AdLens: Efficient Detection of Deceptive Software Ads

Ritik Roongta
New York University
Brooklyn, NY, USA
ritik.r@nyu.edu

Marwan Adnan Darwish
Radboud University
Nijmegen, The Netherlands
marwan.darwishkhabbaz@ru.nl

Masoud Poorghaffar Aghdam
Radboud University
Nijmegen, The Netherlands
masoud.poorghaffaraghdam@ru.nl

Rachel Greenstadt
New York University
Brooklyn, NY, USA
greenstadt@nyu.edu

Gunes Acar
Radboud University
Nijmegen, The Netherlands
g.acar@cs.ru.nl

Abstract

Threat actors increasingly use online advertising as a delivery vector for malware, fraud and scams. A common tactic in these campaigns is to pressure users into installing software by inducing panic through false claims about device infections or by promoting software with unrealistic capabilities. While prior work has detected problematic ads through web scraping, those studies have had limited visibility into the broader ad ecosystem. In this study, we leverage ads available through Google's Ads Transparency Center to detect three categories of deceptive software ads at scale: *scareware* ads that use fake security threats, ads containing *misleading claims*, and ads with *misleading design* that conceal advertiser identity and intent. To efficiently detect problematic ads, we propose *AdLens*, a scalable, multilingual and modular detection pipeline. AdLens uses a two-stage approach for efficiency: a low-cost semantic similarity-based prioritization stage followed by a high-precision LLM-based classifier, based on open-weight models. Applying AdLens to a subset of software-related Google ads, we identify hundreds of deceptive software ads, which received millions of impressions. Beyond detecting deceptive software ads, AdLens can be adapted to identify other types of harmful ads, and support efficient exploratory analysis over ad corpora via embedding-based nearest-neighbor search. Our findings reveal significant gaps in ad moderation, exposing millions of users to deceptive and malicious ads. We provide practical tools, methods and datasets for researchers and policymakers to audit opaque ad ecosystems and improve accountability.

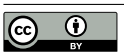
CCS Concepts

• **Security and privacy** → *Web application security*; *Privacy protections*; *Usability in security and privacy*.

Keywords

Advertisements, Malvertising, Scareware, Security, LLMs, Web Measurement

ACM Reference Format:
Ritik Roongta, Marwan Adnan Darwish, Masoud Poorghaffar Aghdam, Rachel Greenstadt, and Gunes Acar. 2026. AdLens: Efficient Detection of Deceptive Software Ads. In *Proceedings of the 2026 ACM SIGSAC Conference on Computer and Communications Security (CCS '26), November 15–19, 2026, The Hague, Netherlands.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3830454.3846781



ACM ISBN 979-8-4007-2871-6/2026/11
https://doi.org/10.1145/3830454.3846781

1 Introduction

The abuse of advertising networks to distribute malware, phishing lures, and other scams is well documented in academic literature and industry reporting. A Reuters investigation reported that internal Meta documents projected that ads for scams and banned goods accounted for approximately 10% of the company's 2024 revenue [38]. Google search ads have likewise been repeatedly abused to distribute malware, including campaigns impersonating popular software downloads [42, 68]. Reporting has further shown that ransomware access brokers used sponsored search results to advertise fake enterprise software as part of initial-access campaigns [5, 10]. Google's 2025 Ads Safety Report also indicates the scale of the problem: "Abusing the Ad Network" and "Misrepresentation" were among the largest policy-enforcement categories in 2025, accounting for roughly 1.7 billion blocked or removed ads in total [45]. These enforcement categories substantially overlap with deceptive or malicious software promotion [32].

Although prior work [42, 65, 68, 84, 109] has attempted to detect and measure harmful advertisements, collecting ads for systematic analysis remains difficult. Geofencing, cloaking, bot-detection mechanisms, and fine-grained targeting all hinder reproducible large-scale sampling. Because ad delivery is individualized and monitoring remains limited, harmful campaigns can persist for extended periods before they are identified [47, 83].

To address these limitations, we present a large-scale analysis of deceptive software advertisements using data available through Google's Ads Transparency Center (ATC) [30, 31]. Compared with scraping live webpages, the ATC offers three advantages. First, ads shown in the European Economic Area (EEA) countries[1] are labeled by topic, which allows us to isolate software and mobile-app advertisements. Second, it retains ads that are currently active or were active during the previous three years, enabling retrospective analysis of past campaigns alongside ongoing ones. Third, it provides advertiser metadata and impression information for ads shown in the EEA — data that are rarely available through live collection or AdChoices disclosures [9]. While several major platforms, including Meta and TikTok provide similar ad libraries, we

[1]EEA includes 27 European Union member states plus Iceland, Liechtenstein, and Norway.

focus on Google because it is the largest search and web advertising platform and has been exploited in multiple past malvertising campaigns [16, 42, 49, 92].

The ATC contains hundreds of millions of ads, which poses a challenge for scalable detection of deceptive software ads. To address this challenge, we propose ***AdLens***, an efficient, modular, and multilingual detection pipeline (refer to Fig 1) that can be adapted to other ad libraries and other classes of harmful advertising. AdLens uses a two-stage pipeline that combines inexpensive high-recall retrieval with high-precision verification. In the first stage, we use a 300M-parameter Gemma 3 embedding model [33] to encode OCR-extracted, translated ad text and rank ads by similarity to reference statements derived from manual analysis and Google policy documents. In the second stage, a VLM ensemble with Qwen3.5-9B [78] and Gemma3-12B [35] classifies candidate ads, while Gemma4-26B [36] adjudicates disagreements. AdLens achieves an F1 score of 0.922 while substantially reducing inference cost by reserving VLM analysis for a small, high-ranked subset of ads.

The input to AdLens detection pipeline is 188k ad creatives we collected by selectively scraping the ATC pages for ads from three topics - *Software*, *Mobile App Utilities*, and *Computer & Consumer Electronics*. We focused on these three topics as preliminary analysis showed Google consistently assigns them to ads promoting software and mobile-apps. Manual analysis of this corpus, guided by Google's ad policies [32], yielded three categories of problematic advertising practice we focus on in this work: scareware, deceptive claims, and misleading ad design. For each category, we identified recurring subthemes and developed reference statements that anchor the similarity-ranking stage of our pipeline. Specifically, reference statements serve as search queries that advertisements are compared against using semantic similarity.

Despite the public availability of ads on ATC, the massive scale of the dataset poses a challenge for researchers, practitioners, and policymakers who need an easy way to explore and search the data. Moreover, the ATC does not support text-based search, accepting only queries by advertiser name and website. To enable more robust search capabilities, we provide a vector database of texts extracted from ad creatives, and a search tool (Appendix B(f)) to efficiently identify semantically similar ads. The search tool, along with AdLens, can be extended to additional ad categories and provides a low-cost and efficient tool for exploratory ad analysis to researchers and policymakers.

Our contributions are as follows:

- We develop AdLens, a multilingual pipeline for efficiently identifying deceptive and malicious software advertisements.
- We identify thousands of deceptive or malicious ads, including campaigns that collectively reached millions of users.
- We build a fully automated system for identifying ads whose landing domains are classified as malicious by prominent protective DNS providers.
- We provide a vector database and search tool based on ad-text embeddings, enabling exploratory analysis of potentially harmful ads by researchers, regulators, and the public.

2 Background

2.1 Google Ads Transparency Center

To address growing transparency concerns and to satisfy regulations such as the *Digital Services Act* (DSA) [1], large platforms such as Google, Meta and TikTok have introduced searchable ad repositories [30, 62]. Among these, Google's Ads Transparency Center (ATC) provides a public repository of ads served across *surfaces* such as Google Search, Gmail, and YouTube, and supports filtering by advertiser or website name [30]. Initially limited to political ads, ATC has since expanded to non-political ads (2023 [31]). To our knowledge, Google's ATC has not yet been systematically used to study potentially deceptive and malicious software ads.

For EEA ads, Google provides a platform-level view via two public BigQuery datasets: `creative_stats` (detailing metadata and impressions for 174M+ active ads) and `removed_creative_stats` (tracking 10M anonymized, removed ads).

2.2 Embeddings

Embeddings map text or images to dense vectors that preserve semantic similarity [51, 85]. They expose signals not apparent in raw inputs and are widely used for retrieval, clustering, classification, and generation [13, 39, 61, 75, 76]. In ad analysis, they enable comparison with policy-violating prototypes without exact lexical overlap. AdLens uses embeddings for high-recall retrieval, narrowing large corpora before more expensive multimodal verification.

2.3 Malvertising

Malvertising uses advertising infrastructure to distribute malware, redirect users, or induce harmful actions through deception [56, 103]. Its scale remains substantial: in 2023 alone, Google reported removing or blocking over 206.5 million ads for misrepresentation, and over one billion ads for abusing their ad network, which includes promoting malware [44]. Independent monitoring entities also report persistent malvertising campaigns across the advertising ecosystem [6, 20].

Common attack patterns include forced redirects to malicious pages [6] and fake software-download ads impersonating tools such as Notepad++, 7-Zip, and PDF utilities [16, 68, 87]. These campaigns have been linked to malware such as IcedID, RedLine Stealer, and Vidar [4]. Sophos documented the TamperedChef campaign, which used Google Ads to distribute a trojanized PDF editor with delayed activation [91]. Related threats include scareware, ransomware, adware, and spyware delivered via ads or promoted software [5, 58, 72, 101, 104].

Detection remains challenging due to limited visibility into the ad supply chain. Malicious advertisers exploit targeting, cloaking, and rapid turnover, while observers see only a fraction of delivered ads [11, 43, 63]. Existing defenses such as ad blockers may reduce risks, but may not completely block malicious campaigns [52]. The ATC partially addresses this by publishing creatives and advertiser metadata directly, enabling campaign-level analysis without relying solely on user-side collection.

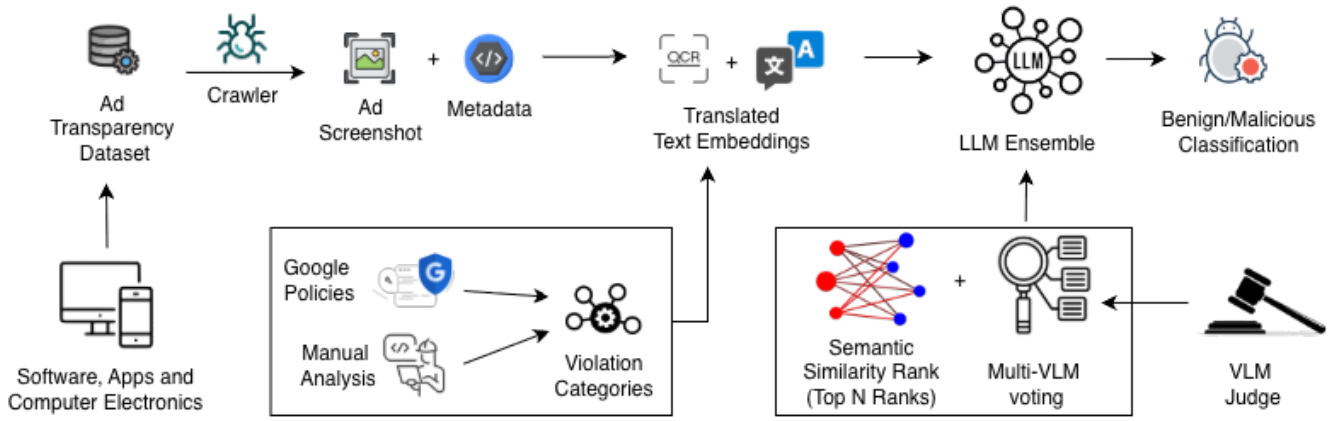


Figure 1: AdLens architecture: ad collection followed by the analysis pipeline. Translated ad text and taxonomy are embedded for semantic ranking, then passed to multi-VLM voting and judge verification.

