Online. https://support.google.com/google-ads/answer/2580383?hl=en

[15] 2024. Acceptable Ads Standard. Online. https://acceptableads.com/

[16] 2024. Ad Blocker Usage and Demographic Statistics. Online. https://backlinko.com/ad-blockers-users#adblock-usage-reasons

[17] 2024. Adblock Plus. Online. https://chromewebstore.google.com/detail/adblock-plus-free-ad-bloc/cfhdojbkjhnklbpkdaibdccddilifddb

[18] 2024. Amazon - Advertising Policies. Online. https://advertising.amazon.com/resources/ad-policy/creative-acceptance/

[19] 2024. Bing Ads - Advertising Policies. Online. https://about.ads.microsoft.com/en-us/cn-policy-for-pilot.pdf

[20] 2024. Coalition For Better Ads. Online. https://www.betterads.org/

[21] 2024. DataReportal Library. Online. https://datareportal.com/library

[22] 2024. Digital Advertising Spend. Online. https://www.statista.com/outlook/dmo/digital-advertising/worldwide#ad-spending

[23] 2024. Ghostery. Online. https://www.ghostery.com/ghostery-ad-blocker

[24] 2024. Google Ads policies - Advertising Policies. Online. https://support.google.com/adspolicy/answer/6008942

[25] 2024. GPT-4o mini: advancing cost-efficient intelligence. Online. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

[26] 2024. IAB - Advertising Policies. Online. https://www.iab.com/guidelines/iab-new-ad-portfolio/

[27] 2024. Jaccard Index / Similarity Coefficient. Online. https://www.statisticshowto.com/jaccard-index/

[28] 2024. Meta ads - Advertising Policies. Online. https://transparency.meta.com/policies/ad-standards/business-assets

[29] 2024. OpenX - Advertising Policies. Online. https://www.openx.com/legal/ad-exchange-supply-policy/

[30] 2024. OutBrain - Advertising Policies. Online. https://www.outbrain.com/advertisers/guidelines/

[31] 2024. pi-hole. Online. https://pi-hole.net/

[32] 2024. Privacy Badger. Online. https://privacybadger.org/

[33] 2024. Pubmatic - Advertising Policies. Online. https://pubmatic.com/legal/aq-policy/

[34] 2024. Taboola - Advertising Policies. Online. https://help.taboola.com/hc/en-us/sections/115001764928-Advertising-Policies

[35] 2024. The Truth Behind Weight Loss Ads | Consumer Advice. Online. https://consumer.ftc.gov/articles/truth-behind-weight-loss-ads

[36] 2024. Ublock Origin. Online. https://chromewebstore.google.com/detail/ublock-origin/cjpalhdlnbpafiamejdnhcphjbkeiagm?hl=en&pli=1

[37] 2024. WebLLMs. Online. https://webllm.mlc.ai/

[38] 2024. Z-Test: Definition, Uses in Statistics, and Example. Online. https://www.investopedia.com/terms/z/z-test.asp

[39] 2025. Consent or pay Summary of call for views . Online. https://ico.org.uk/media2/0zeatdun/20240417-corp-cfv-reply-uk-dpa_redacted.pdf

[40] AAC. 2018. MOBILE ADVERTISING STUDY MEASURING AD-BLOCKING USERS' PERCEPTIONS OF ADVERTISING TYPES ON MOBILE BROWSERS COMMISSIONED BY: Acceptable Ads Commitee. (2018). https://marketingland.com/the-iab-takes-on-ad-blocking-by-first-admitting-the-industry-screwed-up-147235

[41] AAC. 2020. Video advertisement study Measuring ad-blocking users' perceptions of types of video advertisement. (2020).

[42] AAC. 2021. In-article and in-gallery ads survey Measuring ad-blocking users' perceptions of ads placed in primary content The Acceptable Ads Committee. (2021).

[43] AAC. 2022. In-view ad refresh survey Measuring ad-blocking users' perceptions of in-view ad refresh. (2022).

[44] AAC. 2023. User perceptions of sustainable online advertising Qualitative interviews on how internet users perceive sustainable online advertising initiatives. (2023).

[45] AAC. 2024. In-Article Ad Survey: 300x250 Revisited Revisiting ad-blocking users' perceptions of a 300x250 ad placed in primary content. (2024).

[46] Gunes Acar, Christian Eubank, Steven Englehardt, Marc Juarez, Arvind Narayanan, and Claudia Diaz. 2014. The Web Never Forgets: Persistent Tracking Mechanisms in the Wild. (2014). https://doi.org/10.1145/2660267.2660347

[47] Meltem Akçaboy, Saliha Şenel, and Betül Ulukol. 2022. Advertising in the Digital Area: Playing Online Games Pose a Risk to Come Upon Improper Advertisements. *Turkish Archives of Pediatrics* 57, 2 (mar 2022), 241–243. https://doi.org/10.5152/TURKARCHPEDIATR.2022.21335

[48] Muhammad Ali, Angelica Goetzen, Alan Mislove, Elissa Redmiles, and Piotr Sapiezynski. 2023. All Things Unequal: Measuring Disparity of Potentially Harmful Ads on Facebook. (2023).

[49] Athanasios Andreou, Giridhari Venkatadri, Oana Goga, Krishna P. Gummadi, Patrick Loiseau, and Alan Mislove. 2018. Investigating Ad Transparency Mechanisms in Social Media: A Case Study of Facebook's Explanations. *25th Annual Network and Distributed System Security Symposium, NDSS 2018* (2018). https://doi.org/10.14722/NDSS.2018.23191

[50] Anna D ' Annunzio and Antonio Russo. 2017. Ad Networks, Consumer Tracking, and Privacy. (2017). www.RePEc.org

[51] Brandon Atkins and Wilson Huang. 2013. A Study of Social Engineering in Online Frauds. *Open Journal of Social Sciences* 1, 3 (2013), 23–32. https://doi.org/10.4236/jss.2013.13004

[52] Nielsen Rasmus Kleis Atlay Sacha and Fletcher Richard. 2022. Quantifying the "Infodemic": People Turned to Trustworthy News Outlets During the 2020 Coronavirus Pandemic. *Journal of Quantitative Description: Digital Media* (2022). https://journalqd.org/article/view/3617/2703

