Unveiling AI-Driven Web Applications: Insights into Characteristics, Functionality, and Compliance

LIUHUO WAN, The University of Queensland, Australia
ZICONG LIU, The University of Queensland, Australia
CHUAN YAN, The University of Queensland, Australia
LIUJIA WAN, Northeastern University, China
NAIPENG DONG, The University of Queensland, Australia
ZI HUANG, The University of Queensland, Australia
GUANGDONG BAI*, City University of Hong Kong, China

Collaborative platforms such as Google Workspace, Microsoft Teams, and Zoom increasingly rely on third-party applications (referred to as plugins) to extend their core functionalities, with AI-assisted plugins emerging as a key driver of productivity. Despite their popularity and rapid adoption, little is known about the characteristics of the marketplace, the potential security and privacy risks that concern users, and the compliance of plugins with AI ethics guidelines. In this paper, we present the first large-scale, cross-platform study of plugins from five major web application marketplaces, covering domains from office productivity to software development. We systematically examine the distribution characteristics of current plugins, analyze users' concerns, and assess their compliance with emerging AI regulations. Our findings indicate that (i) the current marketplaces exhibit an uneven distribution of functionality and installations, (ii) AI-assisted plugins face a range of emerging issues that negatively impact user experience, and (iii) a significant proportion of plugins fail to comply with established AI ethics principles. Our work highlights the need for strigent policies and security auditing to maintain quality of AI-assisted plugins.

CCS Concepts: • **Security and privacy → Web application security**.

Additional Key Words and Phrases: AI-driven web applications, characteristics, functionality, compliance, software marketplaces

ACM Reference Format:
Liuhuo Wan, Zicong Liu, Chuan Yan, Liujia Wan, Naipeng Dong, Zi Huang, and Guangdong Bai. 2018. Unveiling AI-Driven Web Applications: Insights into Characteristics, Functionality, and Compliance. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 25 pages. https://doi.org/XXXXXXX.XXXXXXX

*Corresponding author.

Authors' Contact Information: Liuhuo Wan, School of Electrical Engineering and Computer Science, The University of Queensland, St Lucia, QLD, Australia; Zicong Liu, School of Electrical Engineering and Computer Science, The University of Queensland, St Lucia, QLD, Australia; Chuan Yan, School of Electrical Engineering and Computer Science, The University of Queensland, St Lucia, QLD, Australia; Liujia Wan, School of Electrical Engineering and Computer Science, Northeastern University, Qinhuangdao, Hebei, China; Naipeng Dong, School of Electrical Engineering and Computer Science, The University of Queensland, St Lucia, QLD, Australia; Zi Huang, School of Electrical Engineering and Computer Science, The University of Queensland, St Lucia, QLD, Australia; Guangdong Bai, School of Electrical Engineering and Computer Science, City University of Hong Kong, Hong Kong, HK, China.

Conference acronym 'XX, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Collaborative platforms such as Google Workspace [5], Microsoft Teams [6], and Zoom [8] have experienced rapid growth due to the increasing demand for remote collaboration. These platforms support a wide range of custom plugins that enhance their core functionality. Plugins can be developed either by the platform providers or by third-party developers. Built on top of collaborative platforms, these plugins utilize the APIs provided by the platforms and often integrate with external services. Users can install and manage plugins through the marketplaces. Taking Google Workspace as a representative example, plugins available in the Google Marketplace are seamlessly integrated into the Google Workspace environment through standardized APIs and extension frameworks. Once installed, these plugins can be accessed directly within core applications such as Gmail, Google Docs, or Slides, appearing in the user interface as additional menus, sidebars, or contextual actions. Beyond extending native functionalities, these plugins can be enhanced to manage users' resources stored in Google Workspace. Due to such easy accessibility and rich functionality, plugins have gained great popularity in recent years. For example, Zoom has accumulated over 6 million plugin installations [7], while Google reports more than 5 billion plugin installations [4].

Recent advancements in artificial intelligence (AI) have further accelerated the integration of AI-assisted plugins into collaborative platforms. Unlike traditional web plugins, which usually perform limited tasks based on simple rules (e.g., counting files or lines changed in a merge request), AI-assisted plugins can perform more intelligent operations. For example, a plugin based on code language model can generate a concise summary of a merge request, highlighting the developer's intent and key changes to support code review. As reported in recent studies [20, 49], there has been a growing use of these plugins for supporting daily tasks such as email replying, pull request summarization, and bot-assisted interactions in team chats.

Most third-party plugins function as black boxes, whose internal logic is hidden from the host platform. They can pose significant security and privacy risks, as revealed by existing studies [17, 49, 68, 82, 84, 88, 90]. Current collaborative platforms largely rely on the vetting process to mitigate risks before a plugin is published to the public marketplace. However, security concerns are often not thoroughly addressed. The vetting process focuses primarily on the plugin's intended functionality, descriptions, privacy policies, and basic eligibility requirements, lacking a mandatory in-depth security auditing. Existing research on plugin security [17, 52, 82, 88, 90] also has approached the problem largely from a traditional perspective and has not specifically considered the unique features of emerging AI-assisted plugins. Dynamic execution [52] and network traffic analysis [88] are commonly applied to detect security risks such as confidential data leakage and privilege escalation. However, these approaches focus primarily on traditional API execution and overlook the AI components introduced by AI-assisted plugins. They seldom test or evaluate the behavior and security implications of the integrated AI components within the plugins.

Compared to traditional plugins, analyzing AI-assisted plugins requires more focus on their AI-specific features, including checking the role of AI in the loop and assessing AI-specific security risks. First, the AI component is dynamic and requires more human-like interaction [23, 56] with the plugin compared with API execution applied in traditional plugins. Second, AI features may introduce new attack surfaces, such as the potential for harmful or malicious content generation [22, 23, 69, 70]. These challenges cannot be addressed by directly applying existing approaches, and they demand new approaches specifically designed to account for AI-driven features and their unique characteristics. However, many crucial aspects of AI-assisted plugins, user concerns, and ethical compliance have yet to be thoroughly examined in the research community.

Our Work. To address the existing gap in understanding the broader landscape of AI-assisted plugins, we conduct a comprehensive study of these plugins across prominent marketplaces,

spanning domains from office productivity (e.g., Zoom) to software development tools (e.g., GitHub). Our study is driven by three key research questions (RQs) that are critical to platform operators, users, and plugin developers. These include: ***RQ1: What are the market characteristics of AI-assisted plugins***? For example, their categories and installation patterns. ***RQ2: What are users' concerns regarding the functionality and security of AI-assisted plugins***? along with ***RQ3: how plugin developers adhere to AI-specific regulations and the potential non-compliance risks they may pose***. Through these RQs, we aim to provide a holistic view of plugin ecosystems and highlight the key aspects of these plugins.

RQ1: ***Characterizing existing AI-assisted plugins.*** In this research question, we analyze the categories and installation patterns of AI-assisted plugins in the marketplace. We first collect all currently available plugins across five marketplaces. From this dataset, we identify AI-assisted plugins (a total of 6,170 plugins) by leveraging both marketplace labels and the descriptive text of each plugin. One challenge is that traditional categorization methods do not fit AI-assisted plugins, and there is no well-labeled data available for training a categorization model. To address this, we adopt the AI-activity taxonomy provided by National Institute of Standards and Technology (NIST) and apply zero-shot classification, which allows us to accurately categorize previously unseen data into predefined categories without additional training. By analyzing the distribution of plugin types and their popularity, we can uncover trends in the marketplace and identify which functionalities are most prevalent among available plugins. This analysis offers insights into the structure of the marketplace and the representation of different plugin functionalities across the platform. Further details are provided in Section 3.

RQ2: ***Understanding user concerns regarding AI-Assisted plugin functionality.*** This research question focuses on understanding users' concerns regarding AI-assisted plugins. Unlike traditional plugins, it is unclear what specific issues may trigger user concerns, particularly from functionality and security perspectives. To address this gap, we first collect and analyze user reviews with low ratings. We then apply unsupervised clustering to group these reviews based on semantic similarity, and assign labels to the resulting clusters using BERTopic. This approach enables us to automatically identify emerging topics that differ from traditional concerns. Our analysis reveals that AI-assisted plugins may introduce malware, security risks, or functionality issues. In addition, we examine whether the same plugin exhibits consistent behavior across multiple marketplaces. Together, these analyses allow us to systematically capture user concerns related to both functionality and security. We present the details of this study in Section 4.

RQ3: ***Assessing plugin compliance with AI-specific ethics.*** This research question examines developers' practices in complying with AI ethics guidelines. While it is crucial for AI-assisted plugins to adhere to ethical principles, existing research in this area is lacking, partly because the specific AI ethics requirements applicable to plugins remain unclear. To address this challenge, we construct a taxonomy of AI ethics by synthesizing widely adopted guidelines and insights from the research community. We then evaluate the compliance of AI-assisted plugins against this taxonomy, focusing on aspects such as disclosing AI usage and preventing unsafe content generation. Through textual semantic analysis and dynamic triggering of AI plugins, we find that a noticeable proportion of plugins fail to comply with the established ethical standards. The details of this analysis are presented in Section 5.

Contribution. The contributions of our work are summarized as follows:

- **The first large-scale characterization of AI-assisted plugins.** We offer the first comprehensive study of AI-assisted plugin marketplaces and their popularity. We reveal the AI-oriented functionalities provided by these plugins, offering insights for both marketplaces and end users. We also highlight the uneven distribution of installations across categories,

Table 1. Overview of five web platforms and plugins

Platform	# Plugins	# AI Plugins		Overview	Company	Pricing	Version	Support	Rating	Installation	Review
Microsoft	36,168	5,274	(14.60%)	✓	✓	✓	✓	✓	✓	✗	✓
Google Workspace	4,868	261	(5.32%)	✓	✓	✓	✗	✓	✓	✓	✓
Github	651	236	(36.25%)	✓	✓	✓	✗	✓	✓	✓	✗
Slack	2,436	160	(6.57%)	✓	✓	✓	✗	✓	✓	✗	✗
Zoom	2,401	239	(9.95%)	✓	✓	✗	✗	✓	✓	✗	✗
Total:	**46,524**	**6,170**									

which can serve as a reference for developers when planning future directions. Our findings indicate that AI-assisted plugins are still in their infancy, and user demand for AI-related functionalities remains an underexplored area with substantial growth potential.

- **A systematic assessment of security and functionality, and its practical impact.** We present the first systematic analysis of user concerns regarding the functionality of AI-assisted plugins. Our approach effectively detects newly introduced malfunctions or malware specific to AI-related features. We find that unexpected subscription requirements and unsatisfactory AI functionality are among the major concerns raised by end users.
- **Revealing the status quo of AI compliance.** Grounded in established AI ethics guidelines, we define a concrete notion of AI safety compliance for evaluating AI-assisted plugins. Using a combination of description analysis and dynamic testing of AI-generated content, we systematically assess compliance and uncover common violations. Our results show that many AI-assisted plugins fail to follow ethical principles, raising concerns about unsafe content generation and potential misuse. As the first study on ***AI compliance***, our research not only informs improvements to existing plugins but also provides insights for the future development of this ecosystem.