2.4 Google Ad Policies

Google's advertising policies define the categories of content and conduct that its moderation systems seek to detect and enforce [32]. We review these policies and derive five policy-violation classes relevant to the software category, summarized in Appendix B(a).

We focus on violations observable from ad creatives and lightweight destination inspection. In particular, we study misrepresentation-related violations-including clickbait, misleading ad design, unreliable claims, and false impressions-as these are often visible in ad content. Other categories, including deceptive installation, covert surveillance, and enabling dishonest behavior, require post-installation analysis and are out of scope. As an auxiliary analysis, we use protective DNS providers [2, 3, 19, 77] to detect malicious landing pages linked from software ads.

3 Related Work

3.1 LLM/VLM-based Classification

Large language models (LLMs) and vision-language models (VLMs) support classification tasks requiring joint inferences over text and images, making them well suited to multimodal online advertisements. ScreenAI demonstrates strong performance on UI understanding tasks such as question answering and summarization [12]. In the advertising domain, Roongta et al. [84] show that a zero-shot GPT-4o-mini classifier can label problematic ads at scale with strong agreement with human annotators, suggesting model-based classification is effective when grounded in a clear taxonomy. Prior work highlights trade-offs between direct visual classification and reasoning-based pipelines. Contrastive VLMs such as CLIP, ALIGN, and SigLIP provide strong zero-shot baselines via image–text similarity [48, 80, 111], while prompt-adaptation methods (CoOp, CoCoOp) improve transfer with limited supervision [113]. Cooper et al. [21] show that additional language-model reasoning improves performance on tasks requiring contextual interpretation, and Bracha et al. [14] demonstrate gains from LLM-generated descriptions. LLM-as-a-judge approaches further resolve ambiguous cases, with strong models approximating human agreement [112], motivating cascade designs that reserve expensive models for uncertain inputs [102]. AdLens follows this paradigm by combining inexpensive retrieval with selective multimodal verification.

3.2 Semantic Similarity

Semantic similarity is widely used as a first-stage filter in retrieval and detection pipelines [55, 69, 71, 81, 100]. Varied content types like texts, images, etc. are embedded into a vector space and compared against reference examples, enabling efficient nearest-neighbor search at scale [80, 82]. This approach is effective for candidate generation, offering strong efficiency–accuracy trade-offs [53], and is widely used in production systems like Amazon and Meta for ranking and matching [40, 70]. However, embeddings alone may miss subtle intent or visual deception. AdLens therefore uses semantic similarity for coarse filtering and relies on multimodal models for final classification.

3.3 Deceptive Advertising

Prior work shows that ads often use deceptive design to manipulate user behavior [56, 65, 84, 108], including promotion of malicious software through misleading claims and visuals [10, 86, 93]. This establishes that ad creatives encode both persuasive intent and signals of harm.

Dark-pattern research provides a foundation: Mathur et al. [60] define such patterns as designs that coerce or deceive users. In ads, these include fabricated urgency, fake warnings, misleading buttons, and obscured disclosures. Empirical studies on health websites and political campaigns confirm prevalence across domains [108, 110], and recent work shows generative systems can reproduce such patterns [18].

Software-related deception is particularly important yet understudied. Scareware campaigns using fake security warnings are well documented [22, 67], but existing datasets and classifiers are often domain-specific or not designed for large-scale multilingual auditing. Foundational datasets such as the Pittsburgh Ads

Dataset [41] do not target software-specific deception in transparency archives. Recent taxonomies provide stronger grounding for detection [84, 105], while regulatory audits highlight persistent transparency gaps [26, 66]. AdLens builds on this work by focusing on deceptive software ads and operationalizing a taxonomy of scareware, deceptive claims, and misleading ad design.

3.4 Ad Library Studies

Recent work uses platform ad libraries to audit advertising ecosystems, primarily focusing on political ads. Leerssen et al. [54] frame ad archives as accountability mechanisms while noting limitations such as incomplete coverage and limited targeting transparency. Subsequent studies analyze political campaigns, delivery patterns, and disclosure gaps at scale [7, 17, 25].

Complementary work develops infrastructure for large-scale collection, such as AdDownloader for the Meta Ad Library [28]. However, prior work largely focuses on political or societal advertising and does not examine deceptive commercial ads such as software-related scams. Our work instead leverages Google's ATC to study deceptive software advertising at scale, extending ad-library research to user-facing harms tied to platform moderation.

4 Deceptive Ad Categories

We derive a taxonomy of deceptive software and app advertisements using a two-stage process. First, we review Google's advertising policies and retain those relevant to software ads that can be assessed from ad creatives, excluding categories requiring behavioral or post-installation analysis (e.g., network attacks or cryptomining) [32]. We also analyze landing pages when available, allowing us to include malware distribution policy violation as well. Relevant policy areas are summarized in Appendix B(a).

Second, we manually annotate ads from our crawled dataset (Section 5.1.1). Three annotators reviewed ~5,000 randomly sampled ads and identified recurring patterns such as urgency cues, unreliable claims, shocking content, and misleading interface elements. Combining policy review and annotation, we derive 10 themes grouped into three categories: *scareware*, *deceptive claims*, and *misleading ad design*. For scareware and deceptive claims, we construct reference statements (Table 1) as seeds for semantic similarity; these represent trivially false or implausible claims for software/apps, corroborated by user reviews from Trustpilot [99] and app stores[2].

Concretely, for each violating ad identified during manual annotation, we query the embedding space for similar creatives and draft a reference statement that covers the resulting cluster, rather than the individual ad; we provide the full annotation codebook alongside Google's policy mapping as supplementary material. A simpler, though coarser, alternative is to seed reference statements directly from violating ad text or known fraud-campaign language, which trades some annotation effort for potentially lower thematic coverage. For misleading ad design, we define button- and interface-based cues. To assess saturation, we annotate an additional 3,000 ads and observe no new themes.

Using this taxonomy, we identify ads based on their claims or interface cues and classify them into one of three categories. It is important to note that an ad may fall into multiple categories, as advertising campaigns often combine these techniques to evade detection and increase user engagement; for instance, ads with deceptive claims can also induce fear (Fig. 2i), overlapping with scareware.

Scareware. Ads that use fabricated threats or warnings to induce urgency, such as claims of device compromise, account suspension, or data exposure. We identify six themes (Table 1): account suspension, storage/performance, malware infection, data breach, legal threats, and privacy exposure. Examples include *Last warning: Your account will be suspended or restricted* and *Your phone is infected with viruses* (Fig. 2a–d).

Deceptive claims. Ads that present plausible but false claims to induce clicks or installs. We identify three themes: photo/data recovery bait, people tracking/surveillance, and social-curiosity bait. Examples include *You deleted 100 photos 3 years ago* and *Track anyone's location by entering their phone number* (Fig. 2h–j).

Misleading ad design. We define this category as ads that rely on low-information call-to-action (CTA) interfaces while obscuring advertiser identity, e.g., creatives dominated by prompts such as *Continue*, *Access Now*, or *Click here*, with no accompanying advertiser name or description (Fig. 2e–g,k). We exclude fake UI elements (e.g., simulated buttons or progress bars) and system-dialog or OS-notification mimicry, as these require scenario-specific reasoning that does not generalize reliably. Since low-information CTAs can also appear in legitimate ads, we require the absence of advertiser identity as an objective criterion. This category broadly aligns with Google's "Misleading Ad Design" policy theme (Appendix B(a)). Table 1 lists common CTA keywords.

5 Methods

AdLens operates in two phases (Fig. 1): Ad Collection and Ad Analysis. The ad collection phase gathers creatives and metadata from Google's ATC; the analysis phase classifies ads into the categories defined in Section 4. Crawling runs on DigitalOcean instances (16 cores, 32GB RAM), and multimodal inference uses NVIDIA L40S GPUs with 46GB VRAM. All VLMs use "temperature=0", "do_sample=False", and "enable_thinking=False" for determinism and efficiency.

5.1 Ad Collection

5.1.1 Crawled Ads. We constructed a corpus of 90,000 ad IDs from the `creative_stats` table, evenly sampled across the Software, Mobile App Utilities, and Computer & Consumer Electronics categories. From each topic, 30,000 ads (with distinct IDs) were selected to ensure balanced representation across topics. To capture recent advertising trends, one third of them i.e., 10,000 creatives per category were selected based on recency, using their `first_shown_date` values closest to the crawl date (March 26, 2026). The remaining 20,000 creatives were randomly sampled from the full dataset, spanning back to early 2023, to preserve diversity and ensure broader temporal coverage.

A pool of 90,000 ad IDs provides us with substantial coverage of the software ads. Moreover, the Software topic contains only about 32,000 available ad IDs, making 30,000 a near upper bound for balanced sampling and a natural choice for a large yet tractable dataset.

[2]https://play.google.com/store/apps, https://www.apple.com/app-store/

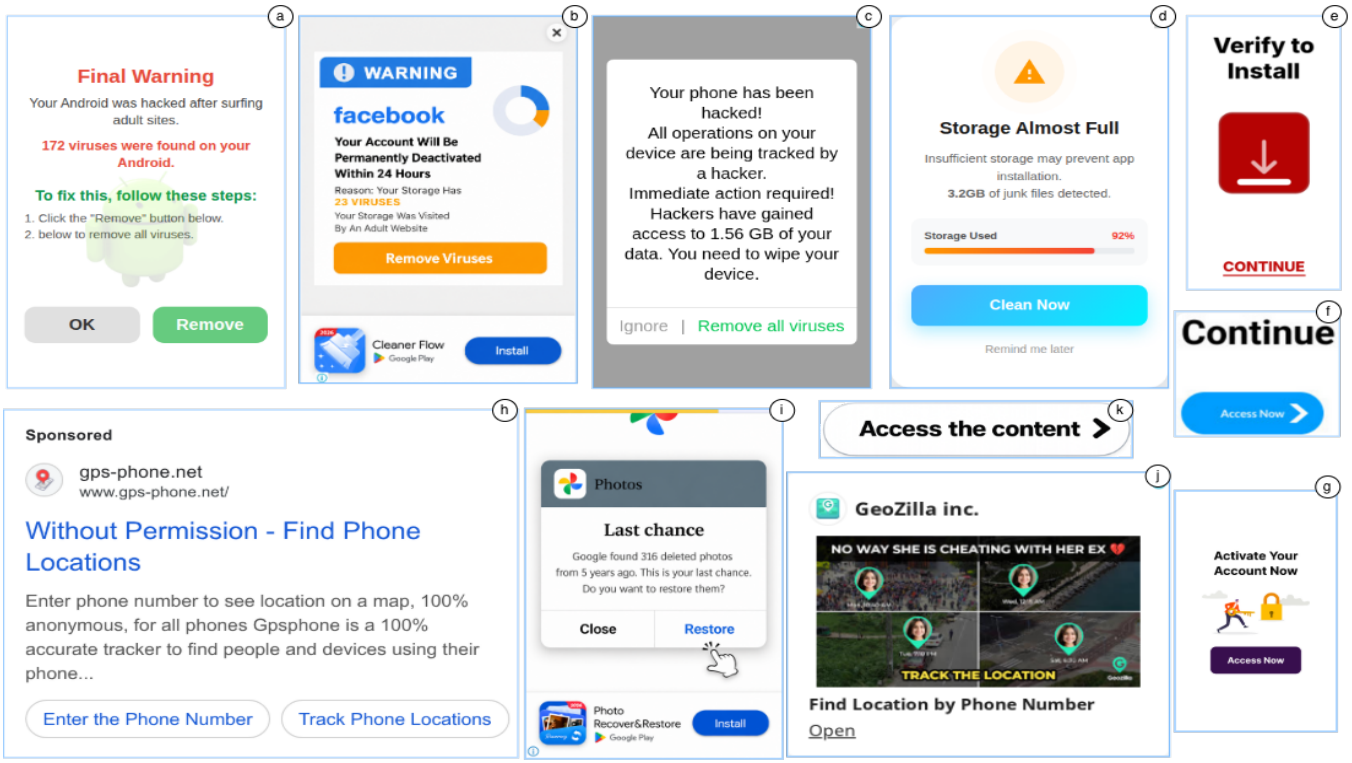


Figure 2: Examples of deceptive software ads identified by AdLens. (a,b,c,d) refer to scareware examples. (e,f,g,k) refer to ads with misleading design, and (h,i,j) refer to ads with deceptive claims.