[53] Mika D Ayenson, Dietrich James Wambach, Ashkan Soltani, Nathan Good, and Chris Jay Hoofnagle. 2011. Flash Cookies and Privacy II: Now with HTML5 and ETag Respawning. *SSRN Electronic Journal* (7 2011). https://doi.org/10.2139/SSRN.1898390

[54] Michele Banko, Brendon Mackeen, and Laurie Ray. 2020. A Unified Typology of Harmful Content. (2020). https://doi.org/10.18653/v1/P17

[55] Paul Barford, Igor Canadi, Darja Krushevskaja, Qiang Ma, and S Muthukrishnan. 2014. Adscape: Harvesting and Analyzing Online Display Ads. (2014). http://www.worldwidewebsize.com/

[56] Glaser Barney and Strauss Anselm. 1999. *Discovery of Grounded Theory: Strategies for Qualitative Research.*

[57] Nataliia Bielova, Laura Litvine, Anysia Nguyen, Vincent Toubiana, and Estelle Hary. 2024. The Effect of Design Patterns on (Present and Future) Cookie Consent Decisions. (2024). https://www.usenix.org/conference/usenixsecurity24/presentation/bielova

[58] Christoph Bösch, Benjamin Erb, Frank Kargl, Henning Kopp, and Stefan Pfattheicher. 2016. Tales from the Dark Side: Privacy Dark Strategies and Privacy Dark Patterns. *Proceedings on Privacy Enhancing Technologies* 4 (2016), 237–254. https://doi.org/10.1515/popets-2016-0038

[59] Christian Böttger, Nurullah Demir, Jan Hörnemann, Bhupendra Acharya, Norbert Pohlmann, Thorsten Holz, Matteo Grosse-Kampmann, and Tobias Urban. 2025. Understanding Regional Filter Lists: Efficacy and Impact. *Proceedings on Privacy Enhancing Technologies* 2025, 2 (apr 2025), 309–325. https:

//doi.org/10.56553/POPETS-2025-0063

[60] Giorgio Brajnik and Silvia Gabrielli. 2010. A Review of Online Advertising Effects on the User Experience. *INTL. JOURNAL OF HUMAN-COMPUTER INTERACTION* 26, 10 (2010), 971–997. https://doi.org/10.1080/10447318.2010.502100

[61] Joshua A. Braun and Jessica L. Eklund. 2019. Fake News, Real Money: Ad Tech Platforms, Profit-Driven Hoaxes, and the Business of Journalism. *Digital Journalism* 7, 1 (jan 2019), 1–21. https://doi.org/10.1080/21670811.2018.1556314/ASSET/65C76EE4-53C2-4FCF-AD1E-18F6CAD06442/ASSETS/IMAGES/RDIJ_A_1556314_F0001_B.JPG

[62] Moniek Buijzen and Patti M Valkenburg. 2003. The effects of television advertising on materialism, parent-child conflict, and unhappiness: A review of research. (2003). https://doi.org/10.1016/S0193-3973(03)00072-8

[63] Moira Burke and • M Burke. 2005. High-Cost Banner Blindness: Ads Increase Perceived Workload, Hinder Visual Search, and Are Forgotten. *ACM Transactions on Computer-Human Interaction* 12, 4 (2005), 423–445.

[64] David Burnes, Charles R. Henderson, Christine Sheppard, Rebecca Zhao, Karl Pillemer, and Mark S. Lachs. 2017. Prevalence of financial fraud and scams among older adults in the United States: A systematic review and meta-analysis. *American Journal of Public Health* 107, 8 (aug 2017), e13–e21. https://doi.org/10.2105/AJPH.2017.303821/FORMAT/EPUB

[65] Selensky J. C. and Carels R. A. 2021. Weight stigma and media: An examination of the effect of advertising campaigns on weight bias, internalized weight bias, self-esteem, body image, and affect. *Body Image* (2021).

[66] Colin Campbell and Pamela E. Grimm. 2019. The Challenges Native Advertising Poses: Exploring Potential Federal Trade Commission Responses and Identifying Research Needs. *Journal of Public Policy and Marketing* 38, 1 (jan 2019), 110–123. https://doi.org/10.1177/0743915618818576/SUPPL_FILE/WEB_APPENDIX_-_818576.PDF

[67] Yinzhi Cao, Song Li, and Erik Wijmans. 2017. (Cross-)Browser Fingerprinting via OS and Hardware Level Features. *24th Annual Network and Distributed System Security Symposium, NDSS 2017* (2017). https://doi.org/10.14722/NDSS.2017.23152

[68] CAPV 2021. 2021. Age-restricted ads online Advertising Guidance (non-broadcast). (2021).

[69] Darion Cassel, Su-Chin Lin, Alessio Buraggina, William Wang, Andrew Zhang, Lujo Bauer, Hsu-Chun Hsiao, Limin Jia, and Timothy Libert. 2022. OmniCrawl: Comprehensive Measurement of Web Tracking With Real Desktop and Mobile Browsers. *Proceedings on Privacy Enhancing Technologies* (2022), 227–252. Issue 1. https://doi.org/10.2478/popets-2022-0012

[70] Quan Chen, Panagiotis Ilia, Michalis Polychronakis, and Alexandros Kapravelos. 2021. Cookie swap party: Abusing first-party cookies for web tracking. *The Web Conference 2021 - Proceedings of the World Wide Web Conference, WWW 2021* (4 2021), 2117–2129. https://doi.org/10.1145/3442381.3449837

[71] Chang-Hoan Cho and Hongsik John Cheon. 2004. Why Do People Avoid Advertising on the Internet? *Source: Journal of Advertising* 33, 4 (2004), 89–97.

[72] Huang Chua, Jonathan Wareham, and Daniel Robey. 2007. The role of online trading communities in managing internet auction fraud. *MIS Quarterly: Management Information Systems* 31, 4 (2007), 759–781. https://doi.org/10.2307/25148819

[73] Coalition for Better Ads. 2016. An experimental methodology to measure consumers' perceptions of online ad experiences. April (2016), 1–73.

[74] Coalition for Better Ads. 2016. An experimental methodology to rank N ad experiences by consumers ' perceptions. April (2016), 1–73.

[75] Coalition for Better Ads. 2017. An experimental methodology to measure consumer perceptions of ads in short-form video. (2017).