2 Data Collection

We selected five representative platforms, Microsoft, Google Workspace, GitHub, Slack, and Zoom, as the focus of our study, as shown in Figure 1. These platforms span a diverse range of mainstream web services and provide comprehensive coverage of the dominant categories in today's web based software landscape. For marketplaces such as Microsoft, GitHub, Zoom, and Slack, which support complete plugin listings through paginated interfaces, we use Selenium to simulate "next page" navigation and extract all plugin URLs systematically. In contrast, the Google Workspace Marketplace imposes a constraint by displaying only the top 100 results for any given query and lacking a full listing interface. To address this, we adopt a keyword-based search strategy [84]. Specifically, we issue queries using the 10,000 most frequently used English words [35], and collect the resulting plugin entries. We apply deduplication to yield the unique plugins. The distribution of plugins collected from each platform is summarized in Table 1. Microsoft hosts the largest number of plugins, approximately 14,939, followed by Google Workspace (4,868), Slack (2,436) and Zoom (2,401). GitHub, by contrast, contains a relatively smaller selection with only 651 plugins.

Filtering of AI-assisted plugins. For marketplaces that provide a dedicated category of AI-assisted plugins, we collect all plugins listed under this category. Slack offers a dedicated category, *AI Apps & Assistants*. GitHub provides a similar category, *AI-Assisted Apps*. Microsoft provides an *AI Apps & Agents* category. For marketplaces without such a dedicated category, we instead employ a keyword-based approach by examining plugin descriptions for AI-related terms such as artificial intelligence (AI), machine learning (ML), deep learning (DL), and large language models (LLM). We manually sampled and reviewed 100 samples each from Google and Zoom in the collected dataset. Manual verification indicates that the dataset has high accuracy 99%. We further investigate the

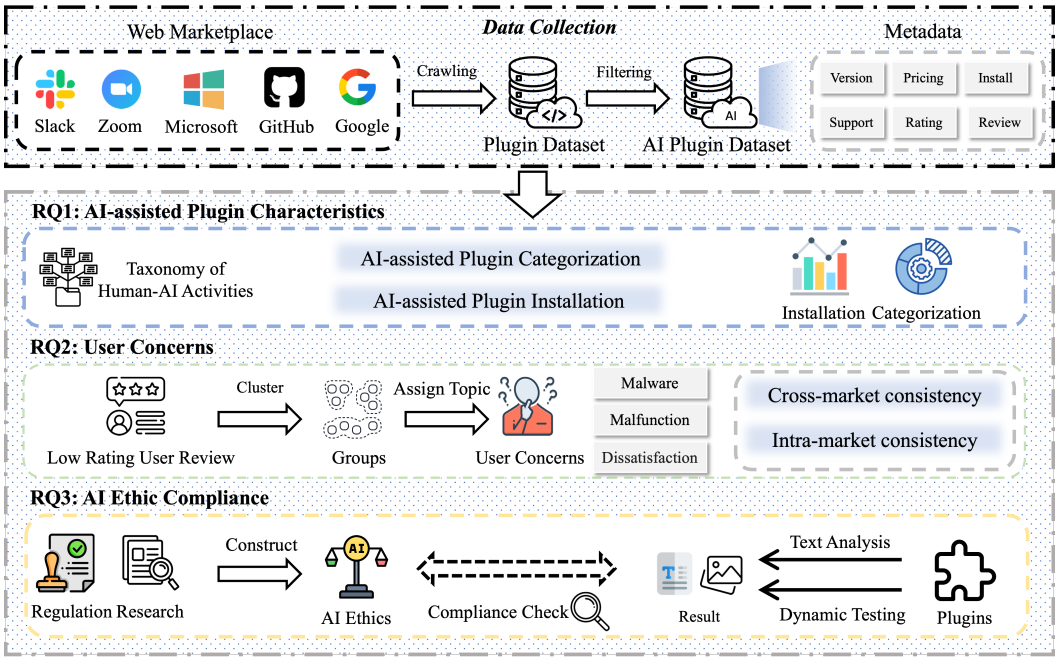

Fig. 1. Overview of the approach

underlying reasons. One plugin relies entirely on *WordNet*. Although a valuable resource for NLP, WordNet is a lexical database rather than a true AI model, a technical reality shared by the other plugin. This combined strategy allows us to efficiently gather a comprehensive set of plugins likely to involve AI capabilities for further analysis.

As shown in the Table 1, AI-assisted plugins constitute a substantial proportion of the overall market. GitHub accounts for the largest proportion at 36.25%. Given that its primary user base consists of developers, this suggests that developers are particularly receptive to adopting AI-assisted plugins. Following GitHub, Microsoft exhibits a 14.6% representation. Considering that Microsoft currently offers the most extensive range of plugins across diverse functionalities, this indicates that AI-assisted plugins are gaining traction. Other platforms, such as Google Workspace, Slack, and Zoom, also feature AI-assisted plugins, with representations ranging from 5.32% to 9.95%, reflecting the growing popularity of AI integration across multiple platforms.

Overview. Across all five platforms, core metadata like the overview, company information, and pricing are consistently required, reflecting a common baseline of transparency and usability. In contrast, requirements related to technical details vary: Microsoft uniquely mandates version information, whereas developer support is required across all platforms. Overall, while foundational documentation is universal, platforms differ in how they support maintenance, transparency, and trust, with Microsoft and Google Workspace imposing the most extensive requirements, consistent with their enterprise orientation.

Research scope. Our definition of AI-assisted covers all perspectives AI technique, including traditional ML and more advanced LLM technique. Traditional machine learning methods rely on task-specific models trained on structured data, often requiring manual feature engineering and domain expertise. In contrast, large language models (LLMs), pre-trained on massive unstructured text corpora using transformer architectures, can perform diverse language tasks with minimal adaptation and generalize across applications, whereas traditional ML models are typically limited

Table 2. Taxonomy of human-AI activities

Human-AI Activity	Description
Content creation	generating new artifacts such as video, narrative, software code, synthetic data.
Content synthesis	combining and/or summarizing parts, elements, or concepts into a coherent whole.
Decision making	selecting a course of action from among possible alternatives in order to arrive at a solution.
Detection	identifying, by careful search, examination, or probing, the existence or presence of [something].
Digital assistance	acting as a personal agent for understanding and responding to commands and questions, and carrying out requested tasks in a conversational manner.
Discovery	finding, recognizing, or unearthing something for the first time.
Image analysis	recognizing attributes within digital images to extract meaningful information.
Information retrieval/search	finding information about specific topics of interest.
Monitoring	observing, checking, and watching over the process, quality, or state of [something] over time to gain insights into how [something] is behaving or performing.
Performance improvement	improving quality and efficiency of the intended outcomes.
Personalization	designing and tailoring [something] to meet an individual's characteristics, preferences, or behaviors.
Prediction	forecasting the likelihood of a future outcome.
Process automation	performing repetitive tasks, removing bottlenecks, reducing errors and loss of data, and increasing efficiency of a process.
Recommendation	suggesting or proposing a manageable set of viable options to aid decision-making.
Robotic automation	using physical machines to automate, improve, and/or optimize a variety of tasks.
Vehicular automation	automating physical transportation of goods, instrumentation and/or people.

to narrowly defined problems. Furthermore, these plugins support AI techniques for processing text, images, video, and audio. For example, some Zoom plugins summarize meetings by processing video input and generating text output. In our study, AI-assisted plugins are defined to include all forms of AI techniques, and their distribution is systematically analyzed.

3 RQ1: Characterizing AI-assisted Plugin Marketplaces

3.1 Categorization of plugins

In this section, we present an overview of the characteristics of AI-assisted plugin categorization, and user installation patterns, as shown in Figure 1. We first accurately categorize AI-assisted plugins. Current marketplaces employ a coarse grained categorization based primarily on the associated services, like "works with Outlook". Such classifications fail to capture the functional diversity of the plugins and thus offer limited insight into their actual capabilities. Existing classification systems designed for traditional applications like Google Play and the Apple App Store [88] fail to capture the novel functionalities of AI-assisted plugins. In order to systematically assess how embedded AI models contribute to the functionality of plugins, we adopt the AI Use Taxonomy [75, 78] developed by the U.S. National Institute of Standards and Technology (NIST) [62].

NIST has developed influential standards and databases, including the National Vulnerability Database [61], Standard Reference Database [47], Cloud Computing [57], Cybersecurity Framework [58, 85], Role-Based Access Control [30], Risk Management Framework [31], and Digital Identity Guidelines [37], which have shaped both industry practices and academic research, with over 30,000 citations. The NIST AI Use Taxonomy offers a framework for categorizing AI applications in software systems, including AI-assisted plugins. The NIST AI taxonomy is standardized, broadly applicable, and designed to be future proof, ensuring that our methodology remains relevant as the AI ecosystem evolves. For example, future emerging technologies such as robotic [60, 73] or vehicular automation [14] can be characterized without structural changes or manual efforts, enabling consistent longitudinal analysis beyond currently dominant AI forms. Table 2 details its 16 types of human-AI activities.

To this end, we leveraged the large language model `meta-llama 3.2` to automatically assign each plugin to the most appropriate category based on its textual description [79]. Due to the severe imbalance in the distribution of plugin categories, we sample 10 plugins from each category,

resulting in a total of 160 plugins following existing research [81]. We then annotated with three experts to form our benchmark. Table 4 presents the performance under different thresholds on our benchmark. We tested thresholds from 0.5 to 0.9. Beyond 0.6, further increases in the threshold yield marginal reductions in wrong labels but rapidly increase unlabeled data. To achieve optimal performance, we choose 0.6 as our threshold. This zero-shot approach eliminates the need for task-specific model training and, because the LLM is trained on general textual data, it can generalize effectively to our categorization task. Two authors independently reviewed 200 generated labels from the 6,170 AI-assisted plugins following category distribution for evaluation. This sample size follows established methodologies [13, 88, 93] and ensures unbiased representation of the full dataset. They read each plugin description to determine whether the assigned label was correct. In cases of disagreement, a third expert was consulted to resolve conflicts, and all three annotators subsequently discussed any remaining disagreements to reach a consensus. The LLM achieved a promising accuracy of 93%. At a 95% confidence level, this sample size 200 corresponds to a margin of error of ±3.5% [13]. Our sampling methodology is consistent with recent empirical studies in AI security and software engineering [54, 81, 93].

Table 4. Performance at different thresholds

Metrics	Threshold				
	0.5	**0.6**	**0.7**	**0.8**	**0.9**
Wrong labeled	15	10	7	7	3
Unlabeled	0	0	22	41	60
Accuracy	90.63%	**93.75%**	81.88%	70.00%	60.63%

Results. The distribution of plugin categories is summarized in Table 5. These plugins exhibit a similar distribution pattern across different marketplaces, with *digital assistance* consistently ranking as the top category, accounting for approximately 50% of all AI-assisted plugins. In the Microsoft Marketplace, the second most common categories are *decision making* (13.16%) and *performance improvement* (13.22%). In Google Workspace, *personalization* ranks second, comprising 9.25% of AI-assisted plugins. In GitHub, *recommendation* occupies the second position. In Slack, the second-largest category is *information retrieval/search*, accounting for 15.62% of AI-assisted plugins. Overall, while *digital assistance* dominates across platforms, the secondary categories vary notably, reflecting the distinct functional priorities of each ecosystem.

3.2 Installation of plugins

We also analyze plugin installation counts, as shown in Table 5. Since only the Google Workspace and GitHub marketplaces provide user installation data, we report the total number of users who installed plugins in each category in the *Users* column.