We also constructed a `test_dataset` of 15,000 ad IDs with `first_shown_date` between March 26, 2026, and April 8, 2026. This dataset reduces the risk of evaluating on ads that informed our reference statements or prompt design, and offers a limited check on temporal stability by assessing whether similar violation patterns persist in newer ads.

5.1.2 Crawler. We implement a Puppeteer-based crawler [29] to extract ad creatives and metadata from ATC webpages. An *ad creative* is a captured ad screenshot, while an *ad ID* is the underlying identifier; each ad may have multiple variants, all of which are collected as different ad creatives.

Given an ad ID and its corresponding advertiser ID from the BigQuery database, our crawler constructs the ATC URL, loads it in a headless Chrome browser, and extracts the ad format, creative screenshots, landing URLs (when available), and additional ads linked to the advertiser [29]. Because multiple creatives for a given ad ID may be preloaded but hidden, the crawler reveals them through DOM overrides before capture.

When ads are clickable, the crawler also retrieves destination URLs. In many cases, the rendered creative does not expose this information, but a linked text-format version on the page reveals additional routing details. We therefore parse both the rendered view and any available text-format content.

For iframe-based creatives, particularly HTML and video ads, we extract structured data such as app store or YouTube links using pattern matching and recover embedded metadata via regular expressions. Screenshots of ad creatives are further processed with OCR and translation, then combined with extracted metadata for downstream analysis.

5.1.3 Crawled Dataset. Using the crawler, we scrape the ATC pages of 90k ads in March 2026, obtaining 188k ad creatives. These 188k creatives and the metadata scraped from 90k ads constitute the main input to our classification pipeline.

5.1.4 OCR and Translation. We use PaddleOCR (PP-OCRv5) for text extraction on 188k ad creatives due to its efficiency and multilingual robustness [23]. We use the English-oriented recognition model due to a large presence of Latin languages, since the English model also recognizes many Latin scripts. In a separate analysis using PaddleOCR-VL [74], we find that fewer than 1.5% of ads contain non-Latin scripts such as Russian, Greek and Arabic, which cannot be recognized by our OCR model. Among the text extracted from ad creatives, English (39.63%), German (13.46%), French (11.01%), Spanish (5.63%), and Italian (5.04%) are most common.

We translate all text into English using TranslateGemma 4B [34], which preserves original English inputs while normalizing multilingual content by correcting minor typos and adding appropriate whitespace.

We further apply lightweight filtering by deduplicating creatives with identical OCR text, yielding ~112,000 unique ad creatives.

Table 1: Table showing 10 different themes grouped into three major categories. Reference statements for scareware and deceptive claims are used for semantic-similarity ranking.

Category	Theme	Reference Statements
Scareware	Account Deactivation / Suspension	Last warning: Your account will be suspended or restricted. Your account will be permanently disabled because it does not comply with our rules. Your account has been flagged for unusual login attempts. Verify immediately or lose access. Action required: Your account will be deleted within 24 hours unless you confirm your identity.
	Storage & Device Performance	Your device storage is full. Your phone is running slow because of junk files. Your software/pdf version is outdated. Your battery is draining fast because of background apps.
	Virus / Malware Infection	Your phone is infected with viruses. Your device is severely compromised because you visited adult websites. Your browser is damaged by Trojan horses downloaded while visiting illegal websites. Warning: 13 viruses have been detected on your device. Clean now. Your phone has been compromised by spyware. Your camera and microphone may be recording.
	Hacker / Data Breach	Your device has been hacked and your personal information is at risk. Someone is monitoring your browsing activity right now. Your passwords have been exposed in a data breach. Change them immediately.
	Legal / Law Enforcement	Your IP address has been flagged for illegal activity. Law enforcement has been notified about suspicious activity from your device.
	Privacy Exposure	Your personal photos are visible to everyone on the internet. Your location is being tracked. Disable tracking immediately. Your browsing history is publicly accessible. Hide it now.
Deceptive Claims	Photo / Data Recovery Bait	You deleted 100 photos 3 years ago. Recover them. Photos deleted 4 years ago were found in memory. Recover them now. We found 347 recoverable files on your device. Restore them.
	People Tracking / Surveillance	Track anyone's location by entering their phone number. Find out who is calling you from unknown numbers. Monitor your partner's private messages and location. Track spirits or spells around you.
	Social Curiosity Bait	Install this app to find out who secretly hates you. Install this app to find out who has a secret crush on you. Find out what your friends really think about you.
Misleading Ad Design	Call to Action Baits	Access Now, Continue, Click here, etc.

5.2 Ad Analysis

The ad analysis pipeline has three stages: embedding-based ranking, ensemble classification, and LLM-based adjudication of disagreements between ensemble models. We evaluate performance using manually annotated golden datasets.

5.2.1 Golden Datasets. We construct two datasets. The *golden main dataset* contains 1,283 ads (with per-category true positive / true negative splits of 198/202 for scareware, 225/224 for misleading content, and 239/195 for ad design) from the crawled dataset. The *golden test dataset* contains 253 ads from the temporally distinct test_dataset, where we collected 50 true positive and true negative ads per violation. The ads used to derive the taxonomy and reference statements (Section 4), the golden main dataset, and the golden test dataset are drawn from entirely separate, non-overlapping ad pools, so none of the evaluation sets informed the reference statements or prompts used to classify them. Unfortunately, our manual analysis only yielded three true positives for misleading ad design violations. Although small, this golden_test_dataset remains useful for evaluating true-negative performance, which is essential for ad moderation platforms to maintain advertiser trust.

Three expert annotators labeled the data after reviewing 5,000+ ads. Inter-annotator agreement on 300 samples using Krippendorff's alpha with Jaccard distance [50] yields a score of $\alpha = 0.91$, with individual topic scores being 0.84, 0.92, and 1.00 for computer, mobile, and software, respectively. The alpha scores indicate strong agreement overall.

5.2.2 Embedding-Based Ranking. For scareware and deceptive claim ads, we use EmbeddingGemma [33], a 300M-parameter text embedding model designed for retrieval and semantic-similarity tasks, making it suitable for this high-recall filtering stage. We embed OCR text from the ads and rank them based on their semantic similarity to the embeddings of the reference statements from Table 1. Ads with very short text (fewer than three words) are excluded from the analysis in these categories to reduce noise, as even a single overlapping word can disproportionately inflate similarity scores.

For misleading ad design, we apply rule-based filtering on the entire crawled dataset and retain short-text ads, because longer text-heavy ads rarely exhibited this violation pattern in our manual review, giving us a pool of 689 ads. Further, we match their OCR

text against 18 call-to-action keywords (after lemmatization and normalization). Minimal-text ads with interface-like structure are also retained.

To aid researchers and policymakers, we additionally provide a semantic search tool for retrieving similar ads from the embedding space using text-based queries. A screenshot of the search tool is given in Appendix B(f).

5.2.3 Ensemble Classification. Candidate ads are passed to a VLM ensemble consisting of Qwen3.5-9B and Gemma3-12B [35, 78]. We also evaluated GLM-4.6V-Flash-9B during model selection [107]. All models operate in a similar parameter range and support multimodal inference over image-text inputs. Based on labeled data performance, we selected Qwen3.5-9B and Gemma3-12B for their balance of precision and recall. We also prioritized diversity across model families developed in different geographic and institutional contexts to mitigate regional and linguistic blind spots. Agreement between the two models is treated as a high-confidence signal, while disagreements are escalated to a larger judge model.

Classification is zero-shot prompt-based and grounded in reference statements for each violation type. When both models agree, their prediction is accepted; otherwise, the case is escalated. This design limits expensive inference to ambiguous cases while maintaining throughput.

We evaluate classifier efficacy on the golden main and golden test datasets. Table 2(a) reports precision, recall, and F1 for the three VLMs. Qwen3.5-9B achieves the highest macro F1 (0.917), followed by Gemma3-12B (0.821) and GLM-4.6V-Flash-9B (0.680). Performance varies by category: GLM slightly outperforms others on deceptive claims, Qwen leads on scareware and misleading design, and Gemma achieves the highest recall on deceptive claims. However, GLM performs poorly on a misleading design (recall 0.109, F1 0.195). These results motivate our ensemble choice: Qwen provides strong overall performance, while Gemma recovers cases Qwen may miss.

5.2.4 LLM as a Judge. To resolve classifier disagreements, we use Gemma4-26B as a judge model [36]. As AdLens is an open-source pipeline, we restrict choices to open-weight models. The judge receives the original prompt along with both models' outputs and reasoning, and produces a final decision. A larger model is used here due to the stronger comparative reasoning required.

We also evaluated Qwen3.5-27B and Mistral-Small3.2-24B [64, 79] as alternative judges, selecting Gemma4-26B based on performance on the golden datasets (Section 5.2.1). Table 2(b) reports results on disagreement cases. Gemma4-26B achieves the highest macro F1 (0.922), outperforming Qwen3.5-27B (0.877) and Mistral-Small3.2-24B (0.865). Qwen performs slightly better on scareware, and Qwen and Gemma tie on misleading design, but Gemma is strongest on deceptive claims.

Disagreement rates between ensemble models are low for scareware (8.5%) and deceptive claims (11.1%), reflecting strong agreement on clear violations. Misleading design is higher (29.0%) due to its subjective nature since whether a creative presents sufficiently clear advertiser information is not always unambiguous. On the full AdLens run, disagreement increases to 30.8% (scareware: 12.6%, deceptive claims: 31.5%, misleading design: 41.6%), consistent with more diverse data. Despite this, judge performance remains strong across categories, confirming robustness under elevated disagreement.

Disagreement between ensemble models is low for scareware (8.5%) and deceptive claims (11.1%), indicating strong agreement on clear violations. It is higher for misleading design (29.0%), likely because these creatives contain little textual information (§5), leaving insufficient evidence to determine whether advertiser information is adequately disclosed. On the full AdLens run, disagreement increases (30.8%; scareware: 12.6%, deceptive claims: 31.5%, misleading design: 41.6%), reflecting greater data diversity. Despite this, the judge model maintains strong performance across all categories (Table 2(b)), demonstrating the robustness of AdLens's disagreement-escalation design.