[76] Coalition for Better Ads. 2017. Determining a Better Ads Standard Based on User Experience Data. (2017), 1–46.

[77] CONFIANT. 2023. Malvertising and Ad Quality Index. (2023).

[78] CONFIANT. 2024. Malvertising and Ad Quality Index. (2024).

[79] Henriette Cramer. 2015. Effects of Ad Quality & Content-Relevance on Perceived Content Quality. (2015). https://doi.org/10.1145/2702123.2702360

[80] Savino Dambra, Iskander Sanchez-Rola, Leyla Bilge, and Davide Balzarotti Eurecom. 2022. When Sally Met Trackers: Web Tracking From the Users' Perspective. (2022). https://www.usenix.org/conference/usenixsecurity22/presentation/dambra

[81] Amit Datta, Michael Carl Tschantz, and Anupam Datta. 2015. Automated Experiments on Ad Privacy Settings. *Proceedings on Privacy Enhancing Technologies* 2015, 1 (2015), 92–112. https://doi.org/10.1515/popets-2015-0007

[82] Lange Rudi De. 2013. The portrayal of slimness through design : an analysis of a misleading weight loss advertisement : research article. *South African Journal of Art History* (2013). https://journals.co.za/doi/epdf/10.10520/EJC149213

[83] Joe Deblasio, Saikat Guha, Geoffrey M. Voelker, and Alex C. Snoeren. 2017. Exploring the dynamics of search advertiser fraud. *Proceedings of the ACM SIGCOMM Internet Measurement Conference, IMC* Part F1319 (2017), 157–170. https://doi.org/10.1145/3131365.3131393

[84] Nurullah Demir, Tobias Urban, Norbert Pohlmann, and Christian Wressnegger. 2024. A Large-Scale Study of Cookie Banner Interaction Tools and Their Impact on Users' Privacy. *Proceedings of Proceedings on Privacy Enhancing Technologies* 2024 (2024).

[85] Tom Dobber, Sanne Kruikemeier, Natali Helberger, and Ellen Goodman. 2023. Shielding citizens? Understanding the impact of political advertisement transparency information. *New Media and Society* (11 2023). https://doi.org/10.1177/14614448231157640/SUPPL_FILE/SJ-DOCX-1-NMS-10.1177_14614448231157640.DOCX

[86] Wei Dong, Shaoyi Liao, and Zhongju Zhang. 2018. Leveraging Financial Social Media Data for Corporate Fraud Detection. *Journal of Management Information Systems* 35, 2 (2018), 461–487. https://doi.org/10.1080/07421222.2018.1451954

[87] Anna D'Annunzio and Antonio Russo. 2019. Ad Networks and Consumer Tracking. *https://doi.org/10.1287/mnsc.2019.3481* 66 (11 2019), 5040–5058. Issue 11. https://doi.org/10.1287/MNSC.2019.3481

[88] Peter Eckersley. 2010. How Unique Is Your Web Browser? (2010). https://panopticlick.eff.org.

[89] Benjamin Edelman. 2010. Adverse selection in online â€œtrustâ€• certifications and search results. (2010). https://doi.org/10.1016/j.elerap.2010.06.001

[90] Laura Edelson, Shikhar Sakhuja, Ratan Dey, and Damon McCoy. 2019. An Analysis of United States Online Political Advertising Transparency. (2 2019). http://arxiv.org/abs/1902.04385

[91] Steven M. Edwards, Hairong Li, and Joo Hyun Lee. 2002. Forced exposure and psychological reactance: Antecedents and consequences of the perceived intrusiveness of pop-up ads. *Journal of Advertising* 31, 3 (2002), 83–95. https://doi.org/10.1080/00913367.2002.10673678

[92] Saiid El, Hajj Chehade, Sandra Siby, and Carmela Troncoso. 2023. SINBAD: Saliency-informed detection of breakage caused by ad blocking. (2023).

[93] Saiid El, Hajj Chehade, Ben Stock, Carmela Troncoso, and § Epfl. 2025. Double-Edged Shield: On the Fingerprintability of Customized Ad Blockers. *USENIX Security* (2025).

[94] Steven Englehardt and Arvind Narayanan. 2016. Online Tracking: A 1-million-site Measurement and Analysis. (2016). https://doi.org/10.1145/2976749.2978313

[95] Gabriel Freedman and Francesca Toni. lfr2024. Detecting Scientific Fraud Using Argument Mining. (lfr2024), 15–28.

[96] FTC. 2015. Enforcement Policy Statement on Deceptively Formatted Advertisements. (2015).

[97] Liza Gak. 2022. The Distressing Ads That Persist: Uncovering The Harms of Targeted Weight-Loss Ads Among Users with Histories of Disordered Eating. (2022). https://doi.org/10.1145/3555102

[98] Tarleton Gillespie. 2020. Content moderation, AI, and the question of scale. *Big Data and Society* 7 (7 2020). Issue 2. https://doi.org/10.1177/2053951720943234

[99] Gustavo Gomez-Mejia. 2020. "Fail, Clickbait, Cringe, Cancel, Woke": Vernacular Criticisms of Digital Advertising in Social Media Platforms. (2020). https://doi.org/10.1007/978-3-030-49576-3_23

[100] Colin M. Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt, and Austin L. Toombs. 2018. The dark (patterns) side of UX design. *Conference on Human Factors in Computing Systems - Proceedings* 2018-April (apr 2018). https://doi.org/10.1145/3173574.3174108

[101] Colin M. Gray, Cristiana Teixeira Santos, Nataliia Bielova, and Thomas Mildner. 2024. An Ontology of Dark Patterns Knowledge: Foundations, Definitions, and a Pathway for Shared Knowledge-Building. *Proceedings of the CHI Conference on Human Factors in Computing Systems* (may 2024), 1–22. https://doi.org/10.1145/3613904.3642436

[102] Savannah Greenfield. 2018. When Beauty is the Beast: The Effects of Beauty Propaganda on When Beauty is the Beast: The Effects of Beauty Propaganda on Female Consumers Female Consumers. (2018). https://digitalcommons.unomaha.edu/university_honors_program

[103] Aniko Hannak, Gary Soeller, David Lazer, Alan Mislove, and Christo Wilson. 2014. Measuring price discrimination and steering on E-commerce web sites. *Proceedings of the ACM SIGCOMM Internet Measurement Conference, IMC* (11 2014), 305–318. https://doi.org/10.1145/2663716.2663744