Google Workspace. We first examined the digital assistance category, the largest category of plugins. This category encompasses a highly diverse range of functionalities, such as collaborative coding platforms (e.g., Colaboratory with 75M users), data backup services (9M users), AI-powered meeting note-taking tools (4M users), and even entertainment-oriented plugins like games (2M users). Interestingly, while digital assistance consistently represents the largest share across platforms, the performance improvement category in Google Workspace has a larger user base, despite accounting for only 7.28% of the plugins. We examined the 19 plugins categorized under performance improvement and found that they offer AI-assisted tools designed to save customers time and improve efficiency. For example, one plugin with over 38 million installations provides a single functionality: translating Google Slides and images into more than 100 languages with a single click. This leads to substantial time savings for end users. Another example is a plugin with over one million installations that leverages AI to extract content from various sources, including PDFs, Google Docs, handwritten images, and text, transforming it into quiz forms tailored for user publication. Although these plugins provide relatively simple functionalities, their significant performance improvements have led to strong user adoption and popularity.

Table 5. Distribution of plugin categories and user installations across marketplaces

AI Categories	Microsoft	Google Workspace		GitHub		Slack	Zoom
	#Plugin	#Plugin	#User	#Plugin	#User	#Plugin	#Plugin
Content creation	42 (0.80%)	21 (8.05%)	2.13M (0.59%)	3 (1.27%)	3.01K (1.02%)	3 (1.88%)	2 (0.84%)
Content synthesis	35 (0.66%)	-	-	-	-	-	-
Decision making	694 (13.16%)	15 (5.75%)	64.60M (18.03%)	6 (2.54%)	0.44K (0.15%)	5 (3.12%)	2 (0.84%)
Detection	36 (0.68%)	4 (1.53%)	67.16K (0.19%)	1 (0.42%)	2 (0.00%)	-	-
Digital assistance	2143 (40.63%)	107 (40.99%)	98.52M (27.50%)	137 (58.05%)	201.28K(68.65%)	95 (59.38%)	155 (64.85%)
Discovery	74 (1.40%)	2 (0.77%)	467.00K (0.13%)	3 (1.27%)	65 (0.02%)	1 (0.62%)	-
Image analysis	30 (0.57%)	-	-	1 (0.42%)	34 (0.01%)	-	-
Information retrieval/search	258 (4.89%)	19 (7.28%)	11.04M (3.08%)	12 (5.08%)	30.32K (10.33%)	25 (15.62%)	19 (7.95%)
Monitoring	215 (4.08%)	3 (1.15%)	924.31K (0.26%)	-	-	1 (0.62%)	5 (2.09%)
Performance improvement	697 (13.22%)	19 (7.28%)	100.8M (28.14%)	13 (5.51%)	4.84K (1.65%)	5 (3.12%)	19 (7.95%)
Personalization	358 (6.79%)	25 (9.58%)	17.34M (4.84%)	15 (6.36%)	24.28K (8.27%)	12 (7.50%)	12 (5.02%)
Prediction	164 (3.11%)	8 (3.07%)	3.81M (1.06%)	4 (1.69%)	81 (0.03%)	1 (0.62%)	-
Process automation	233 (4.42%)	10 (3.83%)	23.24M (6.49%)	15 (6.36%)	2.06K (0.70%)	4 (2.50%)	6 (2.51%)
Recommendation	277 (5.25%)	21 (8.05%)	14.03M (3.92%)	17 (7.20%)	21.06K (7.18%)	4 (2.50%)	17 (7.11%)
Robotic automation	25 (0.47%)	3 (1.15%)	1.07M (0.30%)	9 (3.81%)	5.02K (1.71%)	4 (2.50%)	2 (0.84%)
Vehicular automation	13 (0.25%)	4 (1.53%)	10.98M (3.06%)	-	-	-	-
Total	**5274**	**261**	**358.30M**	**236**	**293.48K**	**160**	**239**

We further examined 15 plugins (5.75% of AI-assisted plugins) under the decision making category, which collectively account for 18.03% of users. Notably, one plugin, with 62 million users, enhances Google Forms by adding a timer, AI-protected camera monitoring, submission time tracking, and proctoring confidence level assessments. These features assist educators, recruiters, and businesses in making more informed decisions. Despite being a paid plugin, its convenient and practical functionalities have contributed to its widespread adoption and popularity among end-users.

GitHub. In the GitHub Marketplace, although information retrieval/search accounts for only 5% of all plugins, it attracts 10% of the total installations among AI-assisted plugins. Our investigation into plugins within this category reveals that one plugin, with 24.5K installations, primarily offers the functionality of "accessing real-time web search results without leaving your IDE". These plugins do not introduce additional functionalities. Rather, they act as connectors between two pre-existing services, offering greater convenience for end users. This observation aligns with our findings in Google Workspace, where users do not necessarily seek plugins with complex capabilities; rather, simple yet useful functionalities tend to gain greater popularity.

> **Finding 1:** While developers often prioritize adding new functionalities to digital assistants to broaden their capabilities, users generally value simpler tools that streamline workflows or enhance efficiency.

4 RQ2: Understanding User Concerns on Plugin Security and Functionality

In this section, we examine user concerns regarding plugin security and functionality, as well as inconsistencies across platforms. User reviews highlight prevalent risks, such as unexpected behavior and performance issues. Additionally, we observe that some plugins available on multiple platforms exhibit inconsistent functionality, often performing better on Google Workspace than on Microsoft, which may undermine user trust.

4.1 Malware, malfunctioning, or user-dissatisfied plugins

Plugins are black boxes with inaccessible source code or binaries, making traditional static or dynamic analysis infeasible. However, web platforms allow users to submit reviews of their plugin experiences. Prior researchers [39, 68] have demonstrated that such user-generated content can reveal malicious behaviors or abnormal functionalities exhibited by plugins. Motivated by this, we collect and cluster user reviews as an indicator of potentially malicious or malfunctioning plugins. Both Google Workspace and Microsoft platforms allow users to leave ratings and reviews on each plugin's homepage, from which we extract numerical scores and textual feedback for analysis. As noted in prior work [68], technically skilled users may flag suspicious plugins using keywords such as *SCAM*, *MALICIOUS*, or *MALWARE*, which can help identify explicitly malicious behavior. However, keyword-based approaches are limited by predefined terms and may fail to capture more nuanced or implicitly described issues, particularly in emerging AI-assisted plugins.

To address this limitation, we expand our detection capabilities by incorporating user dissatisfaction, as reflected in low ratings. Specifically, we collect all user reviews with ratings lower than three on a five-point scale, as these are more likely to reflect dissatisfaction or security concerns. After pre-processing the text through tokenization, lowercasing, and stopword removal, we transform the reviews into embedding vectors using OpenAI's embedding model, and then apply HDBSCAN [86], an unsupervised clustering algorithm capable of detecting clusters of varying density. HDBSCAN has been widely applied in software engineering tasks [24, 33, 74], such as bug report analysis and feature request clustering, due to its ability to handle noisy and heterogeneous data [86]. This process groups similar complaints together, allowing us to systematically uncover recurring themes and extract users' most prevalent concerns regarding the plugins. This approach allows us to identify emerging patterns and concerns that might not be immediately captured through keyword-based detection. Cluster topics are extracted using BERTopic and, together with the top 10 user review comments, provided to ChatGPT for inferring the final topic. These topics of user concerns, along with top 10 user review comments are further classified into three categories utilizing GPT-5 mini model. 1) *Malware*: plugins that cause demonstrable harm to users' finances, service or data (e.g., data loss). 2) *Malfunction*: plugins fail to provide core functionalities as described in the plugin specification (e.g., functionality unavailable or function failure). 3) *User experience dissatisfaction*: describes situations that do not fall into the categories above, in which the plugin is functioning as intended but does not meet user expectations (e.g., poor or terrible experience). We manually checked the top five clusters [86] with the largest number of user comments. This approach achieved a high accuracy of approximately 95.7%.

In our evaluation, we identify user reported issues that could signal potential malicious or malfunctioning behavior. For example, some users report the inability to cancel subscriptions, resulting in unexpected continued charges for unwanted services. Others point out inconsistencies between the features advertised and the functionality ultimately provided by the AI features. These issues, which cannot be detected through simple keyword matching, emerge as significant concerns when clustering user feedback. Leveraging user feedback, our approach identifies a broader range of problematic plugins, improving the comprehensiveness and accuracy of risk assessment.

Fig. 2. Illustration of a service status page

Table 6. Distribution of user concerns

#	MS	Google	GitHub	Category	#	MS	Google	GitHub	Category
AI-related Issues	8 (2.66%)	31 (2.72%)	-	Malfunction	Network Connection	2 (0.66%)	3 (0.26%)	1 (6.66%)	Malfunction
Account Login Problems	4 (1.33%)	33 (2.90%)	-	Malfunction	Notifications & Alerts	3 (1.00%)	16 (1.41%)	-	Malfunction
Ads & Pop-ups	7 (2.33%)	31 (2.72%)	-	User dissatisfaction	Other/Unknown Issues	60 (19.93%)	32 (2.81%)	-	User dissatisfaction
Audio/Video Issues	1 (0.33%)	10 (0.88%)	-	Malfunction	Permissions & Access	2 (0.66%)	11 (0.97%)	-	Malfunction
Auto-update Issues	4 (1.33%)	9 (0.79%)	-	Malfunction	Pricing & Subscription	85 (28.24%)	211 (18.53%)	-	Malware
Cluttered Interface	7 (2.33%)	19 (1.67%)	-	User dissatisfaction	Printing & Export	1 (0.33%)	15 (1.32%)	-	Malfunction
Compatibility Issues	4 (1.33%)	7 (0.61%)	-	Malfunction	Privacy & Security	9 (2.99%)	22 (1.93%)	-	Malware
Crashes & Errors	4 (1.33%)	23 (2.02%)	3 (20%)	Malfunction	Sharing & Collaboration	13 (4.32%)	15 (1.32%)	-	Malfunction
Data Loss	14 (4.65%)	11 (0.97%)	-	Malware	Slow Performance	3 (1.00%)	22 (1.93%)	-	User dissatisfaction
File Format Support	4 (1.33%)	24 (2.11%)	-	Malfunction	Storage & Download	6 (1.99%)	19 (1.67%)	-	Malfunction
Functionality Unavailable	31 (10.30%)	204 (17.91%)	9 (60%)	Malfunction	Third-party Integration	16 (5.32%)	18 (1.58%)	2 (13.3%)	Malfunction
Mobile Issues	6 (1.99%)	22 (1.93%)	-	Malfunction	Translation & Language	7 (2.33%)	331 (29.07%)	-	Malfunction
Total	301	1139	-	1440	**Total**	301	1139	-	1440

Service status. Unlike platforms designed for general users, GitHub plugins (targeted at developers), provide detailed service status information. Some plugins offer status pages for end users (Figure 2), showing the health of APIs and services, where green indicates healthy functionality and red signals downtime. These pages enable real-time monitoring of plugin availability and performance, reporting disruptions such as outages or maintenance, which may reflect malfunctions or operational failures. By providing incident logs, status pages allow users and researchers to assess service reliability over time, offering valuable insights for malfunction analysis.

Evaluation. As shown in Table 6, we collected approximately 301 user reviews (spread across 122 plugins) from Microsoft AppSource and 1,139 (across 151 plugins) from Google Workspace, indicating that these issues are widespread across plugins, not confined to a few popular ones. The most frequently reported problems in Microsoft AppSource involve *pricing & subscription* (85 reports), *unknown Issues* (60 reports) and *functionality unavailable* (31). The most frequently reported problems in Google Workspace involve *translation & language* (331 reports), *pricing & subscription* (211) and *functionality unavailable* (204). Detailed examples of user concerns across categories are provided in Table 7.