We also evaluate the full pipeline on the temporally distinct golden test dataset to reduce overfitting risk. As shown in Table 2(c), performance remains strong across categories. Results on misleading design are nearly perfect, but should be interpreted cautiously due to class imbalance and a few positive examples. This remains informative, as false positives are especially harmful in ad moderation as they erode advertiser trust. Performance on scareware and deceptive claims in the golden test dataset is consistent with the golden main dataset, indicating temporal stability over the short interval separating the two crawls.

5.2.5 Latency Analysis. We report inference latency for each component of the AdLens pipeline to assess scalability, a key constraint in ad moderation systems that process billions of ads daily (see Appendix B(b)). Without parallelism, the pipeline completes in a median of **4.7 seconds per ad** under a 15% disagreement rate from the golden dataset on a single GPU, which fits within typical asynchronous review windows.

Running the two ensemble classifiers in parallel on separate GPUs reduces latency to **3.8 seconds**, as the faster Gemma3 12B call (816 ms) is masked by the slower Qwen3.5 9B call (2.1 s). Processing four ads concurrently on a single GPU improves throughput to **one ad per 1.2 seconds**, a four times gain with no additional hardware. Combining both strategies achieves **under one second per ad** in effective throughput. Since AdLens processes only 11,000 semantically ranked ads rather than the full 188,000, total runtime is reduced to **about 14 hours** without parallelism, compared to **10 days** for full sequential processing, a 17 times improvement.

A text only variant of AdLens (i.e., one that only processes OCR'ed ad text and no images) further reduces latency from 4.7 seconds to **3.2 seconds**, a 32% reduction or about 1.5 seconds saved per ad, due to faster inference on shorter prompts (about 455 ms on Gemma3 12B and about 937 ms on the judge). However, it is less accurate for misleading ad design, which depends on visual signals. For tasks focused on the other categories, a text only pipeline using LLMs offers a more efficient alternative with a very slight accuracy tradeoff.

5.3 Auxiliary Analysis

We process the crawled dataset to quantify data completeness and characterize the advertising ecosystem represented in the crawl. Creative-level labels indicate what an ad claims or how it is visually

Table 2: Performance of AdLens pipeline stages. (a) ensemble candidates and (b) judge candidates, both on the 1,283-ad golden dataset; (c) the full pipeline on a previously unseen test set, with the last row giving overall performance. Best value per column and the chosen model at each step are in bold.

		Deceptive Claims			Scareware			Misleading Design			
Stage	**Model**	P	R	F1	P	R	F1	P	R	F1	**Macro F1**
Ensemble Classifiers (a)	**Qwen3.5:9b**	0.973	0.787	0.870	0.989	0.949	0.969	0.859	0.970	0.911	0.917
	Gemma3:12b	0.873	0.822	0.847	0.994	0.803	0.888	0.966	0.586	0.729	0.821
	GLM-4.6V-Flash:9b	0.994	0.791	0.881	0.930	1.000	0.964	0.963	0.109	0.195	0.680
Judge Models (b)	**Gemma4:26b**	0.897	0.929	0.912	0.967	0.935	0.951	0.861	0.946	0.902	0.922
	Qwen3.5:27b	0.628	0.964	0.761	0.939	1.000	0.969	0.861	0.946	0.902	0.877
	Mistral-small3.2:24b	0.800	0.714	0.755	0.968	0.968	0.968	0.897	0.848	0.872	0.865
Golden Test Dataset (c)	**Qwen3.5:9b**	0.978	0.938	0.957	0.873	0.960	0.914	1.000	1.000	1.000	0.957
	Gemma3:12b	1.000	0.958	0.979	0.978	0.880	0.926	1.000	1.000	1.000	0.968
	Gemma4:26b (judge)	0.980	1.000	0.990	1.000	1.000	1.000	1.000	1.000	1.000	**0.997**

presented, but they do not identify the entities behind those campaigns, how the ads are delivered and its reach. We therefore complement creative-level classification with analyzing key metadata fields, distribution of ad formats across platforms and profiling advertisers using the `creative_stats` table. In addition, we analyze advertisers and traffic routing by identifying frequently occurring advertisers and normalized destination domains, while explicitly tracking missing routing information.

This analysis is intended to characterize abusive advertisers and compare delivery patterns across malvertising categories (as defined in section 4). In particular, it allows us to identify recurring advertisers, compare advertiser behavior across categories, and distinguish broadly distributed campaigns from narrowly targeted ones. To do so, we aggregate advertiser identity and delivery signals, including disclosed names, locations, verification status, ad formats, serving surfaces, advertiser age, impression bounds, and geographic spread. We implement these aggregations in BigQuery and compute profiles separately for each category. Further, we also look into the impression count of the malicious ads to measure their reach. Google provides lower and upper bounds for ad impressions shown in the EEA and Turkey region for ads older than 90 days. We report the *midpoint* of the lower and upper bounds of the ads older than 90 days as a reasonable estimate to prevent over-counting impressions. The impact could be much higher in the real world, since these impressions are delayed and only pertain to views registered in Europe. For ads shown in multiple regions, we sum the impressions across regions to obtain an aggregate.

For the ads where landing page URLs are present, we query them using a quorum of PDNS providers to identify malicious URLs and identifying how many ads link to these overall irrespective of topics. We also look into App Store pages of the flagged ads wherever available through the transparency platform and flag low user ratings, critical user reviews. We also measure the download data to measure the reach of such malicious apps.

6 Results

Having established the efficiency and accuracy of AdLens, we then apply it to 188k ad creatives to estimate the prevalence of malicious ads and analyze the advertisers behind them. We further examine associated URLs to uncover finer signals.

6.1 Crawled Dataset Overview

Table 3 summarizes the 90k ad IDs crawled and analyzed across software, mobile, and computer topics. In total, we collect 188k ad creatives, which reduces to over 112k after exact OCR-text deduplication. The software category contains the largest number of ads (over 41k), all of which are image-based. Mobile app and software categories include all three ad types: image, text and video. Since the crawler takes screenshots of all ad creatives, it also processes text from initial frames of video ads. To contextualize our sampling relative to each topic's full advertiser population in `creative_stats`, the Software, Mobile App Utilities, and Computer & Consumer Electronics topics contain approximately 32K, 303K, and 6.5M ad IDs respectively; our crawl covers all available software ads (near-exhaustive, §5.1.1) but only a sample of the much larger mobile and computer populations A notable finding is the widespread absence of direct landing-page links in ads published on ATC. We found that across all categories, only 33% of ads have a target URL in their HTML attributes. Low prevalence of ad links limits analysis of potentially malicious destinations; thus, we also extract links through OCR to supplement our malicious domain analysis (§6.5).

Table 3: High-level stats for the crawled dataset. An *ad creative* is a screenshot of a rendered ad; an *ad ID* is the underlying transparency-dataset identifier. Dedup. creatives uses exact OCR-text match.

Metric	Computer	Mobile	Software	Total
Total creatives	54,211	61,191	73,432	188,834
Dedup. creatives	35,467	36,217	41,077	112,761
Ad type				
Image	16,527	16,558	41,077	74,162
Text	7,800	9,423	0	17,223
Video	11,140	10,236	0	21,376

6.2 Deceptive Ads: Prevalence

We use AdLens to estimate the prevalence of scareware, deceptive claims, and misleading ad design in the dataset of 188k ad creatives. Figure 3 shows, for each semantically ranked bin, the number of ads flagged by each classifier, their agreement, and the final violation rate after judge resolution. Semantically ranked bins correspond to ads most similar to the reference statements (Table 1). Because scareware is relatively rare, we compute its positive rate in bins of 100 over the top 1,000 semantically ranked ads; for deceptive claims, we use bins of 500 over the top 10,000 semantically ranked ads. These curves help determine stopping thresholds under varying compute budgets.

For scareware, the positive rate drops sharply after rank 200, falling below 20% by around rank 500, which indicates a strong concentration of violations among the top ranked ads (i.e., most similar to reference statements). In contrast, deceptive claims are more numerous and linguistically diverse; their positive rate declines more gradually and does not fall below 20% until approximately rank 8,500.

As shown in Figure 3, for deceptive claim ads, the Gemma4:26b judge line (Purple), which represents the ensemble verdict combined with violations identified during disagreements, initially tracks closely with the Gemma3-12b line (Blue). This suggests that the ensemble captures additional positive cases that were missed by the conservative Qwen3.5-9b model. However, as rank increases, it begins to align more closely with the Qwen3.5-9b line (Orange), indicating that some ads flagged as violations by the Gemma3-12b classifier are no longer considered violations by the ensemble.

Overall, this pattern suggests that higher semantically ranked ads are more likely to be true violations during disagreement resolution compared to lower ranked ones, further supporting the effectiveness of semantic similarity for ranking ads. It also highlights the robustness of the AdLens pipeline: it achieves high recall at the top ranks and captures the dense concentration of violating ads, while shifting toward higher precision at lower ranks which minimizes false positives.

Figure 4 reports violations with and without ad-ID-level deduplication. In the deduplicated setting, only one creative per ad ID, with the highest similarity score to the reference statement, contributes. For scareware, we detect 407 violating ads in the top 1,000 semantically ranked items, corresponding to 238 unique ad IDs. For deceptive claims, we detect 5,558 violating ads in the top 10,000, corresponding to 3,346 unique ad IDs. Misleading ad design yields 261 violating ads and 258 unique ad IDs after OCR-text-length filtering, highlighting lower bounds on the true prevalence of each category in our corpus.

To assess recall beyond our retrieval cutoffs, we analyze the rank distribution of labeled ads in the golden main dataset. True positives appear much higher in the similarity ranking than true negatives, with median ranks of 167 vs. 676 for scareware and 519 vs. 3,426 for deceptive claims (average ranks show the same trend). We further hand-labeled 300 ads sampled uniformly beyond our reporting cutoffs (i.e., after rank 1,000 for scareware and 10,000 for deceptive claims). All 300 scareware ads were benign, and only one deceptive-claims ad was a previously uncounted violation, suggesting that our cutoffs capture the vast majority of violations while requiring only a fraction of the computation needed to score the full corpus.

To further analyze these results, we examine behavior at the Ad ID level by measuring the share of associated creatives flagged for each violating Ad ID. Across all categories, creatives linked to a violating Ad ID are typically also flagged: 88.7% of scareware Ad IDs and 99.2% of misleading design Ad IDs have all associated creatives (i.e., variants) flagged, compared to 76.3% for deceptive claims.

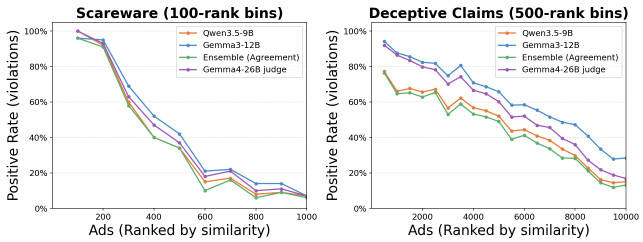


Figure 3: Prevalence of scareware and deceptive-claim ads across 1000 and 10000 ads most semantically similar to reference statements respectively. We see a steep decline for scareware ads beyond the 200 rank showing higher concentration while deceptive claim ads fall gradually. Ensemble represents a common verdict for violating ads between the Qwen3.5 and Gemma3 models. Also, Gemma4:26b judge line is always above the Ensemble since it represents the total violations after addressing the disagreements.

