"I Cannot Write This Because It Violates Our Content Policy": Understanding Content Moderation Policies and User Experiences in Generative AI Products

Lan Gao, Oscar Chen, Rachel Lee, Nick Feamster, Chenhao Tan, Marshini Chetty
University of Chicago

Abstract

While recent research has focused on developing safeguards for generative AI (GAI) model-level content safety, little is known about how content moderation to prevent malicious content performs for end-users in real-world GAI products. To bridge this gap, we investigated content moderation policies and their enforcement in GAI online tools — consumer-facing web-based GAI applications. We first analyzed content moderation policies of 14 GAI online tools. While these policies are comprehensive in outlining moderation practices, they usually lack details on practical implementations and are not specific about how users can aid in moderation or appeal moderation decisions. Next, we examined user-experienced content moderation successes and failures through Reddit discussions on GAI online tools. We found that although moderation systems succeeded in blocking malicious generations pervasively, users frequently experienced frustration in failures of both moderation systems and user support after moderation. Based on these findings, we suggest improvements for content moderation policy and user experiences in real-world GAI products.

1 Introduction

The development of large generative AI (GAI) models, such as large language models (LLMs) and diffusion models, has promoted the production of *AI-Generated Content (AIGC)* — synthetic content, in the form of text, image, audio, video, that are generated by AI models given human instructions.[1] Recently, this has opened up AIGC as a new form for content creation as opposed to human-created content [8, 35, 87, 89].

Simultaneously, the potential harms associated with GAI's generation capabilities, such as producing disturbing, misleading, and privacy/copyright-infringing AIGC [10, 65, 95], have raised attention. Malicious content generated by GAI tools could not only directly compromise the safety of end-users who interact with the system, but also pose security and privacy risks to the public due to its automated and fast generation. In response, AI and Security practitioners have been working on GAI safeguards, such as safety alignments and content filters, to prevent problematic content generation (e.g. [36, 40, 55, 70]). When deploying GAI models into real-world products, service providers also enforce *content moderation* on users' content generation process. Content moderation is a common strategy employed in online communities to reduce problematic user-generated content by detecting and restricting such content and users who publish it [45, 64], and has now been used for GAI products. For example, ChatGPT denied over 250,000 requests for generating images of US political campaigns before the 2024 US election day, to prohibit potential misinformation [11].

Nevertheless, most recent works focus on improving safeguards to GAI content generation safety at the *model level*. Little is known about how and how well content moderation practices are enforced at the *product level* (i.e., real-world GAI products, like ChatGPT). Since GAI is an emerging technology, there has been a lack of legal guidance on how GAI products should engage in content moderation until recently [32], with service providers instead relying on themselves for making policies [47]. Therefore, it is crucial to understand what GAI tool content moderation policies are established by service providers. As seen in online communities [68, 74], content moderation policies reveal how service providers enforce moderation, disclose their practices to the public, and guide user behavior. Meanwhile, given the strong correlation of content moderation with user behavior and experiences [53, 59], user perspectives on content moderation provide valuable insights into how current policy enforcement succeeds or fails in practice. Both processes together could inform the future direction of content moderation policies and practices for GAI products.

We study content moderation policies and their enforcement in consumer-facing GAI products offered through web-based applications (e.g., ChatGPT and its playground, referred

[1] There is no consensus on the definition of AIGC, and therefore the definition we used in this paper is summarized from existing literature (e.g., [8, 89]) and related definitions in law [82].

