Sheep's clothing, wolfish intent

Automated detection and evaluation of problematic 'allowed' advertisements

Link to Paper

Ritik Roongta* NYU (ritik.r@nyu.edu)

Julia Jose NYU

Hussam Habib NYU

Rachel Greenstadt NYU

Introduction

- Internet users adopt privacy-enhancing technologies (PETs) like adblockers to avoid intrusive advertising, but this often disrupts the web's revenue model and draws strong pushback from advertisers and publishers
- The Acceptable Ads Standard (300M users in 2023) aims to strike a balance through allowlisting.
- Our evaluation shows AAS increases problematic ad content, supporting the view that privacy-aware users often pay a higher price for a more private web.

Methodology

- **TAXONOMY**
  - We adopt three guiding principles
    - **balance** user, advertiser, and publisher **interests** (Table 1)
    - enable **multi-label classification** through mutually exclusive categories
    - maximize the **analytical ability** of the annotators and LLMs via **keywords** for each category
- **AD COLLECTION**
  - We collect ads in two different scenarios
    - Under-age (<18) and adult (>18) populations in the US
    - Unauthenticated fresh profiles from Germany and US
  - 7 multi-lingual experts helped to label the ad dataset

Stakeholder	Insights
Ad Exchanges	Age-based and Geographical Regulations; Restricted topics like Health, Financial commodities, etc.; Fraud and scam content, and political content.
Regulators	Publisher and Advertiser compliant requirements around native advertising, weight loss ads, etc.
Ad Consortiums	Aesthetic ad standards around autoplaying ads, creative dimensions etc.; Guidelines for disruptive user experience.
Industrial Studies & Reports	User perceptions of ads, popularity of adblockers, and popular reasons for adblocking.
Academic Studies	Intrusiveness of ad content and its effect on user perceptions; Dark patterns prevalent in online media; Captures domains like psychology, privacy, and security.

Table 1: Different stakeholders and their contributions

Key Findings

- Acceptable Ads are significantly more problematic than Control group ads in all scenarios
- ABP users experience 13.6% higher incidence of problematic ads
- Control group ads are significantly problematic for German(/adult) users compared to US(/under-age)
- Other demographic vectors cease to make an observable impact under the presence of allowlisted ad exchanges.

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

Table 2: Statistical difference in the prevalence of problematic ads

LLM assisted problematic ad detection

We prompted CLIP, GPT-4o-mini, and GPT-4o to classify ad images as problematic based on our taxonomy

MOTIVATION

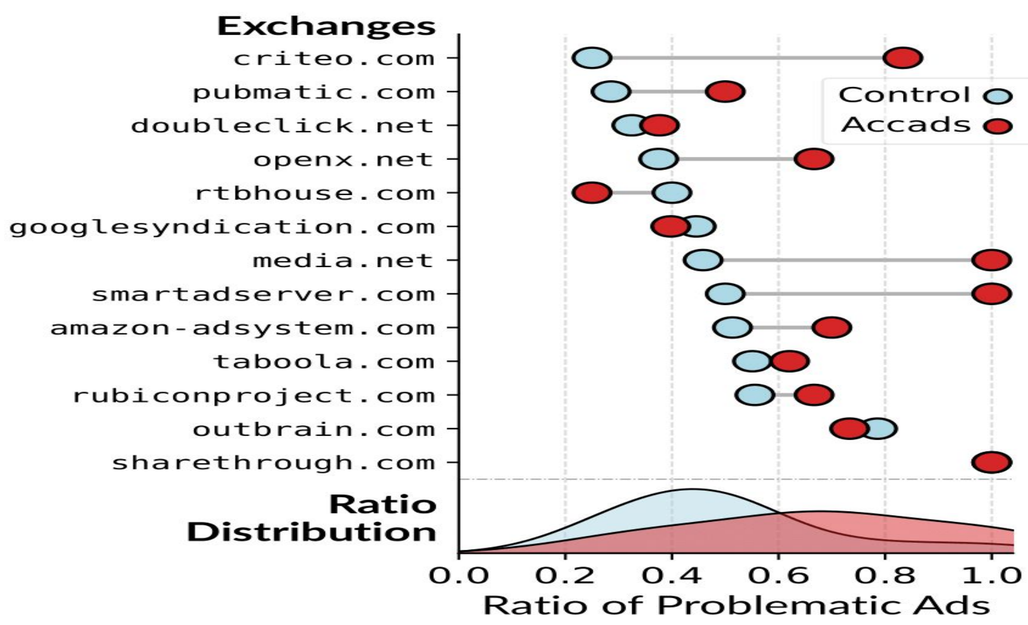
- Ad content is an important aspect of web experience [1,2]
- Ad content moderation is challenging - involves subjective judgement and nuanced definitions
- LLMs combine visual and textual reasoning to capture subtle context, provide human readable explanations
- LLMs are extremely conservative, leading to false positives in some categories

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

Table 3: Krippendorff's Alpha agreement scores between human and LLM labels for each problematic category

Exchange Analysis

- Ad exchanges added(/retained) by the Acceptable Ads allowlist decrease(/increase) problematic ad ratios
- Plausible reasons - Lack of tracking vectors or differential targeting of privacy aware users



Conclusion

- LLMs are effective in detecting problematic ad content - opening avenues for adversaries to fingerprint users with advent of In-browser LLMs
- Acceptable ads being more problematic highlights an important issue - privacy aware users pay a higher cost of privacy when they try to enhance their privacy

References

[1] Zeng et al. What Makes a "Bad" Ad? (CHI 2021)

[2] Ali et al. All Things Unequal: Measuring Disparity of Potentially Harmful Ads on Facebook. (USENIX 2023)