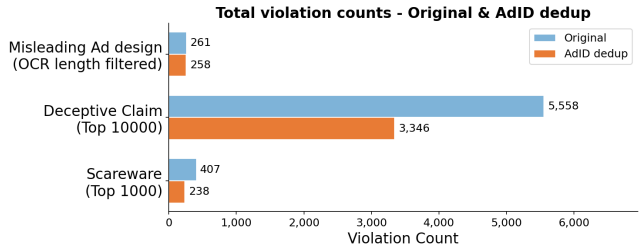


Figure 4: Detected violations with and without deduplication by ad ID for semantically ranked pool of scareware (top 1,000), deceptive claims (top 10,000), and misleading ad design after OCR-text-length filtering. Misleading Ad design has almost always only 1 ad creative per violating ad ID.

6.3 Deceptive Ads: Reach and Advertisers

We present advertiser- and advertisement-level characteristics across the three deceptive ad categories. For advertiser-level analysis, we select the top three advertisers based on the total number of detected ads.

Scareware. The detected 238 (unique ad ID) scareware ads account for more than 16 million impressions (in Europe alone) and come from 54 distinct advertisers. Two of the three most prolific scareware advertisers have more than a billion ad impressions across all of their ads (Table 4). All of top three advertisers have

been active for more than 6 months. While scareware ads are overwhelmingly in video format (99.1%), most only display a static image overlayed by text, as also noted by Breuer et al. [15]. A common scareware type is asserting that the user's device is infected with viruses. For example, an ad for Clean Sweep Plus Android app claimed (in French) that "your browser is 70% damaged, [...] while browsing adult websites". Another ad in Polish (1.1M impressions) stated "Your Android phone was hacked after browsing adult websites. 178 viruses were found on your Android phone."

Deceptive Claims. We detected 3,346 ads with deceptive claims, that received 89.9M impressions in total. These advertisements come from 279 distinct advertisers, with the most prolific advertiser contributing 771 (23%) of the 3,346 ads. The most seen ad with deceptive claims has been active for more than two years and registered 14.5M impressions. The ad promoted an app that can display anyone's GPS location by just entering their phone number. Another ad with 2.8M impressions displayed a (final) warning, claiming it was the user's last chance to restore 284 photos that had been deleted 7 years ago.

Misleading Ad Design. We detected 258 ads with misleading design from 17 distinct advertisers, with one advertiser contributing 169 (65.5%) of the 258 ads. One of the most viewed ads in this category was only shown in France and received 1.6M impressions. The ad had a very minimal design: it contained the text 'Continuer'[3] next to a green shield icon on the right (Appendix B(c)) (b) in the appendices). No identifying information was present about the advertiser or the promoted service. Our crawl metadata showed that clicking the ad opens *govulo.com*, for which several Trustpilot reviewers complained that they were misled by deceptive ads and pop-ups that impersonate other services, leading them to unknowingly enter payment details and sign up for a paid subscription service [97]. The most viewed ad with 2.1M impressions contained what looked like a QR code and text (in Dutch): Start now (Appendix B(c)). The ad contained a link to an unknown sport streaming subscription service (*megastore-online.co*), but the QR code could not be decoded and the ad contained no information on what the promoted product or service was.

Temporal Distribution. Figure 5 presents the temporal distribution of deceptive ads over a three-year period. A noticeable peak occurs around March 2026, largely due to dataset composition, as roughly one-third of the ads were sampled from the most recent ads (§5.1.1). Additional peaks in October 2024 and March 2023 suggest bursts of deceptive ad campaigns during those periods. This pattern aligns with the broader ad trends observed across 90k ad IDs, where overall ad volume is higher during these same intervals.

To assess the ongoing impact on users, we also measure how many of these ads were shown within the past two months, using a February 2026 cutoff for last_seen_date. The cumulative prevalence of these ads (across all categories) is depicted as a line in Fig 5. We observe that prevalence declines rapidly for scareware and misleading ad design, approaching near zero within a year. In contrast, deceptive claims exhibit a longer persistence, indicating that some ads evade detection and remain active over extended periods.

[3]Translates to "Continue", in French.

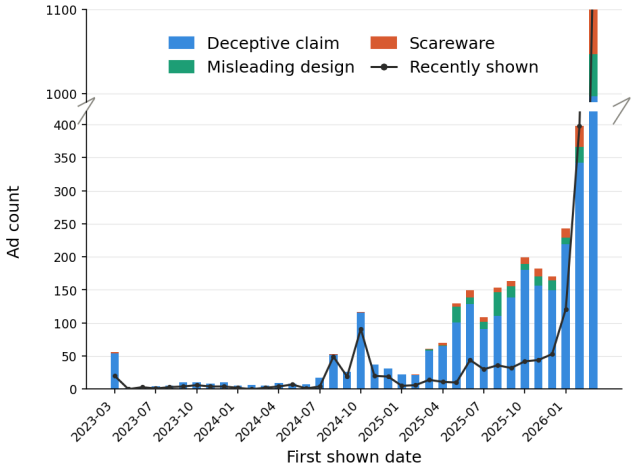


Figure 5: Distribution of deceptive ads by first shown date. The overlaid line counts ads (all categories) still shown on or after February 2026.

6.4 Apps Promoted by Deceptive Ads

We reviewed Google Play and Apple App Store pages linked to violating ads. We identified 198 Ad IDs with app store links, corresponding to 147 unique apps (139 on Google Play and eight on the Apple App Store). Of these, 34 apps (23%) had already been removed despite their ads remaining active, suggesting ad networks lag behind app stores in delisting. Notably, all remaining apps had been updated within the past year.

To classify Google Play apps, we applied a rule-based keyword heuristic on app titles and descriptions. The dataset consists of categories including file recovery apps (42, 40%), phone trackers (35, 33%), phone cleaner/optimization tools (15, 14.2%), and PDF/document readers (6, 5.7%). Among 105 apps with publisher data, we observe 100 unique developer IDs, indicating no publisher with many apps. Among the 105 apps with install data, the most common tier is 1M+ installs (30 apps, 20.4%), followed by 500K+ (21 apps) and 10M+ (19 apps); one app exceeds 100M+. Collectively, installations reach hundreds of millions, indicating widespread adoption of these apps. The mean app rating is 3.95 and the median 4.10. However, user reviews reveal consistent issues aligned with our findings. Reports frequently describe subscription traps (e.g., "*charged $29.99 from debit card without authorization*", "*charging my account*" after uninstalling), misleading claims ("*misleading advertisement that you can track a number by simply entering it – FALSE*"). One reviewer summarized: "*Deceptive Advertising / Demo ZERO STARS ... you MUST INVITE people to track their phone number.*" While anecdotal, these reviews validate our classifications and highlight that qualitative user feedback captures harms that aggregate ratings might obscure.

6.5 Ad Landing Page URLs

In order to detect whether ads in our dataset contain links to malicious domains, we analyze ad landing page URLs. For ads linking to Google's redirecting domains (e.g., googleadservices.com), we

Table 4: Top #3 advertisers per violation category ranked based on violating ads. 'Days' refer to the number of days for which at least one ad of the advertiser was active. EEA countries are counted as a single region in the 'Regions' column. Abbr.: Viol. - Violating, Impr. - Impressions.

Category	Advertiser	Loc.	Viol. Ads (#)	Viol. Impr.	Total Impr.	Regions	Days
Scareware	WELLRESION LIMITED	HK	90	10.96M	1.22B	8	617
	FEMOB TECHNOLOGY LIMITED	HK	17	153.00K	44.32M	7	265
	Tripsoft Global Ltd.	VG	10	174.50K	5.96B	4	340
Deceptive Claims	GeoZilla Inc	US	771	2.68M	204.83M	8	462
	Family Locator, LLC	US	660	1.67M	60.61M	7	468
	APPLYFT LTD	CY	351	428.50K	35.20M	6	202
Misleading Ad Design	ZODIAC TECHNOLOGIES DMCC	AE	169	8.47M	2.30B	1	397
	Antelaria Limited	CY	17	5.00K	392.97M	1	165
	Mobitrans FZ LLC	AE	12	1.50K	598.20M	1	91

Table 5: Malicious domains linked from ads in our dataset. Cloudflare (CF) and Quad9 flag malware; Umbrella and CIRA flag malware and phishing. Only CIRA returns a reason, shown as MW or PH. Linking-ad counts cover all google's ATC ads reaching the domain's eTLD+1.

Domain	Num. of linking ads	Tranco rank	Blocking PDNS
spolecz[...].convertri.com	3K	59K	CIRA (MW), CF, Quad9
dropland.net	0	950K	CIRA (PH), CF
keysoft.store	18	1.6M	CIRA (PH), CF
licenzegenius.it	9	1.6M	CIRA (PH), CF
nerdused.com	62	1.6M	CIRA (PH), CF
eu.yourfavouritedocs.com	900	1.6M	CIRA (MW), CF
pyproxy.com	1	2.6M	CIRA (MW), Quad9
gopdfmanuals.com	300	3.2M	CIRA (MW), CF
kajoyefoqumanalytics.click	1	3.2M	CIRA (MW), CF
pdfscraper.com	42	3.2M	CIRA (MW), CF
pl.getniy.shop	0	3.2M	CIRA (PH), CF

extract the actual landing page URL from the `adurl` parameter. Unfortunately, over 66% of the ads available on the Ad Transparency Center do not contain (href) links in their HTML sources; thus we also extract links from ad OCR texts using the `urlextract` Python library [57]. We query the URL hostnames using four free and widely-used Protective DNS (PDNS) services: Cloudflare 1.1.1.1 for Families [77], Cisco Umbrella [19], Quad9 [3] and CIRA DNS Firewall [2]. The PDNS services work by blocking malicious domains at the DNS level by rewriting responses to return a reserved IP address, or an Extended DNS Error (EDE) code such as 16 (Censored) or 17 (Filtered) [59]. Using `zdns` [46], we query 14,019 distinct hostnames extracted from ads in our 188k creative dataset. To minimize false positives, below we only report 11 hostnames that are flagged by two or more PDNS providers (Table 5).

Our additional verification search revealed concerning results about the PDNS-flagged domains. According to a January 2026 report by Google Threat Intelligence Group [37], *pyproxy.com* is part of *IPIDEA*, the world's largest residential proxy network that enabled various botnets and state-level adversaries. According to Sophos [88, 90] and TrueSec [96], *pdfscraper.com* is part of the Tamperedchef info-stealer malvertising campaign distributed via Google ads. *gopdfmanuals.com* is listed as part of malvertising campaigns by CyberProof [94]. Ads for *pl.getniy.shop* and *dropland.net* have been seemingly removed from ATC since our crawl. The latter domain has a TrustScore of 1.6, with most reviewers labeling the website as scam [98]. In contrast, *keysoft.store*, a graymarket key license key seller, has highly positive reviews on Trustpilot, and we found no clear explanation for it being flagged by two providers, other than a misclassification. We provide example ads associated with the above domains in Appendix B(d). Two of these ads have been taken down by Google after we reported them for linking to malicious domains.

7 Discussion

We choose open-weight models at every stage of the AdLens pipeline based on several key considerations. First, open-source models offer greater control and fewer restrictive guardrails, enabling tools to identify and analyze malicious or deceptive signals more effectively. Second, as AdLens is intended to be accessible to a broad audience, it must remain independent of token-based pricing models, which can become limiting at scale. Our pipeline is designed to run efficiently on mid-tier GPUs, ensuring practical deployability. Third, open-weight models significantly reduce latency compared to closed-source alternatives by avoiding API overheads—an important advantage when processing millions of ads, where small delays become a bottleneck. Finally, this approach aligns with the broader goal of promoting open science.