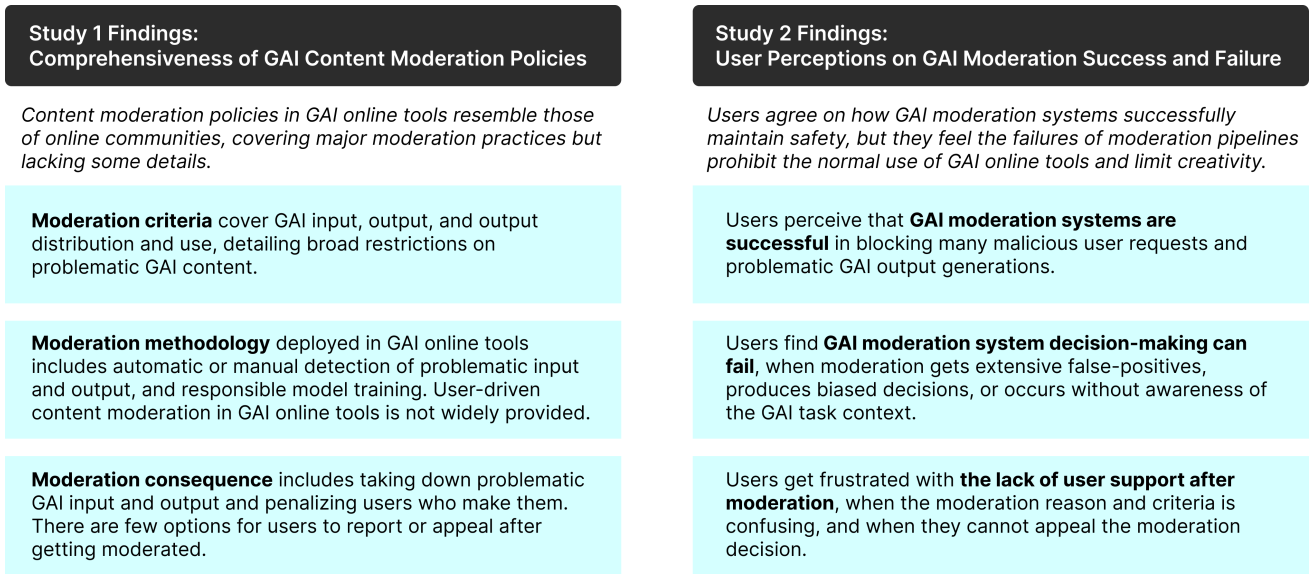

Figure 1: Findings summary of Study 1: Policy Analysis and Study 2: Reddit Study.

to as *GAI online tools* from here on) that enable text and image generation. GAI online tools are widely adopted by end-users regardless of their technical expertise, as compared to locally deployed GAI models and APIs. We focus only on text and image generation tools since they make up many of the GAI online tool markets, despite the growth of video and audio generation tools recently [66, 75]. Through this work, we study the following research questions:

- RQ1 *(Service Provider Policymaking)*: What themes do content moderation policies in GAI online tools cover, and how comprehensive are they?
- RQ2 *(User Experience on Policy Enforcement)*: From user perspectives, in what areas does content moderation policy enforcement for GAI online tools succeed, and in what areas do they fail?

To answer RQ1, we studied the content moderation policies of 14 GAI online tools that were capable of text and image generation (Study 1: Policy Analysis). We found that content moderation policies in GAI online tools are similar to those in online communities [68], with scattered locations in different pages outlining three components of content moderation practices, including moderation criteria on forbidden content, methodology to detect problematic content, and the consequences of problematic content and users. Compared to how online communities rule on user-generated content, policies in GAI online tools focus on governing both user input to the tool and content generated by the tool, with some differences in moderation methodology and consequences for policy violations (See Table 1). We found that GAI online tools are comprehensive in covering major content moderation practices, but these policies lack details on topics such as how users can report problematic content generation and appeal moderation decisions.

To address RQ2, we studied user experiences with content moderation while using GAI online tools to generate new, creative, or crafted AIGC (e.g., requesting ChatGPT to write a poem about cats or using its DALL-E to draw a cat, referred to as *AIGC creative tasks* from here on). Tasks performed through GAI are diverse: LLMs can generate text for dialogue, question-answering, searching, and so on. These tasks result in varied presentations of generated content (e.g., natural speech in dialogues versus structured writing in creative writings), which may affect moderation performance [54] and influence user mental model on content moderation. Therefore, we studied the AIGC creative tasks in particular, which is not only a representative use case of GAI [76], but also heavily depends on GAI's generation capability compared to other usages like searching.

To do so, we analyzed public discussions on content moderation in AIGC creative tasks when using GAI online tools, by looking into user posts and corresponding comments on Reddit covering six different GAI online tools (Study 2: Reddit Study). We found that while user-provided examples highlighted the widespread success of moderation systems in blocking malicious AIGC creation, users also discussed numerous instances of moderation pipeline failures. These failures spanned both moderation systems' shortcomings in making justified moderation decisions, and lack of support for users to understand and appeal moderation decisions. As users posted, failures in the moderation pipeline had severely hampered the capability and usability of GAI online tools, limiting their creativity in AIGC creative tasks as well. See Figure 1 for key findings of the two studies.

Our work provides in-depth insights into the content mod-

eration pipeline in real-world GAI products, a key feature for controlling malicious content generation and user behavior to safeguard security and privacy. Specifically, our work makes the following contributions in AI security, privacy, and safety domains: (1) we provide the first study of GAI products' content moderation policies and existing gaps, (2) we compile two datasets of policies and public discussions on GAI products' content moderation for further studies,[2] and (3) we suggest how to improve policies and user experiences with content moderation when using these products.

2 Related Works

In this section, we review three sets of prior research that are relevant to our study, highlighting how our work extends or contributes to each body of research.