[104] Umar Iqbal, Steven Englehardt, and Zubair Shafiq. 2021. Fingerprinting the Fingerprinters: Learning to Detect Browser Fingerprinting Behaviors. *2021 IEEE Symposium on Security and Privacy (SP)* (2021). https://doi.org/10.1109/SP40001.2021.00017

[105] Umar Iqbal, Zubair Shafiq, and Zhiyun Qian. 2017. The ad wars: Retrospective measurement and analysis of anti-adblock filter lists. *Proceedings of the ACM SIGCOMM Internet Measurement Conference, IMC* Part F131937 (nov 2017), 171–183. https://doi.org/10.1145/3131365.3131387

[106] Umar Iqbal, Peter Snyder, Shitong Zhu, Benjamin Livshits, Zhiyun Qian, and Zubair Shafiq. 2018. ADGRAPH: A Graph-Based Approach to Ad and Tracker Blocking. (2018).

[107] Pablo Jost, Simon Kruschinski, | Michael Sülflow, Jörg Haßler, and | Marcus Maurer. 2022. Invisible transparency: How different types of ad disclaimers on Facebook affect whether and how digital political advertising is perceived. (2022). https://doi.org/10.1002/poi3.333

[108] Tami Kim, Kate Barasz, and Leslie K. John. 2019. Why am i seeing this ad? The effect of ad transparency on ad effectiveness. *Journal of Consumer Research* 45 (2 2019), 906–932. Issue 5. https://doi.org/10.1093/JCR/UCY039

[109] Krippendorf Klaus. 2013. Computing Krippendorff's Alpha-Reliability. *https://doi.org/10.1177/2053951719897945* (2013). https://www.asc.upenn.edu/sites/default/files/2021-03/Computing%20Krippendorff%27s%20Alpha-Reliability.pdf

[110] Tadayoshi Kohno, Andre Broido, and K. C. Claffy. 2005. Remote physical device fingerprinting. *IEEE Transactions on Dependable and Secure Computing* 2 (2005), 93–108. Issue 2. https://doi.org/10.1109/TDSC.2005.26

[111] Sai Venkata Jaswant Kolupuri, Ananya Paul, Rajat Subhra Bhowmick, and Isha Ganguli. 2025. Scams and Frauds in the Digital Age: ML-Based Detection and Prevention Strategies. *ICDCN 2025 - Proceedings of the 26th International Conference on Distributed Computing and Networking* (2025), 340–345. https://doi.org/10.1145/3700838.3703672

[112] Elif Kubilay, Eva Raiber, Lisa Spantig, Jana Cahlíková, and Lucy Kaaria. 2023. Can you spot a scam? Measuring and improving scam identification ability *. (2023).

[113] Tomer Laor, Naif Mehanna, Antonin Durey, Vitaly Dyadyuk, Pierre Laperdrix, Clémentine Maurice, Yossi Oren, Romain Rouvoy, Walter Rudametkin, and Yuval Yarom. 2022. DRAWNAPART: A Device Identification Technique based on Remote GPU Fingerprinting. *29th Annual Network and Distributed System Security Symposium, NDSS 2022* (1 2022). https://doi.org/10.14722/NDSS.2022.24093

[114] Pierre Laperdrix, Quan Chen, Alexandros Kapravelos, Oleksii Starov, and Nick Nikiforakis. 2021. Open access to the Proceedings of the 30th USENIX Security Symposium is sponsored by USENIX. Fingerprinting in Style: Detecting Browser Extensions via Injected Style Sheets Fingerprinting in Style: Detecting Browser Extensions via Injected Style Sheets. (2021). www.usenix.org/conference/usenixsecurity21/presentation/laperdrix

[115] Zengrui Liu, Umar Iqbal, and Nitesh Saxena. 2024. Opted Out, Yet Tracked: Are Regulations Enough to Protect Your Privacy? *Proceedings of Proceedings on Privacy Enhancing Technologies* 2024 (2024). https://doi.org/10.56553/popets-2024-0016

[116] Karen L Malliama. 2009. A Critical Analysis of the Changing Nature of Religious Imagery in Advertising. *Journal of Media and Religion* 8, 3 (2009), 172–190. https://doi.org/10.1080/15348420903091162

[117] Veronica Marotta, Vibhanshu Abhishek, and Alessandro Acquisti. 2019. Online Tracking and Publishers' Revenues: An Empirical Analysis. (2019). http://www.heinz.cmu.edu/Ëœacquisti/cv.htm.

[118] Zahra Moti; Asuman Senol; Hamid Bostani; Frederik Zuiderveen Borgesius; Veelasha Moonsamy; Arunesh Mathur. 2024. Targeted and troublesome: Tracking and advertising on children's websites. (2024). https://ieeexplore.ieee.org/abstract/document/10646733

[119] Jonathan R. Mayer and John C. Mitchell. 2012. Third-party web tracking: Policy and technology. *Proceedings - IEEE Symposium on Security and Privacy* (2012), 413–427. https://doi.org/10.1109/SP.2012.47

[120] Tinhinane Medjkoune, Oana Goga, and Juliette Senechal. 2023. Marketing to Children Through Online Targeted Advertising: Targeting Mechanisms and Legal Aspects. *CCS 2023 - Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security* (nov 2023), 180–194. https://doi.org/10.1145/3576915.3623172

[121] Victor Morel, Cristiana Santos, Yvonne Lintao, and Soheil Human. 2022. Your Consent Is Worth 75 Euros A Year – Measurement and Lawfulness of Cookie Paywalls. *WPES 2022 - Proceedings of the 21st Workshop on Privacy in the Electronic Society, co-located with CCS 2022* (9 2022), 213–218. https://doi.org/10.1145/3559613.3563205

[122] Garrett Morrow, Briony Swire-Thompson, Jessica Montgomery Polny, Matthew Kopec, and John P. Wihbey. 2022. The emerging science of content labeling: Contextualizing social media content moderation. *Journal of the Association for Information Science and Technology* 73 (10 2022), 1365–1386. Issue 10. https://doi.org/10.1002/ASI.24637

[123] Keaton Mowery and Hovav Shacham. 2012. Pixel Perfect: Fingerprinting Canvas in HTML5. (2012). http://www.joelonsoftware.com/items/