Beyond ad classification and detection, our study reveals insights into deceptive strategies and the psychological triggers advertisers exploit. Spyware related ads frequently emphasize themes such as infidelity surveillance, unauthorized account access, and message monitoring. In privacy sensitive categories like file recovery or phone optimization, claims about retrieving deleted photos or restoring lost data are particularly effective [106]. Scareware often relies on alarming messages about malware, storage issues, or declining performance, which mirror common user concerns and are difficult to verify [24], increasing their persuasiveness. Other ads invoke stronger fears such as hacking, tracking, or data exposure, disproportionately affecting vulnerable groups like older adults and teenagers [8, 27]. Together with prior work [73], these

patterns highlight both the mechanics of deceptive advertising and the broader societal anxieties they exploit.

Impression counts, our primary reach metric, measure exposure rather than confirmed harm; establishing infection or financial loss would require access to ad networks and end-user devices beyond our scope. Nonetheless, independent evidence supports real-world impact: reviews linked to flagged ads report unauthorized charges and subscription traps (§6); threat intelligence independently identifies flagged domains as part of malicious campaigns [37, 88, 90, 96] (§6.5); and prior work links sponsored ads to ransomware delivery and deceptive user experiences [5, 10, 109]. AdLens exposes an enforcement gap: Google continues to serve ads from independently confirmed malicious domains (e.g., TamperedChef/pdfscraper and IPIDEA/pyproxy).

Adapting AdLens to a new ad category requires only new reference statements (Table 1), as classification is zero-shot. Supporting another ad library (e.g., Meta) primarily requires replacing the crawler to extract that platform's text and image ads, while video ads may need additional detection methods. AdLens can be deployed server-side by ad platforms for pre-publication screening or client-side by regulators, researchers, and civil society for large-scale auditing. AdLens's low adaptation cost makes it a practical tool for independent oversight.

Our findings suggest that current ad transparency libraries are insufficient for independent oversight. Ad creatives should be easily accessible and downloadable, landing page URLs should be consistently disclosed, and targeting information should be available to enable meaningful investigation of malicious campaigns. Finally, removed-ad archives should retain sufficient detail for retrospective analysis.

We also find that existing ad libraries are difficult to search in ways that support accountability. Transparency alone is not sufficient if researchers cannot effectively query for likely policy violations. To address this, we built and open sourced a search platform over our collected ads (Appendix B(f)). Given a free text query, the system retrieves semantically similar ads, enabling investigators to identify deceptive patterns and themes. This approach can scale to larger datasets with additional crawling, and can be further extended by incorporating image based retrieval using visual embeddings alongside text features.

At the same time, ad moderation remains inherently challenging. No automated system can be fully reliable: even small false positive rates can harm legitimate advertisers, while small false negative rates allow harmful ads to persist. As a result, effective enforcement still depends on human review, making the process slow and resource intensive. Our goal is not to replace human oversight, but to reduce reviewer burden by prioritizing ads that are more likely to violate policy. While our pipeline relies on machine learning models that introduce their own biases and limitations, combining multiple models helps mitigate individual weaknesses, as demonstrated by our ensemble approach balancing precision and recall.

Finally, we submitted violating ads from ten major themes to Google using the "Report this ad" button to better understand its moderation and reporting process. The process was different across user locations: reports from a Google account based in the EU were given an option to receive updates about the process, while report requests from a US-based Google account were not. We therefore used an EU-based personal account owned by a senior coauthor to report over 20 ads. Most ads were deemed non violating by Google, while in some cases violations appeared to be acknowledged but the ads remained visible on the platform, a week after the response. Moreover, ads similar to those confirmed as violations by Google continued to run, raising questions about how Google processes user reports and how effectively its moderation systems group and act on related ads. Notably, after we reported one of the 42 ads linking to *pdfscraper.com*, a domain involved in the Tamperedchef infostealer malvertising campaign (§6.5), the specific ad was removed, but the remaining 41 ads linking to the same domain stayed active. In another surprising case, Google said they were unable to review a scareware ad (Figure2b), which was publicly visible on the Ad Transparency Center and continued getting impressions, as indicated by the last shown date. We plan to continue our reports, while seeking contact with the Google Trust and Safety Team to share our concerns and suggestions.

We also observed that some scareware and deceptive ads were later replaced with benign content under the same advertiser and creative ID between collection and reporting (For example Figure 2(c)). This suggests that advertisers may rotate or swap creatives rapidly, complicating enforcement and accountability.

7.1 Limitations

Our study has a few limitations that future work can address. First, the text-focused semantic ranking and filtering cannot detect non-text modalities such as video and audio based ads, though AdLens is designed to be extensible to these modalities. Our OCR pipeline does not support non-Latin scripts such as Arabic and Cyrillic, though these account for less than 1.5% of our dataset. Since we collect ads exclusively from the Google Transparency Center, our dataset is restricted to ads shown in at least one EEA country, potentially missing newer forms of deception prevalent in other regions.

Similarly, AdLens currently excludes ads shown on Meta, TikTok, and other platforms, but it can be easily adapted to these platforms as well. Our analysis is further constrained to ad creatives, limiting our ability to capture network-level and redirection behavior of live ads that could reveal finer-grained deception signals. Rather than detecting all types of deceptive ads, AdLens focuses on three main problematic practices and ten themes related to software and mobile app ads. It can, however, be easily extended to detect other problematic ads by simply adding new reference statements and classification prompts.

As AdLens relies on instruction-following LLMs, it is susceptible to prompt injection and jailbreaking attacks. Adversaries may also embed malicious instructions within images using adversarial text to evade detection [89]. Practical deployments should therefore screen inputs for malicious prompts [95]. The embedding-based ranking stage is also vulnerable to adversarial manipulation. An attacker could paraphrase or obfuscate deceptive content to reduce embedding similarity, causing a violating ad to fall below the retrieval threshold and evade downstream classification. Since retrieval recall is critical to AdLens, strengthening this stage– for example, through adversarially augmented reference statements or embedding ensembles– is an important direction for future work.

Our ad landing page URL classification (§6.5) relies on PDNS providers, which may incorrectly flag some domains as malicious. We mitigate this by requiring at least two detections and providing additional evidence for detections. Future work should focus on deriving more robust and automated signals to address these gaps. As part of our landing page investigation, we noticed that one of the detected malicious domains was not linked from the ad we crawled, but from an ad that was present in the Ad Transparency Center page DOM as a related ad from the same advertiser. We removed this domain from our results and fixed the bug.

8 Acknowledgements

This work was supported by the National Science Foundation (NSF) under Grant No. 2341206. Acar, Darwish and Aghdam are supported by the Netherlands Organisation for Scientific Research (NWO) through a Vidi grant.

9 Conclusion

Our results show that deceptive advertisers continue to misuse the immense reach provided by advertisement networks. While Google takes down hundreds of millions of ads for violating their policies, we show that many deceptive and malicious ads remain online and reach millions of unsuspecting users. We find hundreds of scareware ads that attempt to induce panic and pressure users into installing apps. We uncover thousands of deceptive ad campaigns that make false claims about the advertised app's capabilities, giving users a false hope. Finally, we reveal many ads that link to domains that were found to distribute malware and classified as malicious by leading DNS providers. Our efficient and versatile detection pipeline, AdLens, takes advantage of a tiered architecture, combining a low-cost, high-recall prioritization stage with an LLM-based classifier based on freely available open-weight language models. In summary, our study brings transparency to the otherwise opaque advertising ecosystem by providing tools, datasets, and methods to researchers, regulators, and civil society organizations.

References

[1] 2022. Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act). http://data.europa.eu/eli/reg/2022/2065/oj. [Accessed 28 Feb. 2023].

[2] 2026. CIRA DNS Firewall Malware protection: Made for Canada – CIRA. https://www.cira.ca/en/cybersecurity/dns-firewall [Online; accessed 28. Apr. 2026].

[3] 2026. Quad9 | A public and free DNS service for a better security and privacy. https://quad9.net/service/service-addresses-and-features [Online; accessed 28. Apr. 2026].

[4] 1Password. 2024. Malvertising on Google Ads: It's hiding in plain site | 1Password. https://1password.com/blog/malvertising-on-google-ads.

[5] Lawrence Abrams. 2023. Ransomware access brokers use Google ads to breach your network. *BleepingComputer* (Jan. 2023). https://www.bleepingcomputer.com/news/security/ransomware-access-brokers-use-google-ads-to-breach-your-network.

[6] Admonsters. 2025. Digital Advertising Malware in 2024: Lessons for 2025 and Beyond - AdMonsters. https://www.admonsters.com/digital-advertising-malware-in-2024-lessons-for-2025-and-beyond/.

[7] Laurenz Aisenpreis, Gustav Gyrst, and Vedran Sekara. 2023. How Do US Congress Members Advertise Climate Change: An Analysis of Ads Run on Meta's Platforms. *Proceedings of the International AAAI Conference on Web and Social Media* 17 (6 2023), 2–11. doi:10.1609/ICWSM.V17I1.22121

[8] Muhammad Ali, Angelica Goetzen, Alan Mislove, Elissa M. Redmiles, and Piotr Sapiezynski. 2023. Problematic Advertising and its Disparate Exposure on Facebook. *32nd USENIX Security Symposium, USENIX Security 2023* 8 (jun 2023), 5665–5682. arXiv:2306.06052 https://arxiv.org/pdf/2306.06052

[9] Digital Advertising Alliance. 2026. YourAdChoices.com | Welcome to YourAdChoices.com. https://youradchoices.com/ [Online; accessed 2026-04-29].

[10] Digital Citizens Alliance and Unit 221B. 2022. UNHOLY TRIANGLE From Piracy to Ads to Ransomware: How Illicit Actors Use Digital Ads on Piracy Sites to Profit by Harming Internet Users. https://www.digitalcitizensalliance.org/clientuploads/directory/Reports/Unholy-Triangle-Report.pdf

[11] Aritz Arrate, José González-Cabañas, Ángel Cuevas, and Rubén Cuevas. 2020. Malvertising in Facebook: Analysis, Quantification and Solution. *Electronics 2020, Vol. 9, Page 1332* 9 (8 2020), 1332. Issue 8. doi:10.3390/ELECTRONICS9081332

[12] Gilles Baechler, Srinivas Sunkara, Maria Wang, Fedir Zubach, Hassan Mansoor, Vincent Etter, Victor Cărbune, Jason Lin, Jindong Chen, and Abhanshu Sharma. 2024. ScreenAI: A Vision-Language Model for UI and Infographics Understanding. *IJCAI International Joint Conference on Artificial Intelligence* (2 2024), 3058–3068. https://arxiv.org/pdf/2402.04615

[13] Simon Baker, Douwe Kiela, and Anna Korhonen. 2016. Robust Text Classification for Sparsely Labelled Data Using Multi-level Embeddings. 2333–2343 pages. https://aclanthology.org/C16-1220/

[14] Lior Bracha, Eitan Shaar, Aviv Shamsian, Ethan Fetaya, and Gal Chechik. 2023. DisCLIP: Open-Vocabulary Referring Expression Generation. (5 2023). https://arxiv.org/pdf/2305.19108

[15] David Breuer, Lucas Becker, and Matthias Hollick. 2026. Ad Personalization and Transparency in Mobile Ecosystems: A Comparative Analysis of Google's and Apple's EU App Stores. *Proceedings on Privacy Enhancing Technologies* (2026).

[16] Alina BÎZGĂ. 2026. Hijacked Google Ads Push Fake 7-Zip, Notepad++ and Office Downloads to Mac Users via Evernote Pages, Bitdefender Labs Warns. https://www.bitdefender.com/en-us/blog/hotforsecurity/hijacked-google-ads-push-fake-7-zip-notepad-and-office-downloads-to-mac-users-via-evernote-pages [Online; accessed 2026-04-29].