Safeguards in GAI for Ensuring Content Safety. AI and security researchers use two automated methods to ensure the safety of GAI model output: internal model fine-tuning and external content guardrails. Prior works have shown the success of widely adopted reinforcement learning methods for LLMs in providing targeted feedback on output toxicity through fine-grained rewards [40, 71, 90]. Recently, AI alignment, also known as Reinforcement Learning from Human Feedback (RLHF), has made further advancements in ensuring desirable model output. This method evaluates the harmfulness and helpfulness of model output with the guidance of human feedback [41]. Following this principle, researchers have enforced safety alignment in LLMs to provide harmless responses and reject problematic model input [5, 13, 40, 60]. Researchers have also focused on adjusting multi-modal generation models for generating safe output [12, 50, 70, 85].

Aside from internal model fine-tuning, content guardrails are also deployed outside the GAI model to monitor both model input and output. Similar to traditional algorithmic moderation in online communities, some guardrails in GAI apply classifiers trained with categorically labeled content on different harms [55, 61]. While other state-of-the-art guardrails used classifier-free guidance — for example, Llama Guard leverages instruction-tuned LLMs and safety taxonomies for problematic content detection [36, 81, 86].

Through improving both internal and external safeguards, previous research has made significant strides in mitigating undesirable outputs within model architectures. However, there is scarce work investigating how content moderation is enforced with end-users in real-world GAI products, except for several works that audited the content moderation endpoint of GAI products [54, 62] To bridge this gap, our work unpacks content moderation practices by examining GAI tool policies and public discussions where users talk about content moderation successes or failures in GAI products.

[2]The datasets are publicly available at https://doi.org/10.6084/m9.figshare.29257187

Content Moderation Policies in Online Communities and GAI. Much research to date focuses on measuring and analyzing content moderation policies in online communities with user-generated content — studying what types of content are forbidden, and how these rules vary around different platforms [4, 7, 18, 19, 42, 74]. Meanwhile, some researchers also analyzed the comprehensiveness of content moderation policies — how and how much online communities disclose content moderation rules in their policies [68, 74]. Schaffner et al., for example, conducted a mixed-method analysis of the comprehensiveness of content moderation policies across 43 platforms hosting user-generated content. Their findings show that while most policies articulate moderation goals, criteria, and practices, there is considerable variation in structure, composition, and legal grounding across platforms. They also found critical shortcomings of these policies, such as the lack of a clear definition of what to moderate, and the absence of user appeals for most moderation cases [68].

Content moderation policies are informed by laws (e.g., The First Amendment and Section 230 in the United States context) while dominated by online communities themselves. Some works have focused on understanding how content moderation policies balance legal requirements, platform motivations, social values, and user experiences [25, 26, 28, 30, 46, 49, 77].

More recent studies have recognized that regulating content generation in GAI products is an emerging issue worthy of attention [2, 32, 69]. Some research works have examined what input and output in GAI usage are forbidden by policies and guidelines [47, 62]. Our work, furthermore, investigates the comprehensiveness of content moderation policies – if and how these policies outline content moderation rules and practices.

User Experience with Content Moderation in Online Communities. Prior research has investigated user experiences and reactions to content moderation in online communities [53]. Some works examined user understanding of content moderation [17, 37, 59], where researchers found users mainly rely on their own interpretation of how the moderation systems work. Prior studies also investigated user behaviors after moderation [38], and how users circumvent or interact with moderation systems [52, 58]. Since policy statements may not align with actual practices [56], prior studies often rely on understanding user experiences to assess whether content moderation policy enforcement is effective or failing [74]. By measuring user behaviors in online communities on a large scale, researchers found content moderation successful in mitigating disturbing content (e.g., [9]). Simultaneously, researchers also found common failures such as biases and inconsistencies of algorithmic moderation [33, 51, 83, 84]; lack of transparency in moderation decisions [44, 59]; and ineffectiveness of user appeals [59]. Researchers also conducted user studies on user preferences and expectations of content moderation to inform the future content moderation policies

and practices [19, 83, 84].

Inspired by prior works investigating users in online community content moderation, we looked at how users experience content moderation in GAI products to understand the successes and failures of policy enforcement. Unlike previous studies, which typically focus on a single platform, our research also spans multiple GAI products, providing a broader and more generalized perspective.

3 Study 1: Policy Analysis Methodology

To answer our first research question, we analyzed content moderation policies of 14 representative GAI online tools capable of text and image generation, inspired by and partly following the approach of Schaffner et al., who analyzed the comprehensiveness of content moderation policies in online communities [68]. For each tool, we manually located its content moderation policies, collected web pages where policies were situated, and qualitatively analyzed those policies.