[124] Sekharan Chandra N. and Vuppala Pavan Sai. 2023. Fine-Tuned Large Language Models for Improved Clickbait Title Detection. (2023). https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10487353&tag=1

[125] Robert C. Nickerson, Upkar Varshney, and Jan Muntermann. 2013. A method for taxonomy development and its application in information systems. *European Journal of Information Systems* 22 (2013), 336–359. Issue 3. https://www.tandfonline.com/action/journalInformation?journalCode=tjis20

[126] Emmanouil Papadogiannakis, Nicolas Kourtellis, Panagiotis Papadopoulos, and Evangelos Markatos. 2025. Welcome to the Dark Side: Analyzing the Revenue Flows of Fraud in the Online Ad Ecosystem. *Proceedings of the ACM on Web Conference 2025* (4 2025), 1522–1535. https://doi.org/10.1145/3696410.3714899

[127] Emmanouil Papadogiannakis, Panagiotis Papadopoulos, Nicolas Kourtellis, and Evangelos P. Markatos. 2021. User tracking in the post-cookie era: How websites bypass gdpr consent to track users. *The Web Conference 2021 - Proceedings of the World Wide Web Conference, WWW 2021* (4 2021), 2130–2141. https://doi.org/10.1145/3442381.3450056

[128] Emmanouil Papadogiannakis, Panagiotis Papadopoulos, Evangelos P. Markatos, and Nicolas Kourtellis. 2024. Ad Laundering: How Websites Deceive Advertisers into Rendering Ads Next to Illicit Content. *WWW 2024 Companion - Companion Proceedings of the ACM Web Conference* (5 2024), 782–785. https://doi.org/10.1145/3589335.3651466

[129] Himangshu Paul and · Alexander Nikolaev. 2021. Fake review detection on online E-commerce platforms: a systematic literature review. *Data Mining and Knowledge Discovery* 35 (2021), 1830–1881. https://doi.org/10.1007/s10618-021-00772-6

[130] Andrew C Pickett, Katie M Brown, and Zack J Damon. 2023. Potentially Misleading Weight Loss Advertisements Targeting Potentially Misleading Weight Loss Advertisements Targeting Men: Examining Influence of Celebrity Athlete Endorsement on Men: Examining Influence of Celebrity Athlete Endorsement on Ad Believability and Purchase Intentions Ad Believability and Purchase Intentions. *Health Behavior Research* 6 (2023). https://doi.org/10.4148/2572-1836.1177

[131] Prabaharan Poornachandran, N. Balagopal, Soumajit Pal, Aravind Ashok, Prem Sankar, and Manu R. Krishnan. 2017. Demalvertising: A Kernel Approach for Detecting Malwares in Advertising Networks. *Advances in Intelligent Systems and Computing* 458 (2017), 215–224. https://doi.org/10.1007/978-981-10-2035-3_23

[132] Wei Qiao, Tushar Dogra, Otilia Stretcu, Yu Han Lyu, Tiantian Fang, Dongjin Kwon, Chun Ta Lu, Enming Luo, Yuan Wang, Chih Chun Chia, Ariel Fuxman, Fangzhou Wang, Ranjay Krishna, and Mehmet Tek. 2024. Scaling Up LLM Reviews for Google Ads Content Moderation. *WSDM 2024 - Proceedings of the 17th ACM International Conference on Web Search and Data Mining* (3 2024), 1174–1175. https://doi.org/10.1145/3616855.3635736

[133] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. (2021). arXiv:2103.00020v1 https://github.com/OpenAI/CLIP.

[134] Nicola Rayner. 2018. The international. *Dancing Times* 109, 1300 (2018), 72–75. https://doi.org/10.4324/9781315756233-6

[135] Roongta Ritik and Greenstadt Rachel. 2024. From User Insights to Actionable Metrics: A User-Focused Evaluation of Privacy-Preserving Browser Extensions. (2024).

[136] Christian Rohrer and John Boyd. 2004. The Rise of Intrusive Online Advertising and the Response of User Experience Research at Yahoo! THE RISE OF INTRUSIVE ONLINE ADVERTISING. (2004).

[137] Ritik Roongta, Mitchell Zhou, Ben Stock, and Rachel Greenstadt. 2024. From Blocking to Breaking: Evaluating the Impact of Adblockers on Web Usability. (2024).

[138] Shadi Sadeghpour and Natalija Vlajic. 2021. Ads and Fraud: A Comprehensive Survey of Fraud in Online Advertising. *Journal of Cybersecurity and Privacy* (2021). https://doi.org/10.3390/jcp1040039

[139] Shadi Sadeghpour and Natalija Vlajic. 2021. Click Fraud in Digital Advertising: A Comprehensive Survey. (2021). https://doi.org/10.3390/computers10120164

[140] Iskander Sanchez-Rola, Xabier Ugarte-Pedrero, Igor Santos, and Pablo G. Bringas. 2017. The web is watching you: A comprehensive review of web-tracking techniques and countermeasures. *Logic Journal of the IGPL* 25 (2 2017), 18–29. Issue 1. https://doi.org/10.1093/JIGPAL/JZW041

[141] Sandra Siby, Umar Iqbal, Steven Englehardt, Zubair Shafiq, and Carmela Troncoso. 2022. WebGraph: Capturing Advertising and Tracking Information Flows for Robust Blocking. (2022). https://www.usenix.org/conference/usenixsecurity22/presentation/siby

[142] Mohit Singhal, Chen Ling, Pujan Paudel, Poojitha Thota, Nihal Kumarswamy, Gianluca Stringhini, and Shirin Nilizadeh. 2023. SoK: Content Moderation in Social Media, from Guidelines to Enforcement, and Research to Practice. *Proceedings - 8th IEEE European Symposium on Security and Privacy, Euro S and P 2023* (2023), 868–895. https://doi.org/10.1109/EUROSP57164.2023.00056 arXiv:2206.14855

[143] Gilberto Atondo Siu, Alice Hutchings, Marie Vasek, and Tyler Moore. 2023. "Invest in crypto!": An analysis of investment scam advertisements found in Bitcointalk. (2023). https://doi.org/10.1109/eCrime57793.2022.10142100