[17] Dominik Bär, Francesco Pierri, Gianmarco De Francisci Morales, and Stefan Feuerriegel. 2024. Systematic discrepancies in the delivery of political ads on Facebook and Instagram. *PNAS Nexus* 3 (6 2024). Issue 7. doi:10.1093/PNASNEXUS/PGAE247

[18] Ziwei Chen, Jiawen Shen, Luna, Hanyu Zhang, and Kristen Vaccaro. 2026. Deception at Scale: Deceptive Designs in 1K LLM-Generated E-Commerce Components. *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems* (4 2026), 1–18. doi:10.1145/3772318.3791063

[19] Cisco. 2026. Point Your DNS to Cisco Umbrella. https://securitydocs.cisco.com/docs/umbrella-dns/olh/146422.dita [Online; accessed 28. Apr. 2026].

[20] Confiant. 2024. The Foremost Benchmark Report on Digital Ad Quality, Security, and Privacy. Malvertising and Ad Quality Index. https://www.confiant.com/hubfs/reports/maq-2024.pdf.

[21] Avi Cooper, Keizo Kato, Chia Hsien Shih, Hiroaki Yamane, Kasper Vinken, Kentaro Takemoto, Taro Sunagawa, Hao Wei Yeh, Jin Yamanaka, Ian Mason, and Xavier Boix. 2024. Rethinking VLMs and LLMs for Image Classification. *Scientific Reports* 15 (10 2024). Issue 1. doi:10.1038/s41598-025-04384-8

[22] Marco Cova, Christopher Kruegel, and Giovanni Vigna. 2010. Detection and analysis of drive-by-download attacks and malicious JavaScript code. *Proceedings of the 19th International Conference on World Wide Web, WWW '10* (2010), 281–290. doi:10.1145/1772690.1772720;WGROUP:STRING:ACM

[23] Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, et al. 2025. Paddleocr 3.0 technical report. *arXiv preprint arXiv:2507.05595* (2025).

[24] Peter R. Darke and Robin J.B. Ritchie. 2007. The defensive consumer: Advertising deception, defensive processing, and distrust. *Journal of Marketing Research* 44, 1 (2007), 114–127. doi:10.1509/JMKR.44.1.114

[25] Laura Edelson, Shikhar Sakhuja, Ratan Dey, and Damon McCoy. 2019. An Analysis of United States Online Political Advertising Transparency. (2019).

[26] FTC. 2022. Bringing Dark Patterns to Light. https://www.ftc.gov/reports/bringing-dark-patterns-light. Accessed: 2026-04-13.

[27] Liza Gak, Seyi Olojo, and Niloufar Salehi. 2022. The Distressing Ads That Persist: Uncovering The Harms of Targeted Weight-Loss Ads Among Users with Histories of Disordered Eating. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 377 (Nov 2022), 23 pages. doi:10.1145/3555102

[28] Paula Alexandra Gitu, Roberto Cerina, Stefanie Vandevijvere, and Roselinde Kessels. 2024. AdDownloader: Automating the retrieval of advertisements and their media content from the Meta Online Ad Library. *SoftwareX* 28 (12 2024), 101919. doi:10.1016/J.SOFTX.2024.101919

[29] Google. 2026. Puppeteer. https://www.npmjs.com/package/puppeteer

[30] Google. 2023. Ads Transparency Center. https://adstransparency.google.com/. Accessed: 2026-04-13.

[31] Google. 2023. Announcing the launch of the new Ads Transparency Center. https://blog.google/innovation-and-ai/technology/ads/announcing-the-launch-of-the-new-ads-transparency-center/. Accessed: 2026-04-13.

[32] Google. 2026. List of ad policies. https://support.google.com/adspolicy/topic/2996750?hl=en&ref_topic=1308156,&sjid=14298733993174336754-NA. Accessed: 2026-02-22.

[33] Google and Hugging Face Team. 2026. google/embeddinggemma-300m. https://huggingface.co/google/embeddinggemma-300m

[34] Google and Hugging Face Team. 2026. google/translategemma-4b-it. https://huggingface.co/google/translategemma-4b-it

[35] Google and Ollama Team. 2026. gemma3:12b. https://ollama.com/library/gemma3:12b

[36] Google and Ollama Team. 2026. gemma4:26b. https://ollama.com/library/gemma4:26b

[37] Google Threat Intelligence Group. 2026. Disrupting the World's Largest Residential Proxy Network | Google Cloud Blog. https://cloud.google.com/blog/topics/threat-intelligence/disrupting-largest-residential-proxy-network [Online; accessed 28. Apr. 2026].

[38] Jeff Horwitz. 2025. Meta is earning a fortune on a deluge of fraudulent ads, documents show. https://www.reuters.com/investigations/meta-is-earning-fortune-deluge-fraudulent-ads-documents-show-2025-11-06. *Reuters* (Dec. 2025).

[39] Chen Huang and Guoxiu He. 2024. Text Clustering as Classification with LLMs. *SIGIR-AP 2025 - Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region* 1 (sep 2024), 374–384. arXiv:2410.00927 doi:10.1145/3767695.3769519

[40] Jui Ting Huang, Ashish Sharma, Shuying Sun, Li Xia, David Zhang, Philip Pronin, Janani Padmanabhan, Giuseppe Ottaviano, and Linjun Yang. 2020. Embedding-based Retrieval in Facebook Search. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (6 2020), 2553–2561. doi:10.1145/3394486.3403305

[41] Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. 2017. Automatic Understanding of Image and Video Advertisements. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017* 2017-January (7 2017), 1100–1110. doi:10.1109/CVPR.2017.123

[42] Ionut Ilascu. 2023. Hackers push malware via Google search ads for VLC, 7-Zip, CCleaner. *BleepingComputer* (Jan. 2023). https://www.bleepingcomputer.com/news/security/hackers-push-malware-via-google-search-ads-for-vlc-7-zip-ccleaner.

[43] Google Inc. 2023. Opening a Can of Whoop Ads: Detecting and Disrupting a Malvertising Campaign Distributing Backdoors | Google Cloud Blog. https://cloud.google.com/blog/topics/threat-intelligence/detecting-disrupting-malvertising-backdoors.

[44] Google Inc. 2024. Our 2023 Ads Safety Report. https://cloud.google.com/vision/docs/detecting-safe-search.

[45] Google Inc. 2026. Ads Safety Report 2026. https://services.google.com/fh/files/blogs/global_2025_adssafetyreport.pdf.

[46] Liz Izhikevich, Gautam Akiwate, Briana Berger, Spencer Drakontaidis, Anna Ascheman, Paul Pearce, David Adrian, and Zakir Durumeric. 2022. ZDNS: a fast DNS toolkit for internet measurement. In *Proceedings of the 22nd ACM Internet Measurement Conference*. 33–43.

[47] Jeff Burt. 2022. Domain aging gang CashRewindo picks vintage sites to push malvertising. https://www.theregister.com/2022/12/02/cashrewindo_scam_domain_aging/. Accessed: 2026-04-13.

[48] Chao Jia, Yinfei Yang, Ye Xia, Yi Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. *Proceedings of Machine Learning Research* 139 (2 2021), 4904–4916. https://arxiv.org/pdf/2102.05918

[49] Kaspersky. 2025. The AMOS infostealer is piggybacking ChatGPT's chat-sharing feature | Kaspersky official blog. https://www.kaspersky.co.uk/blog/share-chatgpt-chat-clickfix-macos-amos-infostealer/29796/ [Online; accessed 2026-04-29].

[50] 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

[51] Boshko Koloski, Senja Pollak, Roberto Navigli, and Blaž Škrlj. 2024. AutoML-guided Fusion of Entity and LLM-based Representations for Document Classification. (sep 2024). arXiv:2408.09794v1 https://arxiv.org/pdf/2408.09794v1

[52] Ludmila Kudryavtseva. 2018. Meet RoughTed — malvertising that can bypass adblockers. https://adguard.com/en/blog/malwertising_bypass_adblockers.html [Online; accessed 2026-04-29].

[53] Lokendra Kumar, Neelesh S. Upadhye, and Kannan Piedy. 2025. Advances and Challenges in Semantic Textual Similarity: A Comprehensive Survey. (12 2025). https://arxiv.org/pdf/2601.03270v1

[54] Paddy Leerssen, Jef Ausloos, Brahim Zarouali, Natali Helberger, and Claes H. de Vreese. 2019. Platform ad archives: Promises and pitfalls. *Internet Policy Review* 8 (2019), 1–21. Issue 4. doi:10.14763/2019.4.1421

[55] Simon Lermen, Daniel Paleka, Joshua Swanson, Michael Aerni, Nicholas Carlini, and Florian Tramèr. [n. d.]. Large-scale online deanonymization with LLMs. ([n. d.]). arXiv:2602.16800v2

[56] Zhou Li, Kehuan Zhang, Yinglian Xie, Fang Yu, and XiaoFeng Wang. 2012. Knowing your enemy: understanding and detecting malicious web advertising. In *Proceedings of the ACM conference on Computer and communications security*. 674–686.

[57] lipoja. 2026. Point Your DNS to Cisco Umbrella. https://pypi.org/project/urlextract/ [Online; accessed 28. Apr. 2026].

[58] Enze Liu, Sumanth Rao, Sam Havron, Grant Ho, Stefan Savage, Geoffrey M Voelker, and Damon McCoy. 2023. No privacy among spies: Assessing the functionality and insecurity of consumer android spyware apps. *Proceedings on Privacy Enhancing Technologies* (2023).

[59] Mingxuan Liu, Yiming Zhang, Xiang Li, Chaoyi Lu, Baojun Liu, Haixin Duan, and Xiaofeng Zheng. 2026. Understanding the Implementation and Security Implications of Protective DNS Services. https://www.ndss-symposium.org/ndss-paper/understanding-the-implementation-and-security-implications-of-protective-dns-services/

[60] Arunesh Mathur, Gunes Acar, Michael J Friedman, Eli Lucherini, Jonathan Mayer, Marshini Chetty, and Arvind Narayanan. 2019. Dark patterns at scale: Findings from a crawl of 11K shopping websites. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–32.

[61] Vivek Mehta, Seema Bawa, and Jasmeet Singh. 2021. WEClustering: word embeddings based text clustering technique for large datasets. *Complex & Intelligent Systems* 7 (2021), 3211–3224. doi:10.1007/s40747-021-00512-9

[62] Meta. 2025. Ad Library. https://www.facebook.com/ads/library/. Accessed: 2026-04-13.

[63] Microsoft Defender Experts Microsoft Threat Intelligence. 2025. Malvertising campaign leads to info stealers hosted on GitHub | Microsoft Security Blog. https://www.microsoft.com/en-us/security/blog/2025/03/06/malvertising-campaign-leads-to-info-stealers-hosted-on-github/ [Online; accessed 2026-04-29].

[64] Mistral and Ollama Team. 2026. mistral-small3.2:24b. https://ollama.com/library/mistral-small3.2:24b

[65] Zahra Moti, Asuman Senol, Hamid Bostani, Frederik Zuiderveen Borgesius, Veelasha Moonsamy, Arunesh Mathur, and Gunes Acar. 2024. Targeted and troublesome: Tracking and advertising on children's websites. In *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE, 1517–1535.

[66] Mozilla and Niemans Lab. 2024. A new Mozilla report exposes major flaws in social media ad libraries | Nieman Journalism Lab. https://www.niemanlab.org/2024/04/a-new-mozilla-report-exposes-major-flaws-in-social-media-ad-libraries/. Accessed: 2026-04-13.