Regulated by regional laws, content moderation policies could be different when providing service in different regions. For example, content moderation of online communities is regulated by the First Amendment in the United States (US) but by the Digital Services Act (DSA) in the European Union (EU). Divergences such as DSA restricting protected speech defined by the First Amendment could lead to distinct policy enforcement between platforms in the US and EU [1, 80]. Acknowledging that regional policy alterations are out of our study scope, we only considered tools with US-based headquarters and accessed all policies through US IP addresses.

3.1 Tool List Creation

We referred to multiple resources to create a representative tool list. First, we relied on a report [66] on popular GAI products, which has been used in other academic research papers (e.g., [23]) and featured by Forbes News [22]. We picked out GAI online tools in the top 15 for text or image generation and with US-based headquarters, resulting in seven qualified tools. We excluded Character.AI,[3] a tool designed for role-playing, which may lead to a different policy focus compared to other multi-tasked or creativity/productivity-focused tools. This process resulted in seven GAI online tools.

We then supplemented our list from a user-curated GAI tool list (i.e., Awesome List [29]), survey and review papers on GAI and AIGC [8, 10], another GAI tool popularity report [21], and AI products produced by big technology companies. We initially selected five tools through this process, which constructed our initial tool list along with the seven tools above. As the study progressed, we incorporated two additional tools based on ongoing tracking of supplemental resources.

[3] https://character.ai/

Our final list contains 14 GAI online tools, with two for text generation: Claude and You.com; five for image generation: CivitAI, Craiyon, DreamStudio, Firefly, and Midjourney; and seven capable of multi-modal generation of both text and image: ChatGPT, Copilot, Gemini, Meta AI, NovelAI, Perplexity AI, and xAI.

3.2 Content Moderation Policies Collection

Using the procedure outlined below, we collected content moderation policies in September 2024 for tools except Meta AI and xAI—two tools added in the middle of the study—-and in December 2024 for Meta AI and xAI.

Defining Policy Scope. We decided to only consider content moderation policies under our research scope — governing activities where users directly interact with GAI through web-based applications. For example, OpenAI offers GPT API [4] to developers and allows people to tailor personalized GPT [5] and provide it to other users. Under this context, there are policies restricting the practices of developers and their customers, which are out of our scope.

Locating Pages with Content Moderation Policies. We followed a process used by Schaffner et al. [68] to identify pages in GAI online tools that contained content moderation policies. For each tool, we manually examined all policy and regulation-related pages and collected two types of pages. First, we selected pages that included policies regulating user interactions with GAI. Second, for those tools provided by companies developing multiple products (e.g., Google which developed Gemini), we also included company's general *Terms of Service* (ToS) that applied to all products.

Next, we visited tool support pages, including the *Help Center* and *Frequently Asked Questions* (FAQ), if they existed. To identify information related to content moderation policies, we referred to Schaffner et al.'s four common elements of content moderation policies for user-generated content: what is moderated, why content is moderated, how the process of content moderation manifests, and who takes the responsibility for content moderation [68]. We then manually checked all support pages and their child pages, collecting those with information that fell into the above four themes. Additionally, we collected pages with information on system errors, which could also relate to content moderation enforcement.

Recording Pages. We saved all static pages to PDF through the MacOS Safari browser, where the web pages retain their original visual appearance and content, with all text editable in the converted PDF file. For those non-static pages in which the text encoding could not be captured through PDF conversion, we took screenshots of the whole pages. We used the Optical Character Recognition (OCR) service on the screenshots to convert all characters into editable text and save them in PDF

[4] https://openai.com/api/

[5] https://openai.com/index/introducing-gpts/

format. We eventually recorded 52 PDF files of 51 pages, forming the policy dataset we analyzed.

3.3 Policy Analysis

To get insights into how content moderation was defined and disclosed in GAI online tools' policies, we conducted a qualitative analysis of the policy dataset. Our analysis was performed deductively with multiple rounds using the analysis tool MAXQDA.[6] To start, the first author developed an initial codebook for the analysis and discussed it with the research team. The design of our initial codebook was based on four key components of how policies describe content moderation, as identified by Schaffner et al. [68]: what content is moderated, why content is moderated, how content moderation is enforced, and who is responsible for content moderation. Using the initial codebook, the first author read all documents and deductively coded all segments that fit into existing themes. Simultaneously, excerpts that related to content moderation but did not fit the initial codebook were tagged with open codes (i.e., adding and removing themes) to build the final codebook.

A second round of coding was then performed by two additional researchers using the final codebook. All documents were equally divided into two sets and assigned to these coders, ensuring that every document was coded by at least two coders at the end. Three coders met regularly to discuss the analysis process, compare each other's codes, and solve coding disagreements. Since our data informed the iterative analysis and we took care to minimize subjectivity and disagreement in the process, we did not calculate the inter-rater reliability (IRR) [3, 57].