[144] Alexander Sjosten, Daniel Hedin, and Andrei Sabelfeld. 2021. EssentialFP: Exposing the Essence of Browser Fingerprinting. *Proceedings - 2021 IEEE European Symposium on Security and Privacy Workshops, Euro S and PW 2021* (9 2021), 32–48. https://doi.org/10.1109/EUROSPW54576.2021.00011

[145] Michael Smith, Peter Snyder, Moritz Haller, Benjamin Livshits, Deian Stefan, and Hamed Haddadi. 2022. Blocked or Broken? Automatically Detecting When Privacy Interventions Break Websites. (mar 2022). arXiv:2203.03528v2 http://arxiv.org/abs/2203.03528

[146] Sara M. Smyth and Rebecca Carleton. 2012. Measuring the Extent of Cyber-Fraud: A Discussion Paper on Potential Methods and Data Sources. *SSRN Electronic Journal* August (2012). https://doi.org/10.2139/ssrn.2020637

[147] Aditya K. Sood and Richard J. Enbody. 2011. Malvertising – exploiting web advertising. *Computer Fraud & Security* 2011 (4 2011), 11–16. Issue 4. https://doi.org/10.1016/S1361-3723(11)70041-0

[148] Blase Ur, Pedro Giovanni Leon, Lorrie Faith Cranor, Richard Shay, and Yang Wang. 2012. Smart, Useful, Scary, Creepy: Perceptions of Online Behavioral Advertising. (2012). http://www.aboutads.info/choices/

[149] Mahalakshmi V. and Thaiyalnayaki M. 2023. Impact Of Harmful Advertisement onChanging BehaviorofChildren –A Study withReference toChennai City. (2023). http://sifisheriessciences.com/journal/index.php/journal/article/view/1005/995

[150] Yash Vekaria, Rishab Nithyanand, and Zubair Shafiq. 2022. The Inventory is Dark and Full of Misinformation: Understanding the Abuse of Ad Inventory Pooling in the Ad-Tech Supply Chain. (10 2022). http://arxiv.org/abs/2210.06654

[151] Yash Vekaria, Rishab Nithyanand, and Zubair Shafiq. 2024. Turning the Tide on Dark Pools? Towards Multi-Stakeholder Vulnerability Notifications in the Ad-Tech Supply Chain. (2024).

[152] David C. Vladeck. 2000. Truth and consequences: The perils of half-truths and unsubstantiated health claims for dietary supplements. *Journal of Public Policy and Marketing* 19, 1 (2000), 132–138. https://doi.org/10.1509/JPPM.19.1.132.16948/FORMAT/EPUB

[153] Joel Weidenmark. 2020. Acceptable Ads guidelines, its effect on user experience and ad-noticeability. *DEGREE PROJECT COMPUTER SCIENCE AND ENGINEERING* (2020).

[154] Bartosz W. Wojdynski. 2016. The Deceptiveness of Sponsored News Articles: How Readers Recognize and Perceive Native Advertising. *American Behavioral Scientist* 60, 12 (nov 2016), 1475–1491. https://doi.org/10.1177/0002764216660140

[155] Claire Wonjeong Jo, Miki Wesołowska, and Magdalena Wojcieszak. 2024. A Taxonomy of Online Harm and MLLMs as Alternative Annotators. (2024).

[156] Apostolis Zarras, Alexandros Kapravelos, Gianluca Stringhini, Thorsten Holz, Christopher Kruegel, and Giovanni Vigna. 2014. The Dark Alleys of Madison Avenue: Understanding Malicious Advertisements. (2014). https://doi.org/10.1145/2663716.2663719

[157] Stan Zdonik, Shashi Shekhar, Jonathan Katz, Lakhmi C Jain, David Padua, V S Subrahmanian, and Newton Lee. 2017. Fraud Prevention in Online Digital Advertising. *SPRINGER BRIEFS IN COMPUTER SCIENCE* (2017). http://www.springer.com/series/10028

[158] Eric Zeng, Paul G Allen, Rachel Mcamis, and Franziska Roesner. 2022. What Factors Affect Targeting and Bids in Online Advertising? A Field Measurement Study. *Proceedings of the 22nd ACM Internet Measurement Conference (IMC '22), October 25â•fi27, 2022, Nice, France* 1 (2022). https://doi.org/10.1145/3517745.3561460

[159] Eric Zeng, Paul G Allen, and Franziska Roesner. 2021. What Makes a "Bad" Ad? User Perceptions of Problematic Online Advertising Tadayoshi Kohno. (2021). https://github.com/eric-zeng/chi-bad-ads-data

[160] Shitong Zhu, Xunchao Hu, Zhiyun Qian, Zubair Shafiq, and Heng Yin. 2018. Measuring and Disrupting Anti-Adblockers Using Differential Execution Analysis. *25th Annual Network and Distributed System Security Symposium, NDSS 2018* (2018). https://doi.org/10.14722/NDSS.2018.23331

A Taxonomy: Term Analysis and Clustering

This section provides a snippet of how the outputs from the term and clustering analysis were arranged. Each term is assigned a maximum cosine similarity score based on its contextual relevance across document sections. Terms are categorized into clusters with assignments of either *Accepted* or *Rejected*, and a rationale is provided to explain the decision.

Table 6 demonstrates this process. For example, terms such as *privacy* and *tracking* received high similarity scores, indicating strong contextual relevance to categories like data privacy. Conversely, terms like *greenwashing* were initially *Rejected* due to their low contextual relevance (similarity score of 0.14). However, through manual analysis, the term *greenwashing* was later added to the taxonomy when environmental claims were included as a relevant subcategory.

The iterative refinement of the taxonomy highlights the importance of both automated methods and manual analysis in ensuring comprehensive coverage. While terms with low scores like *greenwashing* were initially excluded, manual review identified their importance in the context of environmental claims, leading to their inclusion in the final taxonomy.

Table 7 shows the complete taxonomy with categories, definition, keywords and the negative nad positive examples. Positive examples represent the problematic that would classifies in this category and negative examples represent the benign ads that could be confused with this category. This taxonomy is used in the prompt (refer to Figure 6) as well.

B Prompt Engineering

We prompt the LLM with numerous prompts with varied modifications to arrive at the final prompt. The LLM was asked to provide explanations for each of the classifications which helped us to reason about the misclassifications and change the prompt accordingly. The temperature was set to 0 to avoid variable result. The LLM outputs were generated three times to check for variation in the result but none were found.