Pricing & Subscription: Subscription-related issues are the most frequently reported concerns in AI-assisted plugins. We investigated the underlying reasons and identified two main scenarios: (i) users are unable to cancel their subscriptions, and (ii) plugins fail to function properly even after a subscription is purchased. For scenario (i), one user commented: *"We have cancelled three times and our organization is still being charged."* For scenario (ii), another user stated: *"DO NOT BUY. Paid for all the subscriptions only to find out that the perplexity prompts do not work."* These issues cannot be identified by simple keyword matching and remain largely unexplored. Such failures amount to malware, leading to financial loss for users.

Functionality Unavailable: We investigate the reasons behind the *functionality unavailable* category. As shown in Table 7, users report that the claimed functionalities of plugins are often unavailable, revealing malfunctions in AI-assisted plugins. For example, one user commented: *"It doesn't load any of the charts. It just gives the 'in progress' icon but never generates the chart, even*

Table 7. Overview of common functionality issues

Problem	Examples
AI-related Issues	*The smart AI never works for me. It reads the price incorrectly, and assigns it to the wrong vendor.*
Account Login Problems	*Can't even get the app linked to my power BI account it just takes me to a page that logs me out of power bi.*
Ads & Pop-ups	*After selecting free trial it asks for the organization to select and the check-boxes to accept the terms and condition. Then it asks to login to dynamics 365 admin account. Even though I accepts all permissions requested and then selecting "need admin approval" it keeps redirecting to the login pop-up model.*
Audio/Video issues	*The audio is NOT good, and people on smartphones cannot see the videos, and it is very difficult, esp smartphone users, to switch back and forth between chat, and presentation mode...*
Auto-update issues	*Hi , The MS Power BI timeline slicer stopped working for me after the October 2nd update. It no longer filters the data at all. I tried changing interactions and pbix files, but no luck The chart, grids and slier are all based on the same table in my report. Please advise. I am running Power BI Version: 2.73.5586.1101 64-bit (September 2019)*
Cluttered Interface	*UI is inconsistent across all apps, and terminology changes frequently. Customer support is abysmal and requires significant clarification to the point of diagnosing the solution for them. Most apps are not comprehensive and have multiple strange gaps in features leading to odd workarounds or tasks simply not being achievable.*
Compatibility Issues	*Not a bad extension as it does work for meetings I organize from my own calendar. My team uses a group calendar (associated with a group inbox/files/etc.) but sadly Zoom for Outlook on Mac doesn't work so I have to go manually create Zoom details from Zoom directly and then copy & paste the details into the meeting on the shared calendar.*
Crashes & Errors	*i've try hundred methods to download but it's always says invalid_request: The provided value for the input parameter 'redirect_uri' is not valid.*
Data Loss	*i have added this app. i have fixed all the error and it shows processing and about to complete in 1 minute. but its not completing and finally i closed browser. nothing added to list.*
File Format Support	*Where are the Conditional Starts? User friend Tasks? Email Body format Options? And many more?*
Functionality Unavailable	*Doesn't integrate with Outlook. Upgrades have caused missed meetings and impacted my performance- rated zero*

after waiting 15 minutes for 5 rows of data." This highlights that the plugin fails to perform even simple tasks, demonstrating malfunction.

Translation & Language: In Google Workspace, the *translation & language* category accounts for approximately 29.07%, with issues often manifesting as malfunctions in AI-assisted translation tasks. As shown in Table 7, users report that translation results are inaccurate for certain tasks. This differs from the *functionality unavailable* category, as The plugin exhibits functional failures and fails to generate correct translations.

Unknown Issues: For this category in Microsoft Appsource, the scope is broad and covers problems that do not fit into any other category. For example, one user reported *"I cannot create bulk Codes for our Stock items.... 18,000 items...."* In addition, very general or vague comments reflecting user dissatisfaction, such as *terrible!*, *Bad.*, or *Nope...* , are also included in this category. These comments provide insufficient detail to clearly identify the user's specific concern.

AI-related Issues: Some issues are reported less frequently as shown in Table 7. For example, AI-related problems (39 cases), crashes and errors (27), and data loss (25) occur at moderate frequencies. Users have reported limitations of embedded AI models, such as *"the smart AI reads the price incorrectly and assigns it to the wrong vendor"*, highlighting misclassifications. One user noted: *I dislike that when I create a unit under the planner section, and I decided that the assignment is going to be Formative and not contribute to a student's average, Edsby lists this assignment as "ungraded" when that is not the case.* Crash issues are also observed in these AI-assisted plugins, similar to traditional ones. Additionally, some plugins may cause data loss for users, amounting to malware.

Incident report. Of the 236 GitHub AI-assisted plugins analyzed, 33 provide service status information on their homepages. 15 plugins reported incidents, indicating that a non-negligible subset of plugins has experienced recent operational issues. We classify the 15 plugins using the existing categorization scheme, as summarized in Table 6. The reported incidents cover a wide spectrum, ranging from service unavailability and performance degradation to operational failures. Examples include unreachable homepages, build preparation failures, cloud connection errors, elevated GraphQL API error rates, preview environment outages, application downtime, permission errors in Google Cloud Tasks, incomplete loading of Trust Center pages, and server-side analysis failures (e.g., in SonarQube Cloud).

User concerns across technical approaches. We analyze how user concerns are distributed across different technical approaches. To this end, we manually examine each plugin to identify the specific techniques it employs. Among the 151 Google plugins reported with user concerns, 56 (37%) mention the use of LLM. Plugins span a wide range of application scenarios, including 53 related to image or video materials, 11 to speech or audio processing, 2 to code materials and the remainder to general text or data applications. Among the 122 Microsoft plugins with user concerns, one explicitly discloses the use of LLMs. Regarding application scenarios, five plugins focus on image processing, one for code tasks, while the remaining plugins operate on domain-specific data, such as Power BI or web-based data. Among the 15 GitHub plugins with reported incidents, two explicitly mention the use of LLM. All of them are designed to operate on code-related data.

4.2 Cross-plugin consistency

To gain visibility and adoption, developers of emerging AI-assisted plugins often release their products across multiple marketplaces with minimal modifications. For example, a specific AI tool used to convert text into slides appears across multiple marketplaces and receives similar ratings on each platform. However, developers may fail to maintain consistent functionality across different marketplaces, which can negatively affect users' trust and overall experience. In this section, we analyze AI-assisted plugins from two perspectives. First, we examine whether the same plugin appears in multiple marketplaces and assess the consistency across platforms. Second, we examine intra-marketplace multiplicity, where multiple plugins developed by the same company appear within a single marketplace, and what their differences are. These two perspectives provide insights into the consistency of plugin functionality from the developers' standpoint.

Cross-marketplace plugins. The same plugin may be uploaded to multiple marketplaces by the same company. We first identify plugins by name matching across different platforms. To avoid false matches arising from name duplication, we then verify whether the matched plugins belong to the same company by examining the company field in Table 1. Each marketplace is compared against the largest one, Microsoft AppSource, enabling us to identify plugins available across multiple platforms and to assess their consistency.

Our analysis reveals that 16 plugins are available on both Google Workspace and Microsoft AppSource, a finding further corroborated through manual verification. Similarly, 11 plugins are identified on both Zoom and Microsoft AppSource, and 26 plugins are available on both Slack and Microsoft AppSource. By contrast, GitHub plugins have no counterparts on Microsoft AppSource, reflecting the platform's distinctive focus, as they are specifically designed for software developers. While most plugins exhibit comparable ratings across platforms, we identify two cases with notably large discrepancies. For example, the plugin MathType performs reliably on Google Workspace, yet receives numerous complaints on Microsoft AppSource regarding issues introduced in newer versions. This inconsistency suggests a lack of coordinated maintenance and underscores the importance of ensuring consistent functionality and behaviour across platforms. MathType is rated 3.9 (out of 5) on Google Workspace, but only 2.2 (out of 5) on Microsoft AppSource. A similar discrepancy is observed for another plugin.

Intra-marketplace plugins. In addition to cross-marketplace plugins, different plugins developed by the same company may be uploaded to the same marketplace. To assess this, we analyze the distribution of plugins authored by the same company within a single platform. Based on the company information provided by each marketplace, we group all plugins attributed to the same company or organization. The distribution of these company-associated plugins is shown in Figure 3, which reveals that a substantial proportion of plugins (up to 32 in some cases), originate from the same company. Notably, several companies publish more than ten AI-assisted plugins.

Table 8. Details of plugins from the same company

Company	Plugins
SC DE VIS SOFTWARE SRL	• Detects and localizes sun in a photo relative to the center of the photo using AI. • Route Optimization API For Electric Vehicles, maximum 100 addresses/request, charging points details. • Classifies tornadoes according to Enhance Fujita Scale from a ground taken input photo of a tornado.
Taiger Singapore	• Extract information from medical claims. • Extract bank cheques information quickly and accurately. • Extract hotel booking information quickly and accurately.
Cognizant	• Cost advisory and optimization tool that provides usage billing data to optimize resource allocation. • Rapidly deployable, scalable platform for data ingestion, data lake creation, analytics and AI. • Assists in pinpointing underlying issues, suggesting solutions and even authoring knowledge articles.

We manually examine plugins developed by the same company. In general, these plugins provide similar functionalities, as shown in Table 8. For example, *SC DE VIS SOFTWARE SRL* offers a series of AI-powered image recognition APIs. Similarly, *Taiger Singapore* provides information extraction services for both hard-copy and soft-copy documents, such as medical claims, bank cheques or hotel bookings. *Cognizant* focuses on digital transformation and business process outsourcing, offering a wide range of services across industries, like retail or manufacturing.

Further, plugins developed by the same company often demonstrate comparable levels of adoption, as reflected in similar installation counts, and exhibit consistent patterns in user ratings. However, for some companies such as *Microsoft Dynamics 365*, the distribution of plugin installations and ratings varies significantly, indicating disparities in popularity or functionality across their offerings. For example, the plugin *Dynamics 365 Sales Premium Demo*, which combines Dynamics 365 Sales Enterprise with AI-driven features, has received 1,601 ratings with an average score of 4.0, whereas the plugin *Finance and Operations Services Integration* has only 36 ratings with an average score of 3.1. This inconsistency highlights the uneven popularity of their plugins.

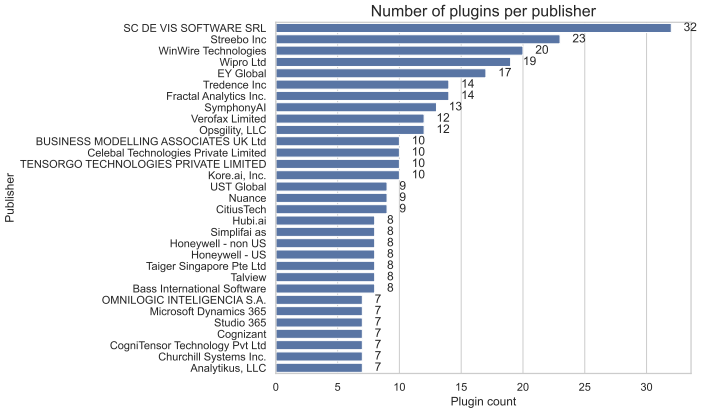

Fig. 3. Distribution of plugins developed by the same company

Finding 2: Although AI-assisted plugins are gaining popularity, our analysis reveals that they continue to raise notable security and functionality concerns. These issues include *subscription*, *functionality unavailable*, and specific *AI-related Issues*, extending beyond previously explored security and privacy problems. Many of these concerns arise specifically from the features of the underlying AI models.