Locations of Content Moderation Policies. Within 51 pages recorded, 44 of them were coded during policy analysis, indicating that these pages contain content moderation policies. Similar to where content moderation policies in online communities are located [68], we observed that those policies for GAI online tools are scattered across various policy and support pages. This might be due to, in our understanding, the lack of standardization of how content moderation policies should be presented for GAI online tools and online communities. Below, we briefly describe the common locations of content moderation policies in GAI online tools.

The most common page for all tools that includes content moderation policies is the ToS, except Gemini which did not have a separate ToS page. Beyond ToS, 5/14 tools have a separate policy page for *Acceptable Use Policy* (AUP, also named alternatively in some tools, such as *Prohibited Usage Policy* or *Usage Policy*) defining allowed and forbidden usage of the tool, including the criteria of content moderation. For the five tools provided by companies developing multiple products, the ToS of four companies includes policies applicable to content moderation in GAI online tools, and two of these also have AI/GAI product-specific rules regulating content generation activities. Other policy pages we found content moderation policies in are *Service Terms* (ChatGPT), *Community Guidelines* (Midjourney), and *Safety Center* (CivitAI). We also found 9/14 tools have information on content moderation rules in their support pages. In addition to these nine tools, four other tools have support pages but do not contain content moderation policies.

[6] https://www.maxqda.com/

3.4 Limitations

Our policy analysis study has several limitations. Our tool list is only representative of commonly used GAI online tools in the US context. Since the market of GAI online tools is monopolized, most customers use only a few types of tools [22, 66], and we cover most of them. Moreover, our policy collection process relied on manual checking on specific pages, meaning we might have missed pages with information on content moderation policies.

4 Study 1: Policy Analysis Findings

In this section, we present findings about what and how complete the content moderation policies of GAI online tools are (RQ1), given that these tools are still evolving when compared with online communities whose policies have been established for much longer and cover various topics [68]. We observed that policies in GAI online tools, like those in online communities, cover the three major components of content moderation practice outlined by Singhal et al. [74]: content moderation criteria — providing definitions of forbidden content (§4.1); content moderation methodology — explaining safeguarding approaches to detect problematic content generation (§4.2); and consequences of content moderation enforcement — detailing the platform responses to problematic content generation and corresponding mechanisms for user appeal (§4.3). However, we did not find clear presentations in GAI online tool policies of 'why content is moderated' which online communities specify — justifications on why GAI online tools engage in content moderation, beyond general statements on how platforms value user safety and responsible AI development.

Overall, both online communities and GAI online tools specify similar rules regarding what is not allowed and their moderation strategies at each stage, with GAI online tools modeling their content moderation policies on the relatively mature content moderation frameworks used in online communities. However, online communities have policies that cover user-generated content, while GAI online tools' policies cover user input to the tool (referred to as *input* or *prompt* from here on) in addition to content generated by the tool (referred to as *output* from here on). In most online communities, user-generated content is posted or uploaded directly. Instead, users have less control over the randomness of how the GAI

Moderation Stages	Similarity in Policies of Online Communities and GAI Online Tools	Uniqueness in Policies of GAI Online Tools
Moderation Criteria	Both online communities and GAI online tools detail a wide prohibition on different problematic content (§4.1.2)	GAI online tools regulate input, output, as well as output distribution and secondary use (§4.1.1)
Moderation Methodology	Both online communities and GAI online tools describe their moderation systems as a mixture of automatic detections and human reviews (§4.2.1) Both online communities and GAI online tools acknowledge their content moderation is not infallible and place responsibility for content on users (§4.2.1)	GAI online tools enforce content flagging on both user input and GAI output (§4.2.1) Some GAI online tools mentioned safety measures in model training as a moderation method (§4.2.1) Only a few GAI online tools provided methods for user-driven content moderation (§4.2.2)
Moderation Consequence	Both online communities and GAI online tools take down problematic content and punish users (§4.3.1) Both online communities and GAI online tools leave users with few options once they have been moderated (§4.3.2)	GAI online tools take actions on both user input and GAI output, with the goal of preventing the generation and presentation of output (§4.3.1) Except legal violations, moderation enforcements on content in GAI online tools are the same across different types of content policy violations (§4.3.1)

Table 1: Major similarities and differences between content moderation policies in online communities and GAI online tools.

online tools generate output with their input, for which users may unintentionally generate undesired output [24, 73]. Table 1 summarizes major policy similarities and differences in the two types of platforms. Next, we present details of how GAI online tool policies describe the three content moderation stages.