[67] Terry Nelms, Roberto Perdisci, Manos Antonakakis, and Mustaque Ahamad. 2016. Towards Measuring and Mitigating Social Engineering Software Download Attacks. (2016). https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/nelms

[68] Hacker News. 2023. Malvertisers Using Google Ads to Target Users Searching for Popular Software. https://thehackernews.com/2023/10/malvertisers-using-google-ads-to-target.html.

[69] Massimo Nicosia and Alessandro Moschitti. 2017. Learning Contextual Embeddings for Structural Semantic Similarity using Categorical Information. *CoNLL 2017 - 21st Conference on Computational Natural Language Learning, Proceedings* (2017), 260–270. doi:10.18653/V1/K17-1027

[70] Priyanka Nigam, Yiwei Song, Vijai Mohan, Vihan Lakshman, Weitian (Allen) Ding, Ankit Shingavi, Choon Hui Teo, Hao Gu, and Bing Yin. 2019. Semantic Product Search. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (7 2019), 2876–2885. doi:10.1145/3292500.3330759

[71] Mark Ormerod, Jesús Martínez del Rincón, and Barry Devereux. 2021. Predicting Semantic Similarity Between Clinical Sentence Pairs Using Transformer Models: Evaluation and Representational Analysis. *JMIR Medical Informatics* 9, 5 (may 2021), e23099. doi:10.2196/23099

[72] Harun Oz, Ahmet Aris, Albert Levi, and A Selcuk Uluagac. 2022. A survey on ransomware: Evolution, taxonomy, and defense solutions. *ACM Computing Surveys (CSUR)* 54, 11s (2022), 1–37.

[73] Ayse Bengi Ozcelik and Kaan Varnali. 2019. Effectiveness of online behavioral targeting: A psychological perspective. *Electronic Commerce Research and Applications* 33 (jan 2019), 100819. doi:10.1016/J.ELERAP.2018.11.006

[74] PaddlePaddle. 2026. PaddlePaddle/PaddleOCR-VL. https://huggingface.co/PaddlePaddle/PaddleOCR-VL

[75] Rajvardhan Patil, Sorio Boit, Venkat Gudivada, and Jagadeesh Nandigam. 2023. A Survey of Text Representation and Embedding Techniques in NLP. *IEEEA* 11 (2023), 36120–36146. doi:10.1109/ACCESS.2023.3266377

[76] Alina Petukhova, João P. Matos-Carvalho, and Nuno Fachada. 2025. Text clustering with large language model embeddings. *International Journal of Cognitive Computing in Engineering* 6 (dec 2025), 100–108. arXiv:2403.15112 doi:10.1016/j.ijcce.2024.11.004

[77] Matthew Prince. 2025. Introducing 1.1.1.1 for Families. https://blog.cloudflare.com/introducing-1-1-1-1-for-families [Online; accessed 28. Apr. 2026].

[78] Qwen and Ollama Team. 2026. qwen3.5:27b. https://ollama.com/library/qwen3.5:9b

[79] Qwen and Ollama Team. 2026. qwen3.5:27b. https://ollama.com/library/qwen3.5:27b

[80] 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. *Proceedings of Machine Learning Research* 139 (2 2021), 8748–8763. https://arxiv.org/pdf/2103.00020

[81] Gábor Recski, Eszter Iklódi, Katalin Pajkossy, and András Kornai. 2016. Measuring semantic similarity of words using concept networks. (2016), 193–200. http://github.com/recski/wordsim

[82] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference* (8 2019), 3982–3992. doi:10.18653/v1/D19-1410

[83] Rein Kuffel. 2024. Trojan Extension Malware: 3-Year Campaign, 300k Infections. https://keepaware.com/blog/trojan-extension-malware-3-year-campaign-300k-infections. Accessed: 2026-04-13.

[84] Ritik Roongta, Julia Jose, Hussam Habib, and Rachel Greenstadt. 2025. Sheep's clothing, wolfish intent: Automated detection and evaluation of problematic'allowed'advertisements. *Proceedings on Privacy Enhancing Technologies* (2025).

[85] Christian Schlarmann, Francesco Croce, Nicolas Flammarion, and Matthias Hein. 2025. FuseLIP: Multimodal Embeddings via Early Fusion of Discrete Tokens. (jun 2025). arXiv:2506.03096 https://arxiv.org/pdf/2506.03096

[86] Scientific American. 2023. Online Ads Can Infect Your Device with Spyware. https://www.scientificamerican.com/article/online-ads-can-infect-your-device-with-spyware/. Accessed: 2026-04-13.

[87] Krebs On Security. 2024. Using Google Search to Find Software Can Be Risky – Krebs on Security. https://krebsonsecurity.com/2024/01/using-google-search-to-find-software-can-be-risky/.

[88] Sushmita Shetty. 2026. TamperedChef serves bad ads, with infostealers as the main course | SOPHOS. https://www.sophos.com/en-us/blog/tamperedchef-serves-bad-ads-with-infostealers-as-the-main-course [Online; accessed 29. Apr. 2026].

[89] Congzheng Song and Vitaly Shmatikov. 2018. Fooling OCR systems with adversarial text images. *arXiv preprint arXiv:1802.05385* (2018).

[90] Sophos. 2026. GitHub | IoCs/TamperedChef_IOCs.csv at master · sophoslabs/IoCs. https://github.com/sophoslabs/IoCs/blob/master/TamperedChef_IOCs.csv#L30 [Online; accessed 29. Apr. 2026].

[91] SOPHOS. 2026. TamperedChef serves bad ads, with infostealers as the main course | SOPHOS. https://www.sophos.com/en-us/blog/tamperedchef-serves-bad-ads-with-infostealers-as-the-main-course.

[92] Anthony Spadafora. 2021. Google Ads abused by hackers for major cryptocurrency heist | TechRadar. https://www.techradar.com/news/google-ads-abused-by-hackers-to-steal-hundreds-of-thousands-in-cryptocurrency [Online; accessed 2026-04-29].

[93] Arthur W Stephens. 2000. Alas for Adware-We Know It Too W ell A Thesis Of The Nature of Adware In Practice as Practicum in partial fulfillment of requirements for: The GIAC Security Essentials Certification The SANS Organization. http://www.sans.org

[94] CyberProof Research Team. 2026. Infostealers Strike Again: Malicious Installers Pass Through EDRs Undetected. https://www.cyberproof.com/blog/infostealers-strike-again-malicious-installers-impersonate-legitimate-productivity-tools

[95] Meta Llama Team. 2024. Prompt-Guard-86M. https://huggingface.co/meta-llama/Prompt-Guard-86M. Accessed: 2026-02-08.

[96] Truesec. 2026. Malicious Appsuite PDF Editor Spreads Tamperedchef Malware. https://www.truesec.com/hub/blog/tamperedchef-the-bad-pdf-editor

[97] Trustpilot. 2025. Govulo Reviews | Read Customer Service Reviews of govulo.com. https://www.trustpilot.com/review/govulo.com [Online; accessed 2026-04-30].

[98] Trustpilot. 2026. dropland recordings Reviews | Read Customer Service Reviews of dropland.net. https://www.trustpilot.com/review/dropland.net [Online; accessed 2026-04-30].

[99] Trustpilot A/S. 2026. Trustpilot. https://www.trustpilot.com/. Accessed: 2026-04-15.

[100] Tim vor der Brück and Marc Pouly. 2024. Estimating Text Similarity based on Semantic Concept Embeddings. (jan 2024). arXiv:2401.04422 https://arxiv.org/pdf/2401.04422

[101] Chao Wang, Tianming Liu, Yanjie Zhao, Lin Zhang, Xiaoning Du, Li Li, and Haoyu Wang. 2024. Towards demystifying Android adware: dataset and payload location. In *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering Workshops*. 167–175.

[102] Zixuan Wang, Jinghao Shi, Hanzhong Liang, Xiang Shen, Vera Wen, Zhiqian Chen, Yifan Wu, Zhixin Zhang, and Hongyu Xiong. 2025. Filter-And-Refine: A MLLM Based Cascade System for Industrial-Scale Video Content Moderation. (7 2025). https://arxiv.org/pdf/2507.17204

[103] Wikipedia contributors. 2026. Malware. https://en.wikipedia.org/wiki/Malware. Accessed: 2026-04-14.

[104] Wikipedia contributors. 2026. Scareware. https://en.wikipedia.org/wiki/Scareware. Accessed: 2026-04-14.

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

[106] Murooj Yousef, Sharyn Rundle-Thiele, and Timo Dietrich. 2023. Advertising appeals effectiveness: a systematic literature review. *Health Promotion International* 38, 4 (aug 2023). doi:10.1093/HEAPRO/DAAB204

[107] Z-ai and Hugging Face Team. 2026. zai-org/GLM-4.6V-Flash. https://huggingface.co/zai-org/GLM-4.6V-Flash

[108] Eric Zeng, Tadayoshi Kohno, and Franziska Roesner. 2020. Bad news: Clickbait and deceptive ads on news and misinformation websites. In *Workshop on Technology and Consumer Protection*. 1–11.

[109] Eric Zeng, Tadayoshi Kohno, and Franziska Roesner. 2021. What makes a "bad" ad? User perceptions of problematic online advertising. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–24.

[110] Eric Zeng, Xiaoyuan Wu, Emily N Ertmann, Lily Huang, Danielle F Johnson, Anusha T Mehendale, Brandon T Tang, Karolina Zhukoff, Michael Adjei-Poku, Lujo Bauer, Ari B Friedman, and Matthew S Mccoy. 2025. Measuring Risks to Users' Health Privacy Posed by Third-Party Web Tracking and Targeted Advertising. *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26-May 1, 2025, Yokohama, Japan* 1 (2025). doi:10.1145/3706598.3714318

[111] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*. 11975–11986.

[112] Lianmin Zheng, Wei Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems* 36 (6 2023). https://arxiv.org/pdf/2306.05685

[113] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to Prompt for Vision-Language Models. *International Journal of Computer Vision* 130 (10 2022), 2337–2348. Issue 9. doi:10.1007/s11263-022-01653-1

A Open Science and Ethical Considerations

We rate-limited all crawling and PDNS lookups, and collected only publicly accessible data from ad transparency repositories, leaving publisher revenue and ad delivery unaffected. We disclosed a sample of violating ads to Google and we continue to report ads of the types Google acted on. Sharing ad creatives may implicate copyright, but we consider their inclusion here fair use, consistent with prior work [65, 108]; we release all code and data publicly at https://github.com/Racro/ccs-2026-AdLens.

B Supplementary Material

The following artifacts are hosted in the paper's public repository at https://github.com/Racro/ccs-2026-AdLens/blob/master/ARTIFACTS.md and the project website is hosted at https://racro.github.io/project/adlens/.

(a) Google Ads policy violations: the full table of Google Ads policies applicable to software ads, organized by category and policy theme, from which we derive the taxonomy in §4.

(b) Pipeline latency: per-stage p50/p95 latency over 300 ads on a single GPU (§5).

(c) Misleading ad design examples: creatives carrying no advertiser information, including the two highest-impression ads discussed in §6.

(d) Ads on PDNS-flagged domains: ads linking to domains flagged by multiple Protective DNS providers (§6.5).

(e) Misconfigured ads: ads whose creative is garbled or absent, often replaced by a white box or a bare AdChoices icon. Though benign, these reflect gaps in the robustness of Google's transparency platform.

(f) Semantic search tool: a screenshot and description of the web interface for embedding-based ad retrieval (§5).