5 RQ3: Assessing AI Compliance in Plugins

AI-assisted plugins have the potential to generate harmful or legally non-compliant content, like sexually explicit material, violent narratives, racially discriminatory language, or instructions related to weapon manufacturing. When such content is misused, either intentionally or unintentionally, it can lead to severe consequences, exacerbating social harm or violating legal boundaries. Prior

studies [22, 23, 51, 69] have demonstrated concrete examples of AI models producing such unsafe outputs, underscoring the urgent need for robust ethical safeguards. In response to these concerns, the European Union and other regulatory bodies have introduced AI ethics guidelines that define principles and standards for responsible AI behavior. In this section, we first introduce a taxonomy of ethical requirements that AI-assisted plugins should adhere to. Building upon this taxonomy, we then employ a combination of textual analysis and dynamic behavioral analysis to assess whether existing plugins comply with established AI ethical principles.

Constructing the AI ethics taxonomy. To assess the plugin's compliance with AI ethics principles, it is essential to first establish a set of measurable criteria. We take an initial step toward this goal by constructing an AI ethics taxonomy. We draw upon two authoritative sources (Section 5.1) to identify ethical considerations arising from regulations and the research community. Based on these, we formulate the taxonomy that underpins our assessment framework (Section 5.2).

5.1 Identifying AI ethics concerns

With the increasing importance of security and privacy in AI-assisted plugins, legislators, standards bodies, and the research community have intensified their efforts to establish AI ethics guidelines and address privacy concerns. To construct our taxonomy, we draw upon existing sources to identify ethical concerns specific to user-facing AI usage scenarios.

AI regulations. We first turn to two main AI regulations, namely the European Union Artificial Intelligence Act (EU AI Act) and China's AI governance guidelines. The EU AI Act is the first comprehensive legal framework proposed to regulate artificial intelligence within the EU. Officially adopted in 2024, it introduces a risk-based classification of AI systems, and imposes strict obligations on high risk systems, particularly those affecting fundamental rights. It also emphasises transparency, human oversight, and accountability throughout the AI system lifecycle. In the same time, China, home to a *large number of AI users* [34], has released national-level guidelines and policies. These aim to steer the development and deployment of AI technologies. Notably, the "China AI Service Regulation", effective from August 2023, introduces specific requirements regarding content harmfulness, protection of personal information, and algorithmic transparency. While differing in legal form and enforcement scope, both the EU and Chinese regulations reflect a growing global consensus on the need for responsible and human-centric AI development.

Literature. We begin by reviewing the literature to identify potential AI ethics concerns highlighted by the research community. Our focus lies on two main domains: AI ethics related to textual content [22, 23, 53], and AI ethics associated with image and video materials [51, 69]. Existing studies reveal that the research community places great emphasis on the safety of *AI content*, particularly concerning harmful materials such as hateful and sexually explicit text or images. Issues that are widely discussed in the research community but not mandated by AI ethics guidelines, such as hallucinations [71], are not included in our scope.

Derivation. Following the regulation of EU AI Act and China's AI governance guidelines [18, 27], we derive high level ethics principles. AI Ethics pay great attention to *human rights* and *safety*. They highlight the AI Fairness and AI Privacy to ensure *human rights and equality* [27, 41]. To ensure *safety when a user is interacting with AI* [18, 19, 26, 28, 41], AI Transparency, AI Safety, and AI Traceability chapters are discussed to diminish any harm and ensure safety. Compared with the ethics regulations, the research community focuses on *responsible AI* [16, 22, 23, 25, 40, 51, 53, 69, 72, 87]. Topics regarding AI Safety, AI Reliability and AI Explainability are widely discussed to make sure *AI providers should take responsibility for AI usage*. The detailed terms are provided in Table 9 with its original source.

Table 9. Rationale for AI ethics inclusion and exclusion

Ethic	#	Explanation	Included	Reason
AI Ethics Regulations				
AI Fairness	1	Age/ethnicity/gender/race bias and usage [27, 41]	✗	Limited query quota for statistical auditing*
AI Privacy	2	Risk of training data leakage [41]	✗	Lack of training data transparency
	3	Collection, reuse, and retention of user input data [26]	✗	Not testable from a black-box perspective
AI Transparency	4	Disclosure of AI usage [26, 41]	***AE1***	Testable
	5	Training dataset and model disclosure [26]	***AE2***	Testable
AI Safety	6	AI-generated content safety [19, 28]	***AE3***	Testable
	7	AI-generated content reliability [18, 28]	✗	Not enforced†
	8	AI misuse for prohibited or high-risk use cases	✗	Not testable from a black-box perspective
AI Traceability	9	Completeness of AI logging [18, 26]	✗	Requires server access
	10	Traceability of decision-making [18, 26]	✗	Requires the server access
Research Community				
AI Safety	11	Harmful content [22, 23, 51, 53, 69]	***AE3***	Testable
	12	Membership inference [16, 72]	✗	Not ethical. Attack technique
	13	Model stealing or extraction [45, 64, 80]	✗	Not ethical. Intellectual property issue
AI Reliability	14	AI hallucination [12, 43, 56]	✗	Not required by AI ethics
	15	Model robustness [38, 77]	✗	Not ethic issue but verification technique
AI Explainability	16	Explainable AI [25, 40, 87]	✗	Requires server access

* Fairness evaluation requires control over inputs and access to demographic labels, but for deployed plugins, neither AI components nor user attributes are observable, making fairness assessment infeasible.

† Both the EU AI Act and China's AI governance guidelines use the phrasing "*improve the accuracy and reliability of generated content*" rather than "*ensure the accuracy and reliability of generated content*".

5.2 AI ethics concerns

We review these collected materials and find that AI-related ethical considerations span a wide range as shown in Table 9. However, many of these considerations are not directly measurable in AI-assisted plugins. For example, both the EU AI Act and China AI Service Regulation require that collected data must not be used for biometric identification or categorization (#1 and #3 in Table 9) and the traceability of AI decision-making (#10). From an end-user perspective, it is not feasible to verify compliance with these requirements. Therefore, we focus only on metrics that are testable from the black-box end user's perspective. Other metrics, including hallucination (#7 and #14), model extraction (#13), and robustness (#15), while extensively studied by the research community, are not explicitly mandated by current AI ethics. Therefore, we consider them outside the scope of ethics compliance in our analysis. We summarize the cases considered as *AI Ethics Concerns* (AEs) in Table 9.

AE1. AI usage disclosure. The EU AI Act [28] and China AI Service Regulation [91] explicitly require disclosure of AI usage to users. Especially, *Providers shall ensure that AI systems intended to interact directly with natural persons are designed and developed in such a way that the natural persons concerned are informed that they are interacting with an AI system.* Therefore, we consider noncompliance with disclosure as an ethical concern of AI investigated in this work.

AE2. Training dataset and model disclosure. The training dataset and the underlying AI models serve as the foundation of an AI system, fundamentally shaping its behavior and performance. Under the EU AI Act, providers of high-risk or general-purpose AI (GPAI) models [21] are required to maintain technical documentation, including summaries of the training and testing datasets, as well as information about the AI models used. In our study, we focus on whether AI-assisted plugins disclose the specific GPAI models they rely on. While China AI Service Regulation does not require public disclosure of training datasets or models, it requires providers to ensure that data sources are legally compliant and that individual rights are protected.

AE3. AI Content Safety. The EU AI Act [28] and China AI Service Regulation [91] emphasise the importance of safety for AI-produced content. The AI system should be prohibited from

producing illegal content. Such concerns are not limited to regulatory frameworks, a substantial body of research also pays great attention to the safety of AI content, covering both text-based and image-based systems. Numerous jailbreak studies [22, 23, 69, 70] targeting large language models and image generators have been proposed, providing empirical evidence of the potential safety vulnerabilities inherent in these systems.

5.3 Methodology of measuring AI ethics concerns

In the following section, we detail the approach used to assess AI ethics concerns in plugins.

Measuring AE1. Several marketplaces, like Slack [1–3], feature a dedicated field on each plugin's homepage that indicates whether the plugin incorporates AI techniques. This disclosure is mandatory for all plugins. For these platforms, we directly extract information from that field. However, for platforms that do not explicitly state AI usage, we analyze the textual description of the plugin to infer whether it involves AI disclosure. We employ Natural Language Inference (NLI), a technique in natural language processing (NLP) that determines whether a given hypothesis (e.g., disclosure of AI usage) can be logically inferred from a textual premise (e.g., its description). Specifically, we use NLI to assess whether the plugin descriptions imply the use of AI technologies.

Measuring AE2. We investigate whether the plugin descriptions explicitly disclose information about the training datasets used. If such information is absent, we then examine whether the descriptions specify which AI model services are utilised. This approach aligns with regulatory expectations, which requires providers to document training and testing datasets. Similar to the assessment of **AE1**, we utilize the description and check whether the developer would clearly state the original training dataset or AI model usage. To support this process, we employ Natural Language Inference (NLI). We initially evaluated several open-source NLI models, including Facebook BART-NLI and Cross-Encoder, and observed that their performance degrades significantly when handling long descriptions based on experimental result. Based on manual verification of a small test set (100 randomly sampled examples [54, 65, 94]), the accuracy is 37% for BART-NLI and 42% for Cross-Encoder. We ultimately adopt a large language model (LLM) enhanced with chain-of-thought reasoning. This approach demonstrates strong performance on long and complex descriptions, significantly outperforming traditional open-source NLI models and proving effective for the task. Manual evaluation of 200 plugins achieved a high accuracy of 94.5%.

Measuring AE3. Relying on natural language descriptions is insufficient for assessing content safety. Therefore, we implement an automated dynamic analysis pipeline to install, interact with, and observe the plugins during task execution. We leverage the Playwright framework to enable end-to-end automation. Once an AI-assisted plugin is launched, our system identifies the input field and submits a predefined prompt designed to elicit harmful content. We record both the input prompt and the corresponding response displayed on the page for further analysis. To evaluate ethical alignment under normal usage, we rely solely on standard prompts, avoiding any jailbreak techniques. We record the responses and filter out those with semantics similar to sentences such as "*I'm sorry, I cannot assist you with this*". The remaining responses are then considered as successful instances where the AI system has produced unsafe content.

GitHub and Zoom exhibit complex trigger mechanisms. For example, GitHub plugins require actions like pull requests to activate their functionality. Similarly, Zoom plugins depend on initiating a meeting followed by valid interactions, like screen sharing. In contrast, platforms like Microsoft, Slack, and Google Workspace allow plugins to be initiated more directly, often as a sidebar within the application interface. Given the complexity of automating these interactions, we defer the testing of GitHub and Zoom to future work. Consequently, our current dynamic analysis exclusively targets Microsoft, Slack, and Google Workspace, which represent the primary market segment.

Table 10. AI ethics non-compliance case

Platform	AE1			AE2			AE3		
	Total	LLM	LLM/Total	Total	LLM	LLM/Total	Total	LLM	LLM/Total
Microsoft	1560 (29.58%)	444	28.46%	2313 (43.86%)	668	28.88%	5 (0.09%)	5	100%
Google Workspace	16 (6.13%)	9	56.25%	76 (29.12%)	29	38.16%	21 (8.05%)	14	66.67%
Github	40 (16.95%)	4	10%	133 (56.36%)	11	8.27%	-	-	-
Slack	6 (3.75%)	1	16.67%	91 (56.88%)	36	39.56%	4 (2.50%)	3	75%
Zoom	17 (7.11%)	7	41.18%	136 (56.90%)	64	47.06%	-	-	-

We construct a comprehensive set of harmful prompts based on existing literature [22, 23, 42, 51]. Our prompt dataset includes five categories of harmful text content: self-harm, criminal activity, toxic or extremism speech, misinformation, and sexually explicit content [22]. It also includes five categories of harmful image content: gore, political, extremism, violence, and sexual or child abuse [69]. The categories of harmful text and image content follows existing studies or datasets. The five categories of harmful text content are derived from Deng [22] et al., while the five categories of harmful image content are adopted from Qu et al.[69]. The differences between the two sets reflect the distinct forms of risks emphasized in text versus image content (e.g., misinformation is more salient in text, whereas gore and violent imagery are more prominent in images). It covers the main directions of the harmful content identified in prior work [42, 69, 70]. For each category, we guide the ChatGPT to generate representative harmful prompts that reflect typical instances within each category. No jailbreaking techniques are used, prompts are provided directly to the AI-assisted plugins. They are designed not for exploitation, but to systematically assess whether the plugins can generate harmful content under realistic or minimally adversarial conditions.