4.1 Content Moderation Criteria

4.1.1 How are Moderation Criteria Specified?

Similar to how online communities describe prohibited user-generated content in community guidelines [42], content moderation criteria in GAI online tools are mostly described by articulating what behaviors are acceptable or prohibited when using the service, in the separate AUP, acceptable use guidelines in ToS, and support pages. Interestingly, none of the tools except CivitAI has a separate *Content Policy* or related sections in policy and support pages that specify all rules about forbidden content. In short, the GAI online tool policies are often unstructured, and as such, we found content moderation criteria are presented in a mixed and complicated manner. For example, some rules apply across multiple aspects, including input, output, output distribution, and, in some cases, other content hosted within the tool's corresponding services.

Policies Implicitly Regulate Content Generation Requests. For all 14/14 tools we studied, each has several acceptable use guidelines using descriptive language to broadly define acceptable or prohibited tool usage, usually starting with *"Do not use the service to ..."*. These guidelines apply to all user behaviors within GAI online tools, including when they generate content using the tool. Notably, all 14 tools we studied forbid users from overcoming system restrictions (or 'jailbreaking' the system), which implicitly rules input.

Policies Explicitly Target Content. In addition to descriptive acceptable use guidelines, each GAI online tool lays out rules on what content is prohibited that are scattered over policy and support pages, much like in online community content moderation policies [68]. GAI online tools typically talk about 'content' which encompasses both input and output, represented by the following definition: *"You may provide input to the Services ('Input'), and receive output from the Services based on the Input ('Output'). Input and Output are collectively 'Content'."* (ChatGPT's ToS). Many rules solely use the term 'content' when describing prohibitions, and they do not distinguish if there are different rules for different modalities of content (e.g., text or image) or depending on different tasks (e.g., dialogue versus creative writing).

Furthermore, some image generation tools are integrated into or built up with online communities for sharing user-created AIGC (i.e., Firefly, Craiyon, CivitAI, and Midjourney). We found that policies there just describe rules about any content users have as input/output in GAI online tools and uploaded content in online communities solely as 'content'. As Midjourney defines 'content' in its ToS: *"Inputs, Assets, and other content such as messages, photos, videos, and documents that you may provide to the Services (such as through uploading, posting, sharing, or chat messages) are collectively, 'Content'."* (Midjourney's ToS) Therefore, it is sometimes unclear whether a rule applies to all or specific types of content defined by these tools.

Prohibitions on Output, Input, and Output Distribution. Rules clearly governing output often start with *"Do not generate/create content that ..."* if without directly stating the term 'output'. We noticed that all tools we studied except NovelAI (13/14 tools) elaborate on what output is restricted by defining rules in their policies. Meanwhile, 10/14 tools included specific rules outlining criteria for input. Besides restricting input that leads to prohibited output, these rules

mostly outline the same restrictions as rules targeting output, with many defining forbidden input along with forbidden output in the same place. For example: *"Do not create images or use text prompts that are inherently disrespectful, aggressive, or otherwise abusive."* (Midjourney's *Community Guideline*).

We also found that 12/14 GAI online tools specify rules that address the future distribution and secondary use of output, despite the limited practical authority they have to enforce these rules. Similar to rules governing input, rules on output distribution often reiterate the same restrictions as those applied to output and, in many cases, are presented alongside the output restrictions: *"[Do not] Creating, generating or distributing content that depicts gratuitous violence, cruelty, abuse, sex or gore."* (You.com's AUP). Additionally, we noticed special restrictions that applied to output distribution only. 8/14 tools forbid users to mislead others that the output from the GAI online tool is created by humans: *"[You will not] Represent any Output (defined below) as human generated when they are not."* (xAI's ToS). 5/14 tools prohibit users from using the output to develop machine-learning models: *"[You may not] Use Output to develop models that compete with OpenAI."* (ChatGPT's ToS).

4.1.2 What Types of Content Are Forbidden?

In addition to the loose and scattered structure, content moderation criteria are articulated using a variety of terms that cover a broad range of restrictions, similar to how community guidelines in online communities outline content prohibitions [42]. This observation on the varied use of terms in moderation criteria also aligns with findings on AUP of foundation models [47], where more than 120 prohibited behaviors are described using diverse terms. However, we noticed that forbidden content defined in the GAI online tool policies could be grouped into broader categories, based on the nature of each prohibition. Drawing on major content moderation topics in online communities recognized in prior works [25, 68, 74], we mapped out four types of prohibited content at a high level. The four categories apply to all content span input, output, output distribution, and include rules that do not specify the type of content: **harmful content** that harms individuals, groups, and public safety, regardless of legal or illegal; **content that violates other's rights** with privacy and intellectual property/copyright violations; **misleading content** such as mis/disinformation and deceptive content, including content with misleading nature; and **content that is not appropriate for everyone** which depicts sexual and violence. We found that all tools except NovelAI outline all four types of prohibitions in their policies.