The prompt used is shown in Figure 6.

Table 6: Snippet of term analysis and clustering results.

Term	Max Cosine Similarity	Cluster Assignment	Reason
privacy	0.82	Accepted	Strongly associated with privacy concerns and tracking issues.
tracking	0.79	Accepted	Frequently appears in discussions about data privacy and behavioral targeting.
misleading	0.75	Accepted	Central to categories like health claims and financial fraud.
cookies	0.70	Accepted	Relevant to tracking and data collection themes.
deceptive	0.67	Accepted	Key term in identifying manipulative ads and misleading content.
sustainability	0.15	Rejected	Low similarity with other terms; unclear contextual relevance.
clickbait	0.61	Accepted	Related to dark patterns and manipulative design practices.
greenwashing	0.14	Rejected	Rarely linked to other terms; requires refinement or additional evidence.
subscriptions	0.68	Accepted	Strongly tied to fraudulent schemes like subscription fraud.
explicit	0.52	Accepted	Frequently linked to inappropriate or offensive content.
...	...	...	...

LLM Prompt for Ad Classification

```
messages = [{
  {"role": "user",
    "content": [{
      {"type": "text",
        "text": """Analyze the attached ad and determine if it is problematic or non-problematic strictly based on the taxonomy. If problematic, return the category title you find it to be belonging to as the key value. Otherwise return an empty string as the key value. Analyse all possible sub ad images present as well."""
      }, {
      "type": "text",
        "text": f"{ocr_text}"
      }, {
      "type": "image_url",
        "image_url": {
          "url": f"data:image/png;base64,{base64_image}"
        }
      }
    ]}, {
  "role": "system",
    "content": """You are an expert in evaluating online display advertisements based on a taxonomy derived from ad exchange policies and user perception studies. Analyze the provided advertisement screenshot to determine if it is problematic, strictly using the provided taxonomy. For the ad images, you should also identify advertiser information. For Taboola and Outbrain ads, consider the advertiser as the publisher’s name.
    ##
    The text present (OCR output) in the ad creative is also provided as an additional input for processing and identifying any problematic behaviour.
    ##
    The taxonomy includes a title, definition, keywords, and notes for each category. Strictly use the definitions, keywords, and notes as primary guides. The keywords are exhaustive and represent the themes covered in each category.
    ##
    Be conservative in your classification: Only classify if the ad clearly and confidently fits a category. If in doubt, classify it as non-problematic.
    ##
    Since it is a multi-label classification, multiple categories can be chosen. Return a JSON object with the following structure:
    1. A categories key containing a list of applicable categories. If non-problematic, return an empty list as the value.
    2. An explanations key providing a rationale for each identified category. Include detailed reasoning for each flagged category.
    3. An advertiser key containing the brand name identified in the ad image. If multiple ads exist, report ’multiple advertisers’. If no advertiser identification is possible or no subtitle is present, report *None* in the value.
    ##
    For screenshots with multiple ads, return the superset of categories covering all ads.
    ###
    Taxonomy:
    1. Regulations
      Definition: Ads deemed inappropriate for younger audiences (Below 18 years) due to product or content nature.
      Keywords: sexually suggestive, Cannabinoids, alcohol, gambling/sweepstakes, cosmetic interventions, hazardous goods, drug use, Weight-reduction regimes or establishments, Lottery, E-Cigarettes, Prescription-only medicines, Dating sites
      #
      Note:
        - Classify only if clear evidence like explicit exposure, provocative poses, or overtly suggestive context exists; avoid bias towards modeling images.
        - Sexually suggestive ads should be classified into both Regulations and Inappropriate or Offensive Content if the content is derogatory or explicit.
    ##
    Other categories mentioned in a similar fashion ...
    """
  } ]
}]
```

Figure 6: Prompt used for LLM evaluation of advertisements

Table 7: The taxonomy for identifying and classifying problematic ads on the web is structured into four key columns: Category (Broad themes identified in the literature), Definitions, Set of Keywords (to aid in classification), and Positive/Negative Examples (to clarify the types of ads included in each category). Certain categories are further divided into subcategories to enhance clarity and broaden the scope of classification, ensuring a more nuanced understanding of problematic ad content.

Category	Definition	Keywords	Examples
Regulations	Age-Based (Inappropriate for below 18 audiences) - Ads targeting or featuring content deemed inappropriate for younger audiences or sensitive groups due to the nature of the product or service.	sexually suggestive, Cannabinoids, alcohol, gambling/sweepstakes, cosmetic interventions, hazardous goods, drug use, Weight-reduction regimes or establishments, Lottery, E-Cigarettes, Prescription-only medicines, Dating sites	Positive: Ads promoting alcohol, gambling, or cosmetic surgery services, especially if displayed on platforms accessible to minors. Negative: Ads for non-alcoholic beer branded for a general audience.
	Geographical (might be offensive in certain geographies) - Ads targeting or featuring content deemed inappropriate for specific demographies due to local laws and regulations	pharmaceuticals, prescription drugs, tobacco, weapons, explosives, illegal products, non-compliant content	Positive: Ads for prescription drugs are displayed in countries where such ads are illegal Negative: Ads for over-the-counter medicines that comply with local advertising laws
Inappropriate or Offensive Content	Ads containing language, visuals, or themes that may be offensive, graphic, or disrespectful towards certain individuals or communities	Sexually explicit, offensive language, violent acts, hate speech, hookup, graphic images, racially insensitive, conspiracy theories, disrespectful religious/sacred content/profanity, trafficking, social issues, hacking	Positive: Ads using racial stereotypes or derogatory language targeting specific communities. Negative: Ads that are not targeted at specific communities or individuals.