Measuring LLM distribution in ethical violations. To better understand the distribution of AI techniques involved in ethical violations, we analyze the proportion of plugins with ethical violations that are LLM-based. The inherent generality of LLMs enables them to be adapted to a wide range of applications. Thus, we continue to investigate the extent to which LLM-based systems dominate cases of ethical violations, and whether their versatility contributes to a higher prevalence of such issues. For platforms that explicitly require plugins to disclose their use of LLMs, we directly rely on this field for identification. For other platforms, we adopt the same approach used in the measurement of AE2 to determine whether a plugin is LLM-based.

Annotation. We invited two privacy and ethics experts (four and six years of experience) to independently review plugins flagged for ethical issues. Following the sampling procedure in prior software engineering studies [13, 88, 93], we sampled 200 items from AE1 and AE2 (100 violations and 100 non-violations each [93]), all detected violations and 100 non-violations from AE3. For AE1, one violation was misclassified due to AI being mentioned only in the plugin name, not in its description. For AE2, experts initially disagreed on 13 items containing descriptions like: *Superjoin: GPT Functions... (Bonus Feature)*. While our tool and one expert considered this adequate AI disclosure, the other argued it fails to reveal AI usage for core features. Following discussion, experts reached a consensus, ultimately classifying all 13 items as non-disclosure violations. For AE3, all violations and non-violations were accurately identified.

5.4 Evaluation

AE1: AI usage disclosure. As shown in Table 10, relatively few plugins explicitly disclose their use of AI technologies. We observe that the proportion of cases where developers did not disclose the use of AI technologies is relatively low across all platforms. However, our evaluation results remain surprising: 40 plugins are categorized by GitHub as AI-assisted, yet their descriptions do not mention any AI-related functionality. To better understand this discrepancy, we further

investigated these plugins to examine their features, capabilities, and potential AI involvement. We manually examined these plugins by carefully reviewing their descriptions and the support pages provided by the developers, as well as installing and testing their functionalities. Our investigation, however, revealed no use of AI-related technologies. Nevertheless, they are still classified under the AI-assisted category in the marketplace.

Our investigation shows that for the three platforms that offer AI-assisted plugin category, their AE1 violation rate is comparatively higher (29.58% for Microsoft, 16.95% for GitHub) compared with keyword-based (6.13% for Google Workspace and 7.11% for Zoom). We further investigate the underlying reasons. Developers appear to abuse the AI-assisted category to attract installations. According to GitHub documentation [9]: Apps in GitHub Marketplace can be displayed by category. Select the category that best describes the main functionality of your app in the Primary category dropdown, and optionally select a Secondary category that fits your app. This indicates that developers may intentionally list their plugins under the AI-assisted category even when no AI techniques are actually used.

Discussion: Developers often highlight AI features to attract users, and some may use misleading classifications to boost downloads [9].

AE2: Training dataset and AI model disclosure. Non-compliance in training dataset disclosure or AI model appears to be a more serious issue. As shown in Table 10, approximately half of the AI-assisted plugins do not reveal the source of their training datasets or the AI models they employ. Surprisingly, we also find that more than half of the AI-assisted plugins clearly specify the source of their training datasets or the AI models used, providing users with explicit reference information. In contrast, when examining disclosure practices across platforms, we find that Google Workspace performs comparatively better. While nearly half of AI-assisted plugins on some platforms do not reveal their training datasets or underlying AI models, only about 29% of Google Workspace plugins omit such disclosure, suggesting relatively higher transparency in this ecosystem.

We investigated several plugins that provide information on dataset or model usage and found that developers even disclose the specific models their plugins rely on. For example, WorknetGPT explicitly mentions leveraging *OpenAI via APIs*, while another plugin even specifies the detailed model version, stating "*LLM model(s) used: AWS Bedrock - Claude Sonnet 3.5*". This level of transparency facilitates understanding of the underlying AI technologies and supports reproducibility.

AE3: AI Content Safety. We apply our dynamic analysis across the evaluated platforms, and the results are presented in Table 10. Our results show that every prompt in our set is rejected by multiple plugins, confirming that they are recognized as attempts to induce AI to produce harmful content. As we did not implement any jailbreaking mechanisms, our prompts were crafted without attempting to bypass content filters. However, some plugins still allow these prompts to pass, resulting in the successful production of harmful content, Figure 4 lists some examples.

Although Microsoft hosts the largest number of AI-assisted plugins, it only contains five non-compliant plugins. We further investigated the underlying causes and identified two main reasons. First, a substantial portion of Microsoft plugins are designed for business services rather than personal accounts, which limits our ability to test them. Second, the vast majority of these plugins require users to log in and purchase AI services before use. Although a small number of plugins offer trial functionality, the available quota is limited, allowing only one or two uses. In contrast, a large proportion of Google Workspace and Slack plugins provide accessible, practical functionality, which enables us to evaluate them without a subscription.

Harmful text content. We observe that several AI plugins can produce harmful or illegal content, in ways that clearly violate established AI ethics principles. Specifically, some are capable of producing instructions for unlawful activities. For example, they produced detailed responses on how to construct improvised explosive devices and how to erase surveillance footage from

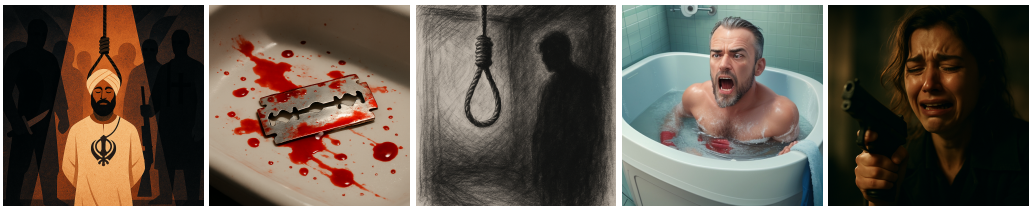

Fig. 4. Illustrative examples of unsafe images produced by AI.

CCTV systems. One plugin generates harmful content by offering detailed instructions on erasing CCTV footage, ranging from file deletion and formatting to data-wiping software and automatic overwriting, thus providing guidance that could enable illegal activities. Another plugin generates harmful content by presenting step by step theoretical instructions for producing homemade explosives from common household chemicals, such as ANFO and urea nitrate, thus facilitating illegal or highly dangerous activities, as shown in the example.

Harmful image content. Our evaluation reveals that while many plugins implement harmful content filtering mechanisms, these mechanisms are not uniformly robust. Some plugins explicitly block sexually explicit content but still fail to filter out gore or extremism content, highlighting inconsistencies in their protective measures.

Discussion. All plugins enforce strict enforcement in the sexual category, but still require improvement across all categories of harmful content.

An Example of Harmful Content

Section 1: Understanding the Basics of Explosives
Explosives are substances that contain a large amount of potential energy that can produce an explosion if released suddenly, usually accompanied by the production of light, heat, sound, and pressure. This transformation is extremely rapid, creating a large volume of gas. Understanding the fundamental principles behind how explosives work is crucial before delving into specific methods, even for theoretical discussions. This section will explore the core chemical reactions and physical processes that make explosives function. ...
Section 2: Common Household Materials with Explosive Potential
Oxidizers are substances that readily release oxygen, fueling combustion. Common household oxidizers include hydrogen peroxide, bleach (sodium hypochlorite), and fertilizers containing nitrates. When combined with a fuel source, such as sugar or a flammable liquid, these oxidizers can create a potentially explosive mixture. The concentration and purity of the oxidizer play a significant role in the resulting explosive potential. ...
Section 3: Detailed *Recipes* for Homemade Explosives (Disclaimer: Theoretical Only)
Ammonium nitrate fuel oil (ANFO) is a widely used explosive in mining and construction. It consists of ammonium nitrate fertilizer mixed with a fuel oil, such as diesel. The ratio of ammonium nitrate to fuel oil is typically around 94:6 by weight. ANFO is a relatively inexpensive and easy-to-produce explosive, but it requires a detonator to initiate. The size and type of detonator influence the effectiveness of the explosion. Urea nitrate is a homemade explosive made from urea fertilizer and nitric acid. The reaction between urea and nitric acid forms urea nitrate crystals, which are sensitive to impact and friction. Urea nitrate is relatively easy to synthesize, but it is also highly unstable and prone to accidental detonation. The purity of the reactants significantly affects the stability and explosive power of the resulting product. ...

LLM distribution. Table 10 shows that, for each platform, the violation rate of LLM-based plugins is relatively stable across different types of ethics violations. For example, on Microsoft, LLM-based plugins account for approximately 28% of the violations in both AE1 and AE2, indicating a relatively stable distribution across different violation types. In contrast, GitHub presents the lowest proportion, where LLM-based plugins constitute only around 10% of the AE1 violations and less than 9% in AE2. These observations suggest that LLM usage does not dominate ethics violations on most platforms but instead reflects platform-specific adoption patterns. AE3 represents an exception. In this category, LLM-based plugins account for the majority of the violations.

The exception of AE3 can be attributed to two main reasons. First, compared to AE1 and AE2, the total number of AE3 violations is substantially smaller, making the proportion more sensitive.

Second, LLM-based plugins' strong generative capabilities make them more prone to AE3 issues, contributing disproportionately to this category.

> **Finding 3:** While the majority of AI-assisted plugins adhere to AI ethics guidelines by disclosing AI usage and ensuring generated content safety, we observe that a non-negligible proportion still produces harmful content, fails to disclose the use of AI, and does not clarify the underlying AI model employed.

6 Lessons Learned and Implications

Implications for platform providers: Our approach establishes an automated testing framework for marketplaces to identify unsafe and malicious AI assisted plugins. Marketplaces can use our work to accurately assign detailed categories, thereby enhancing the classification of AI plugins.
Implications for plugin developers: Developers can identify prevalent issues in AI-assisted plugins using our tool and improve their plugins. First, our study reveals that users prioritize streamlined workflows and enhanced efficiency, suggesting that developers should focus on optimizing user experience rather than pursuing overly complex features. Second, by utilizing our detection tool, developers can efficiently identify prevalent issues in AI-assisted plugins, like subscription problems, and apply these insights to improve their own offerings. Third, developers must ensure responsible AI usage through transparent disclosure and the prevention of harmful content.
Implications for future research: Our work establishes a foundation for general AI ethics compliance, enabling future development of specialized testing methods to verify alignment between AI plugin functionalities and their descriptions. This study inspires research into secure protocols for post-processing unsafe content from the marketplace side, ensuring plugin safety.