We note that the first three categories are frequently addressed in online community policies, where most mainstream platforms enforce strict prohibitions [68, 74]. In contrast, the restrictions of the fourth category, which are strictly prohibited in many GAI online tools, largely depend on the platform's nature when applied to online communities, where personal content moderation (e.g., personalized filtering) is commonly used instead of direct bans from the platform [39, 74].

4.2 Content Moderation Methodology

4.2.1 How is Problematic Output Detected or Prevented?

GAI online tools outline various approaches used for their moderation systems in their policies. These include interventions outside the GAI model for content flagging with similar strategies used in online communities [68, 74], as well as improving the GAI model itself to generate safer output, which is a GAI-specific strategy. Meanwhile, GAI online tools also acknowledge their lack of ability to moderate output as well as online communities do in moderating user-generated content [68], putting all liability of input and output on users.

As with the moderation method disclosed in policies of online communities, GAI online tools outline themselves as employing both automated detectors or filters (7/14 tools) and human reviews (5/14 tools) in their moderation systems. These approaches are enforced in both input and output, represented in the following policy: *"Your prompts and the results generated [...] may be reviewed through both automated (e.g., machine learning) and manual methods for abuse prevention and content filtering purposes."* (Firefly's *User Guidelines*). Nevertheless, we noticed that many tools tend to be vague on technical details and nuances of these interventions. For example, another policy in Firefly's *User Guidelines* only explains its output flagging as *"use available technologies, vendors, or processes"*.

Simultaneously, we found five tools driven by self-developed foundation models that claim themselves dedicated to training and improving their models for output safety. As exemplified by the following instance: *"We also work to make our models safer and more useful, by training them to refuse harmful instructions and reduce their tendency to produce harmful content."* (ChatGPT's *Usage Policy*). Occasionally, policies refer readers to model documentation for further details of these approaches: *"Limitations and bias in AI are still being researched and we're working actively on this subject. You can learn more in the DALL·E mini model card."* (Craiyon's FAQ section on the home page).

Meanwhile, 13/14 tools have disclaimers when it comes to who is responsible for the output generated using the GAI online tool and its moderation. In these tools, these disclaimers about the guaranteed safety of generated output frequently cite the well-known unpredictability of GAI output [24, 73]. As illustrated by the following policy: *"This use of AI is relatively new and still evolving. As a result, while we have taken - and continue to take - efforts to preclude your creation of extreme content, we cannot guarantee the suitability or appropriateness of the resulting images you generate."* (DreamStudio's ToS). Some tools also acknowledge the potential failure of

their safeguards, similar to the online community acknowledgments of challenges in moderating user-generated content: *"These [safety] features are not failsafe, and we may make mistakes through false positives or false negatives."* (Claude's *Help Center*). 12/14 tools say users are ultimately liable for the input, output, and output distribution and use of GAI online tools. For instance: *"You are solely responsible for your Input [...] You are solely responsible for the creation and use of the Output and for ensuring the Output complies with the Terms."* (Firefly's *ToS*).

4.2.2 How Can Users Combat Problematic Output?

In addition to putting users in charge of their input and output, service providers sometimes allow users to engage in the moderation process by reporting problematic output to them. Some reporting channels specifically target the copyright infringement of output, as mentioned by 6/14 tools. Similar to what was observed in online communities [68], GAI online tool policies usually claim a unified reporting pipeline for copyrighted content, grounded by existing laws (i.e., Digital Millennium Copyright Act, DMCA), and to fulfill legal requirements.

Beyond copyright infringement reports, we only found 5/14 tools mentioning how users can combat problematic output in general. This is surprisingly contrary to online platforms, which heavily rely on user-driven moderation even beyond copyright infringement [68, 72]. Although channels for user engagement in content moderation are limited in GAI online tools, some tools provide feedback mechanisms integrated into the tool interface (e.g., functions to instantly 'thumb-up' and 'thumb-down' output) and direct contacts with the service team (e.g., report forms or contact email). As summarized by this example quote: *"Users can report problematic or illegal content via the Feedback button or the Report a Concern function."* (Copilot's ToS).