Category	Definition	Keywords	Examples
Deceptive Claims and Exaggerated Benefits	Health Claims - Ads that make unverified or exaggerated health claims about a product’s or service’s effectiveness, often intending to mislead the consumer	No disclosure, Reasons for conditions, miracle cure, weight loss, scientifically proven, Dietary supplements, cosmetic beauty treatments, mental health services, false vaccine information, cheap substitutes	Positive: Ads claiming "miracle cure for diabetes" without evidence Negative: Ads stating "may help reduce symptoms" with disclosures
	Financial Claims - Ads that make unverified or exaggerated financial claims about a product’s or service’s effectiveness, often intending to mislead the consumer	get rich quick, financial freedom, investment returns, guaranteed profits, False Tax Promises, Crypto Gain Misrepresentation, debt relief	Positive: Ads claiming "get rich quick" with no supporting evidence Negative: Ads for financial services with clear disclosures like "results may vary"
	Environmental and ethical claims - Ads that make unverified or exaggerated environmental claims about a product’s or service’s effectiveness, often intending to mislead the consumer	No disclosure, greenwashing (eco-friendly, sustainable, green product, carbon-neutral, environmentally safe), eco-friendly, ethical sourcing, organic, waste reduction claims	Positive: Ads claiming "100% sustainable" without verification Negative: Ads mentioning "supports sustainability efforts" with clear disclosures
	Other Impossible claims - Ads that make unverified or exaggerated claims about a product’s or service’s effectiveness, often intending to mislead the consumer that doesn’t fall in the above categories	overpromising, instant results, transform your life, best in the world/market, one-of-a-kind, never-before-seen, guaranteed satisfaction, exclusive deal, legal claims	Positive: Ads promoting a product as "the only one of its kind" without credible proof. Claims of "100% customer satisfaction" without valid supporting data Negative: Claims may be permissible if supported by verifiable evidence, appropriate disclosures, or regulatory/legal endorsements
Dark patterns and manipulative design	Ads that use deceptive design techniques to manipulate user behavior, such as clicking, subscribing, or sharing information unintentionally	clickbait, social engineering, scarcity tactics, confirmshaming, countdown timers, fake buttons, sensationalism, fake testimonials, fake celebrity endorsements, urgency, last chance, your data is at risk, fear tactics, emergency, danger, don’t miss out, Incomplete sentences using …	Positive: Ads with countdown timers suggesting artificial urgency to push users into impulsive purchases Negative: Ads using urgency (e.g., "Limited time only!") but labeled as promotional content.
User Experience Disruption	Ads that degrade the user experience by being overly intrusive, annoying, or difficult to navigate	annoying, intrusive (revealing locations), auto-playing video, difficult-to-close pop-ups, disruptive ad formats, ad loading speed, unclear labeling of sponsored content. Ad quality (image), No Advertiser Information (text/image) - for context	Positive: Ads that automatically play sound or video, interrupt user activity Negative: Ads with easily muted videos that don’t hinder browsing
Fraud and Scam Content	Ads promoting fraudulent schemes or products, often aimed at financially exploiting or deceiving users	subscription fraud, fake certificates, counterfeit currency, fake review manipulation, scam behavior, crypto scams, brand impersonation, unauthorized use, copyright violation, stolen images, Repetitive images	Positive: Ads offering “certificates” or “licenses” with no valid accreditation or proof of authenticity Negative: Ads for legitimate online courses offering verifiable certificates
Political Content and Propaganda	Ads related to political topics that could polarize or manipulate public opinion, especially during sensitive times like elections	No disclosure (endorsements), election campaigns, fake endorsements, fake news, propaganda (climate change, LGBTQ rights, racial justice, religious freedom, abortion, immigration policy), defaming candidates	Positive: Ads supporting specific political candidates or parties, especially if they use misinformation to sway opinion Negative: Public service announcements or verified campaigns promoting social awareness.

C Threshold Calibration for CLIP Model

We calibrated the per-category thresholds using sigmoid-transformed similarity scores, optimizing for precision (to reduce false positives). Table 8 summarizes the optimal thresholds for each ad category along with the corresponding precision scores.

Table 8: Threshold calibration for the CLIP baseline using precision on sigmoid-transformed raw scores.

Label	Optimal Threshold	Precision
Dark Patterns and Manipulative Design	0.57	0.3182
Deceptive claims and Exaggerated Benefits	0.57	0.2500
Inappropriate or Offensive Content	0.57	0.2500
Non-Problematic	0.57	0.7143
Political and Socially Sensitive Topics	0.56	0.1556
Regulations	0.56	0.0769
User Experience Disruption	0.58	1.0000

D Evaluation Metrics for LLM and CLIP Ad Classification

We evaluated the performance of four models– GPT-4o (with keywords), GPT-4o-mini (with and without keywords), and CLIP. Table 9 summarizes the binary classification performance for problematic ad content detection across these four models. Table 10 presents the precision, recall, and F1-scores for each ad category across these four models. These metrics show that LLM-based models like GPT-4o-mini with keywords outperform CLIP as well as LLMs without keywords.

Table 9: Binary classification metrics for automated problematic ad detection.

Model	Precision	Recall	F1 Score
GPT-4o-mini (with keywords)	**0.88**	0.84	**0.86**
GPT-4o (with keywords)	0.81	**0.88**	0.84
GPT-4o-mini (without keywords)	0.57	0.82	0.67
CLIP	0.63	0.31	0.41

Table 10: Evaluation Metrics for Ad Classification Models. For each category, precision (Prec.), recall (Rec.), and F1 score are reported for four models: GPT-4o-mini (with keywords), GPT-4o (with keywords), GPT-4o-mini (without keywords), and CLIP.

Label	**GPT-4o-mini (with keywords)**			**GPT-4o (with keywords)**			**GPT-4o-mini (No keywords)**			**CLIP**		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
Dark Patterns and Manipulative Design	0.82	0.70	**0.76**	0.77	0.72	0.74	0.68	0.53	0.60	0.31	0.20	0.25
Deceptive claims and Exaggerated Benefits	0.75	0.63	0.68	0.67	0.80	**0.73**	0.46	0.80	0.58	0.40	0.04	0.08
Inappropriate or Offensive Content	0.50	0.88	0.64	0.58	0.88	**0.70**	0.30	0.60	0.40	0.50	0.09	0.15
Non-Problematic	0.92	0.94	**0.93**	0.93	0.89	0.91	0.88	0.69	0.77	0.66	0.89	0.76
Political and Socially Sensitive Topics	0.65	1.00	**0.79**	0.65	1.00	**0.79**	0.46	1.00	0.63	0.10	0.27	0.14
Regulations	0.74	0.87	**0.80**	0.63	0.87	0.73	0.44	0.52	0.48	0.09	0.11	0.10
User Experience Disruption	0.94	0.73	**0.82**	0.78	0.66	0.71	0.22	0.52	0.31	0.00	0.00	0.00