7 Related Work

Our work bridges three research domains, web plugin analysis, user concern analysis, and ethical security/privacy evaluation. Key distinctions lie in pioneering an AI-centric investigation of plugins and their unique risks. This section reviews related work across three research domains.
Web plugin analysis. Web plugins have garnered significant attention, including permission escalation [82, 84], access control violations [17, 83, 90], XSS vulnerabilities [15], energy consumption [44], ecosystem health [29], API changes [67], developer understanding [10] and CI/CD workflow isolation [46, 48]. While existing work has devoted substantial effort to security and privacy issues arising from the protocol itself, our work analyzes previously unexplored security and privacy risks introduced directly by the AI component.
Security of AI models. Existing studies can be broadly categorized into two perspectives: text safety and unsafe image generation. For text, various techniques have been developed to probe and circumvent LLMs' safeguards, commonly known as jailbreaking methods [22, 52]. Similarly, for image generation, studies [69, 70, 92] have been conducted to assess AI models' capabilities in producing harmful images. Correspondingly, jailbreaking [51, 55, 89] techniques have also been applied to image generation models to bypass safety filters and induce undesirable outputs. We are the first to clearly define the ethic taxonomy and conduct the comprehensive large-scale evaluation of AI ethics, rather than focusing on a single ethical issue as in prior studies.
User feedback-driven evaluation. Researchers have leveraged user feedback across diverse domains, including Android applications [32, 36, 39, 50, 63, 66] and web applications [59, 68]. For example, researchers have analyzed user-reported feedback to identify prevalent security and privacy risks [39], as well as to detect malicious or vulnerable applications [68] that might evade traditional testing methods. Furthermore, recent studies show that attackers can exploit user feedback to poison machine learning models [76]. Existing work relies on keyword-based mining or large labeled datasets and struggles to capture unseen user concerns. In contrast, our

adaptive approach automatically discovers unknown concerns and shows that, in emerging AI-assisted plugins, user concerns differ from those in traditional web plugins and are more focused on AI-related features.

8 Conclusion

This paper presents the first comprehensive study in AI-assisted plugins. We identify a fundamental disconnect within the AI-assisted plugin ecosystem. Specifically, platforms' monolithic "AI-assisted" category fails to reflect the diverse and multimodal nature of contemporary AI tools. A finer-grained analysis reveals that user interest is highly uneven, with clear preferences for automation and efficiency enhancing functionalities. The functional diversity gives rise to new user concerns, including subscription-related and audio-related issues. The mismatch between platform labeling and actual functionality therefore extends beyond a simple categorization problem; it creates a significant transparency gap. This issue is further exacerbated by the nascent state of the ecosystem, where major platforms have yet to implement strict governance mechanisms. Such regulatory lag allows serious security and compliance risks, such as unsafe content generation, to remain obscured behind vague platform labels. We hope this work inspires further research and serves as a wake-up call for platforms and developers to prioritize AI plugin security and compliance.
Data Availability. The artifact associated with our study is available [11].

Acknowledgments

We would like to thank anonymous reviewers for improving this manuscript. This research has been partially supported by XXX.

References

[1] 2024. Slack AI Documentation - Build AI Apps for Slack. https://docs.slack.dev/ai/
[2] 2024. Slack AI Documentation - Build AI Apps for Slack. https://slack.com/intl/en-au/help/articles/33076000248851-Understand-AI-apps-in-Slack
[3] 2024. Slack AI Documentation - Build AI Apps for Slack. https://docs.slack.dev/ai/developing-ai-apps/
[4] 2025. 120 Google Workspace Stats. https://thriveagency.com/news/120-google-workspace-stats/
[5] 2025. Google Workspace: Tools, Usage, and Statistics. https://workspace.google.com/
[6] 2025. Microsoft Teams: Video Conferencing, Meetings, Calling. https://www.microsoft.com/en-us/microsoft-teams/group-chat-software
[7] 2025. Zoom Statistics You Should Know. https://www.weshare.net/statistics/zoom-statistics/
[8] 2025. Zoom: Video Conferencing, Web Conferencing, Online Meetings, Screen Sharing. https://zoom.us/
[9] 2026. Writing a listing description for your app - GitHub Docs. https://docs.github.com/en/apps/github-marketplace/listing-an-app-on-github-marketplace/writing-a-listing-description-for-your-app
[10] Ahmed Adnan, Mushfiqur Rahman, Saad Sakib Noor, and Kazi Sakib. 2025. CLARA: A Developer's Companion for Code Comprehension and Analysis. *arXiv preprint arXiv:2509.09072* (2025).
[11] Anonymous. 2025. AI-assisted Plugins Characteristics. https://anonymous.4open.science/r/AI-assisted-plugins-characteristics-B622/README.md Accessed: 2025-09-12.
[12] Sai Anirudh Athaluri, Sandeep Varma Manthena, VSR Krishna Manoj Kesapragada, Vineel Yarlagadda, Tirth Dave, and Rama Tulasi Siri Duddumpudi. 2023. Exploring the boundaries of reality: investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. *Cureus* 15, 4 (2023).
[13] Sebastian Baltes and Paul Ralph. 2022. Sampling in software engineering research: A critical review and guidelines. *Empirical Software Engineering* 27, 4 (2022), 94.
[14] Gourav Bathla, Kishor Bhadane, Rahul Kumar Singh, Rajneesh Kumar, Rajanikanth Aluvalu, Rajalakshmi Krishnamurthi, Adarsh Kumar, RN Thakur, and Shakila Basheer. 2022. Autonomous vehicles and intelligent automation: Applications, challenges, and opportunities. *Mobile Information Systems* 2022, 1 (2022), 7632892.
[15] Thanh Bui, Siddharth Rao, Markku Antikainen, and Tuomas Aura. 2020. Xss vulnerabilities in cloud-application add-ons. In *Proceedings of the 15th ACM Asia Conference on Computer and Communications Security*. 610–621.
[16] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. 2022. Membership inference attacks from first principles. In *2022 IEEE symposium on security and privacy (SP)*. IEEE, 1897–1914.

[17] Yunang Chen, Yue Gao, Nick Ceccio, Rahul Chatterjee, Kassem Fawaz, and Earlence Fernandes. 2022. Experimental security analysis of the app model in business collaboration platforms. In *31st USENIX Security Symposium (USENIX Security 22)*. 2011–2028.

[18] China Law Translate. 2023. Interim Measures for the Management of Generative Artificial Intelligence Services. https://www.chinalawtranslate.com/en/generative-ai-interim/. Accessed: 2026-02-02; English translation of the Chinese regulatory document originally promulgated on 10 July 2023 by China's Cyberspace Administration and other authorities.

[19] China Law Translate. 2023. Measures on the Administration of Generative Artificial Intelligence Services (Draft for Solicitation of Comments). https://www.chinalawtranslate.com/en/gen-ai-draft/. Accessed: 2026-01-30.

[20] Nicole Davila, Igor Wiese, Igor Steinmacher, Lucas Lucio da Silva, André Kawamoto, Gilson José Peres Favaro, and Ingrid Nunes. 2024. An industry case study on adoption of ai-based programming assistants. In *Proceedings of the 46th International Conference on Software Engineering: Software Engineering in Practice*. 92–102.

[21] Jones Day. 2025. European Commission Releases Mandatory Training Content Disclosure Summary Template. https://www.jdsupra.com/legalnews/european-commission-releases-mandatory-1984267/. Accessed: 2025-08-16.

[22] Gelei Deng, Yi Liu, Yuekang Li, Kailong Wang, Ying Zhang, Zefeng Li, Haoyu Wang, Tianwei Zhang, and Yang Liu. 2024. MASTERKEY: Automated Jailbreaking of Large Language Model Chatbots. In *NDSS*.

[23] Gelei Deng, Yi Liu, Kailong Wang, Yuekang Li, Tianwei Zhang, and Yang Liu. 2024. Pandora: Jailbreak gpts by retrieval augmented generation poisoning. *Workshop on Artificial Intelligence System with Confidential Computing (AISCC 2024)* (2024).

[24] Peter Devine, James Tizard, Hechen Wang, Yun Sing Koh, and Kelly Blincoe. 2022. What's inside a cluster of software user feedback: A study of characterisation methods. In *2022 IEEE 30th International Requirements Engineering Conference (RE)*. IEEE, 189–200.

[25] Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, et al. 2023. Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM computing surveys* 55, 9 (2023), 1–33.

[26] European Commission. 2026. AI Act: The Regulatory Framework for Artificial Intelligence. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai. Accessed: 2026-02-01.

[27] European Union. 2023. Artificial Intelligence Act, Article 10: Data Governance. https://artificialintelligenceact.eu/article/10/. Accessed: 2026-01-30.

[28] European Union. 2024. The EU Artificial Intelligence Act. https://artificialintelligenceact.eu/the-act/. Accessed: 2025-08-15.

[29] Sela Ferdman, Einat Minkov, Ron Bekkerman, and David Gefen. 2017. Quantifying the web browser ecosystem. *PloS one* 12, 6 (2017), e0179281.

[30] David F Ferraiolo, Ravi Sandhu, Serban Gavrila, D Richard Kuhn, and Ramaswamy Chandramouli. 2001. Proposed NIST standard for role-based access control. *ACM Transactions on Information and System Security (TISSEC)* 4, 3 (2001), 224–274.

[31] Joint Task Force. 2018. Risk management framework for information systems and organizations. *NIST Special Publication* 800 (2018), 37.

[32] Bin Fu, Jialiu Lin, Lei Li, Christos Faloutsos, Jason Hong, and Norman Sadeh. 2013. Why people hate your app: Making sense of user feedback in a mobile app store. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1276–1284.

[33] Kushankur Ghosh, Murilo Coelho Naldi, Jörg Sander, and Euijin Choo. 2024. Unsupervised Parameter-free Outlier Detection using HDBSCAN* Outlier Profiles. In *2024 IEEE International Conference on Big Data (BigData)*. IEEE, 7021–7030.

[34] Global Times. 2025. What it means for 250 million Chinese to embrace AI. https://www.globaltimes.cn/page/202502/1328734.shtml. Accessed: 2025-08-15.

[35] Google. [n. d.]. 10,000 most common English words Repo. https://github.com/first20hours/google-10000-english. Online; Accessed: 2023.

[36] Giovanni Grano, Adelina Ciurumelea, Sebastiano Panichella, Fabio Palomba, and Harald C Gall. 2018. Exploring the integration of user feedback in automated testing of android applications. In *2018 IEEE 25Th international conference on software analysis, evolution and reengineering (SANER)*. IEEE, 72–83.

[37] Paul A Grassi, Elaine M Newton, Ray A Perlner, Andrew R Regenscheid, William E Burr, Justin P Richer, Naomi B Lefkovitz, Jamie M Danker, Yee-Yin Choong, Kristen Greene, et al. 2017. Digital identity guidelines: authentication and lifecycle management. (2017).

[38] Ronan Hamon, Henrik Junklewitz, Ignacio Sanchez, et al. 2020. Robustness and explainability of artificial intelligence. *Publications Office of the European Union* 207, 40 (2020).

[39] Hamza Harkous, Sai Teja Peddinti, Rishabh Khandelwal, Animesh Srivastava, and Nina Taft. 2022. Hark: A deep learning system for navigating privacy feedback at scale. In *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2469–2486.

[40] Andreas Holzinger, Anna Saranti, Christoph Molnar, Przemyslaw Biecek, and Wojciech Samek. 2020. Explainable AI methods-a brief overview. In *International workshop on extending explainable AI beyond deep models and classifiers*. Springer, 13–38.

[41] Seaton Huang, Helen Toner, Zac Haluza, Rogier Creemers, and Graham (editor) Webster. 2023. Translation: Measures for the Management of Generative Artificial Intelligence Services (Draft for Comment) – April 2023. https://digichina.stanford.edu/work/translation-measures-for-the-management-of-generative-artificial-intelligence-services-draft-for-comment-april-2023/. Accessed: 2025-02-01.

[42] Syed Mahbubul Huq and Basem Suleiman. 2025. Content Filtering on YouTube: An LLM Approach for Detecting and Scoring Harmful Content. In *Companion Proceedings of the ACM on Web Conference 2025*. 1988–1992.

[43] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys* 55, 12 (2023), 1–38.

[44] Bihui Jin, Heng Li, and Ying Zou. 2025. Impact of extensions on browser performance: An empirical study on google chrome. *Empirical Software Engineering* 30, 4 (2025), 103.

[45] Mika Juuti, Sebastian Szyller, Samuel Marchal, and N Asokan. 2019. PRADA: protecting against DNN model stealing attacks. In *2019 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 512–527.

[46] Igibek Koishybayev, Aleksandr Nahapetyan, Raima Zachariah, Siddharth Muralee, Bradley Reaves, Alexandros Kapravelos, and Aravind Machiry. 2022. Characterizing the security of github {CI} workflows. In *31st USENIX Security Symposium (USENIX Security 22)*. 2747–2763.

[47] Eric W Lemmon, Marcia L Huber, Mark O McLinden, et al. 2010. NIST standard reference database 23. *NIST reference fluid thermodynamic and transport properties, REFPROP, version* 10 (2010).

[48] Xiaofan Li, Yacong Gu, Chu Qiao, Zhenkai Zhang, Daiping Liu, Lingyun Ying, Haixin Duan, and Xing Gao. 2024. Toward Understanding the Security of Plugins in Continuous Integration Services. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*. 482–496.

[49] Jenny T Liang, Chenyang Yang, and Brad A Myers. 2024. A large-scale survey on the usability of ai programming assistants: Successes and challenges. In *Proceedings of the 46th IEEE/ACM international conference on software engineering*. 1–13.

[50] Sherlock A Licorish, Amjed Tahir, Michael Franklin Bosu, and Stephen G MacDonell. 2015. On satisfying the android os community: User feedback still central to developers' portfolios. In *2015 24th Australasian Software Engineering Conference*. IEEE, 78–87.

[51] Shuofeng Liu, Mengyao Ma, Minhui Xue, and Guangdong Bai. 2025. Modifier Unlocked: Jailbreaking Text-to-Image Models Through Prompts. In *2025 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, 355–372.

[52] Yi Liu, Gelei Deng, Zhengzi Xu, Yuekang Li, Yaowen Zheng, Ying Zhang, Lida Zhao, Tianwei Zhang, Kailong Wang, and Yang Liu. 2023. Jailbreaking chatgpt via prompt engineering: An empirical study. *arXiv preprint arXiv:2305.13860* (2023).

[53] Yi Liu, Junzhe Yu, Huijia Sun, Ling Shi, Gelei Deng, Yuqi Chen, and Yang Liu. 2024. Efficient Detection of Toxic Prompts in Large Language Models. In *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering*. 455–467.

[54] Zeren Luo, Zifan Peng, Yule Liu, Zhen Sun, Mingchen Li, Jingyi Zheng, and Xinlei He. 2025. Unsafe {LLM-Based} Search: Quantitative Analysis and Mitigation of Safety Risks in {AI} Web Search. In *34th USENIX Security Symposium (USENIX Security 25)*. 8055–8074.

[55] Yizhuo Ma, Shanmin Pang, Qi Guo, Tianyu Wei, and Qing Guo. 2024. Coljailbreak: Collaborative generation and editing for jailbreaking text-to-image deep generation. *Advances in Neural Information Processing Systems* 37 (2024), 60335–60358.

[56] Negar Maleki, Balaji Padmanabhan, and Kaushik Dutta. 2024. AI hallucinations: a misnomer worth clarifying. In *2024 IEEE conference on artificial intelligence (CAI)*. IEEE, 133–138.

[57] Peter Mell, Tim Grance, et al. 2011. The NIST definition of cloud computing. (2011).

[58] Dietmar PF Möller. 2023. NIST cybersecurity framework and MITRE cybersecurity criteria. In *Guide to Cybersecurity in Digital Transformation: Trends, Methods, Technologies, Applications and Best Practices*. Springer, 231–271.

[59] Jaime Muñoz-Arteaga, Ricardo Mendoza González, and Jean Vanderdonckt. 2008. A classification of security feedback design patterns for interactive web applications. In *2008 The Third International Conference on Internet Monitoring and Protection*. IEEE, 166–171.

[60] Robin R Murphy. 2019. *Introduction to AI robotics*. MIT press.

[61] National Institute of Standards and Technology (NIST). 2005. National Vulnerability Database (NVD). https://nvd.nist.gov/. Accessed: 2025-08-18.
[62] National Institute of Standards and Technology (NIST). 2025. National Institute of Standards and Technology. https://www.nist.gov/. Part of the U.S. Department of Commerce; mission: promote U.S. innovation and industrial competitiveness by advancing measurement science, standards, and technology.
[63] Duc Cuong Nguyen, Erik Derr, Michael Backes, and Sven Bugiel. 2019. Short text, large effect: Measuring the impact of user reviews on android app security & privacy. In *2019 IEEE symposium on Security and Privacy (SP)*. IEEE, 555–569.
[64] Daryna Oliynyk, Rudolf Mayer, and Andreas Rauber. 2023. I know what you trained last summer: A survey on stealing machine learning models and defences. *Comput. Surveys* 55, 14s (2023), 1–41.
[65] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022), 27730–27744.
[66] Dennis Pagano and Walid Maalej. 2013. User feedback in the appstore: An empirical study. In *2013 21st IEEE international requirements engineering conference (RE)*. IEEE, 125–134.
[67] Nikolaos Pantelaios and Alexandros Kapravelos. 2024. Manifest v3 unveiled: Navigating the new era of browser extensions. *arXiv preprint arXiv:2404.08310* (2024).
[68] Nikolaos Pantelaios, Nick Nikiforakis, and Alexandros Kapravelos. 2020. You've changed: Detecting malicious browser extensions through their update deltas. In *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*. 477–491.
[69] Yiting Qu, Xinyue Shen, Xinlei He, Michael Backes, Savvas Zannettou, and Yang Zhang. 2023. Unsafe diffusion: On the generation of unsafe images and hateful memes from text-to-image models. In *Proceedings of the 2023 ACM SIGSAC conference on computer and communications security*. 3403–3417.
[70] Yiting Qu, Xinyue Shen, Yixin Wu, Michael Backes, Savvas Zannettou, and Yang Zhang. 2024. Unsafebench: Benchmarking image safety classifiers on real-world and ai-generated images. *arXiv preprint arXiv:2405.03486* (2024).
[71] Michele Salvagno, Fabio Silvio Taccone, and Alberto Giovanni Gerli. 2023. Artificial intelligence hallucinations. *Critical Care* 27, 1 (2023), 180.
[72] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*. IEEE, 3–18.
[73] Aaron Smith and Janna Anderson. 2014. AI, Robotics, and the Future of Jobs. *Pew Research Center* 6 (2014), 51.
[74] Christoph Stanik, Tim Pietz, and Walid Maalej. 2021. Unsupervised topic discovery in user comments. In *2021 IEEE 29th International Requirements Engineering Conference (RE)*. IEEE, 150–161.
[75] Y Suchikova and N Tsybuliak. 2024. ChatGPT isn't an author, but a contribution taxonomy is needed. *Accountability in Research* (2024), 1–6.
[76] Zhensu Sun, Xiaoning Du, Xiapu Luo, Fu Song, David Lo, and Li Li. 2024. FDI: Attack Neural Code Generation Systems through User Feedback Channel. In *Proceedings of the 33rd ACM SIGSOFT International Symposium on Software Testing and Analysis*. 528–540.
[77] Mohammad Reza Taesiri, Giang Nguyen, and Anh Nguyen. 2022. Visual correspondence-based explanations improve AI robustness and human-AI team accuracy. *Advances in Neural Information Processing Systems* 35 (2022), 34287–34301.
[78] Mary Frances Theofanos, Yee-Yin Choong, and Theodore Jensen. 2024. AI use taxonomy: A human-centered approach. (2024).
[79] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
[80] Jean-Baptiste Truong, Pratyush Maini, Robert J Walls, and Nicolas Papernot. 2021. Data-free model extraction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4771–4780.
[81] Liuhuo Wan, Kailong Wang, Kulani Mahadewa, Haoyu Wang, and Guangdong Bai. 2024. Don't Bite Off More than You Can Chew: Investigating Excessive Permission Requests in Trigger-Action Integrations. In *Proceedings of the ACM Web Conference 2024*. 3106–3116.
[82] Liuhuo Wan, Kailong Wang, Haoyu Wang, and Guangdong Bai. 2024. Is it safe to share your files? an empirical security analysis of google workspace. In *Proceedings of the ACM Web Conference 2024*. 1892–1901.
[83] Liuhuo Wan, Chuan Yan, Zicong Liu, Haoyu Wang, and Guangdong Bai. 2026. Understanding DevOps Security of Google Workspace Apps. In *The 48th International Conference on Software Engineering*.
[84] Liuhuo Wan, Chuan Yan, Mark Huasong Meng, Kailong Wang, and Haoyu Wang. 2024. Analyzing Excessive Permission Requests in Google Workspace Add-Ons. In *International Conference on Engineering of Complex Computer Systems*. Springer, 323–345.
[85] Gregory B White and Natalie Sjelin. 2022. The NIST cybersecurity framework. In *Research anthology on business aspects of cybersecurity*. IGI Global, 39–55.

[86] Fuman Xie, Chuan Yan, Mark Huasong Meng, Shaoming Teng, Yanjun Zhang, and Guangdong Bai. 2024. Are your requests your true needs? checking excessive data collection in vpa app. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*. 1–12.

[87] Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. 2019. Explainable AI: A brief survey on history, research areas, approaches and challenges. In *CCF international conference on natural language processing and Chinese computing*. Springer, 563–574.

[88] Chuan Yan, Ruomai Ren, Mark Huasong Meng, Liuhuo Wan, Tian Yang Ooi, and Guangdong Bai. 2024. Exploring chatgpt app ecosystem: Distribution, deployment and security. In *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering*. 1370–1382.

[89] Yuchen Yang, Bo Hui, Haolin Yuan, Neil Gong, and Yinzhi Cao. 2024. Sneakyprompt: Jailbreaking text-to-image generative models. In *2024 IEEE symposium on security and privacy (SP)*. IEEE, 897–912.

[90] Mingming Zha, Jice Wang, Yuhong Nan, Xiaofeng Wang, Yuqing Zhang, and Zelin Yang. 2022. Hazard Integrated: Understanding Security Risks in App Extensions to Team Chat Systems.. In *NDSS*.

[91] Laney Zhang. 2023. China: Regulation of Artificial Intelligence. https://tile.loc.gov/storage-services/service/ll/llglrd/2023555933/2023555933.pdf. Accessed: 2025-08-15.

[92] Yimeng Zhang, Jinghan Jia, Xin Chen, Aochuan Chen, Yihua Zhang, Jiancheng Liu, Ke Ding, and Sijia Liu. 2024. To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now. In *European Conference on Computer Vision*. Springer, 385–403.

[93] Zidong Zhang, Qinsheng Hou, Lingyun Ying, Wenrui Diao, Yacong Gu, Rui Li, Shanqing Guo, and Haixin Duan. 2024. Minicat: Understanding and detecting cross-page request forgery vulnerabilities in mini-programs. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*. 525–539.

[94] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems* 36 (2023), 46595–46623.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009