4.3 Content Moderation Consequence

4.3.1 What Happens to Problematic Content and Users?

GAI online tool policies elaborate on how they respond to problematic input and output, as well as users engaging in problematic output, similar to rules disclosed in policies of online communities [68]. As summarized by the following policy: *"Content that violates our rules, or attempts to circumvent our content restrictions, will result in appropriate actions, which may include content removal, flagging of the account, suspension of access to the image generation feature, or a ban from the platform."*(CivitAI's *Safety Center*)

User-targeted moderation responses in GAI online tools are nearly the same as those in online communities. Although most GAI online tools say account restrictions target general term violations or can be enforced for any reason, we found 5/14 tools relate these account-level actions to problematic content generation specifically. For example: *"We have adopted a policy of terminating, in appropriate circumstances, Users who are deemed to be repeat infringers [of copyright]."* (xAI's ToS). 3/14 tools also issue warnings to certain accounts with users prompting for content that violates the content rules, exemplified by the following policy: *"As part of our safety process, we warn users if we believe their prompts are violating our Usage Policy."*(Claude's *Help Center*).

In comparison, content-targeted responses (10/14 tools) share a similar strategy with those in online communities but have a different goal. Instead of focusing solely on removing problematic content like in online communities, GAI online tool policies aim to refuse the GAI processing of problematic input and prevent the presentation of undesirable output. Thus, flagged input is typically blocked or removed from the system, with an error or warning message returned instead of the requested output. If problematic output is generated, it may be removed or blurred afterward. As represented by the following quote: *"Image Creator may block prompts that violate the Code of Conduct, or that are likely to lead to the creation of material that violates the Code of Conduct. Prompts or Creations that violate the Code of Conduct may be removed."*(Copilot's ToS of Image Creator). We found no distinction in these enforcements across policy violations, unlike online communities which, for example, investigate harmful speech and misinformation before content removal but remove copyright violation content immediately [68, 74]. One exception is certain legal violations, where service providers may take additional legal action on both input and output. For instance: *"We report apparent child sexual abuse material (CSAM) to the National Center for Missing and Exploited Children."*(ChatGPT's *Usage Policy*).

4.3.2 What Users Can Do After Being Moderated?

Similar to policies of online communities [68], we noticed policies of GAI online tools give users few options after they get moderated. While most of the tools have terms for legal disputes, user appeal after general content moderation is not well specified in the current GAI online tool policies for the tools we examined – only 5/14 tools we studied clearly outline non-legal appeals users can engage in after getting moderated. Furthermore, most appeals are only applicable to redress account restrictions, rather than content-targeted enforcement like blocked input and removed output.

Beyond simple user appeals, there is still little that can be done after content moderation. We observed that none of the tools except three talk about user feedback if they think any content-targeted responses of moderation are questionable. We even found an extreme case, where the service provider does not offer an appeal or ask for any feedback but asks users to try again if the output gets moderated: *"Q: Why are some*

of my images blurred? [...] [T]he model will blur out any content that may be considered inappropriate or offensive. [...] If you are unhappy with the results, you can always try again with a different prompt." (DreamStudio's FAQ).

5 Study 2: Reddit Study Methodology

To answer our second research question, we conducted a case study focusing on content moderation when users engage in AIGC creative tasks such as creating fiction and art. To do so, we analyzed Reddit posts on discussions about content moderation experiences in AIGC creative tasks using GAI online tools. Our approach follows one that is commonly used in prior works to gain real-time insights from people, such as understanding user perceptions and reactions after content moderation in online platforms [48, 52], by qualitatively analyzing online discussions. We performed a keyword-based search in GAI online tool-related subreddits and then manually filtered out irrelevant posts, creating the final Reddit dataset for analysis. Finally, we performed qualitative analysis on randomly selected posts and corresponding comments from our dataset.

5.1 Reddit Post Collection

We performed Reddit post collections in October 2024 through keyword-based searching. We started by deciding which subreddits to focus on and creating a keyword list of content-moderation-related words. Next, we scraped Reddit posts through the Python Reddit API Wrapper (PRAW),[7] and did another round of manual filtering to get the posts we wanted.

Subreddits Choice. Our selection of subreddits for data collection follows two criteria. First, we only included subreddits that solely discuss tools in our list from Study 1 (§3.1), to exclude discussions on content moderation beyond GAI online tools. For example, most posts discussed AIGC creation in r/aiArt, but we did not include this subreddit since it was difficult to determine if a discussion was about using GAI online tools. We also only considered subreddits within the top 5% of all subreddits, to collect high-quality discussions from active communities. Complying with the two criteria above, we finally selected seven subreddits to perform data collection (ordered in size): r/chatgpt, r/midjourney, r/dalle2, r/claudeAI, r/Bard, r/dalle, and r/perplexity_ai.[8] Discussions on these subreddits span six GAI online tools: ChatGPT, Claude, Copilot (DALL-E image generator only), Gemini, Midjourney, and Perplexity AI.

Keyword List Creation. Referring to prior works on analyzing public discussions of online platform content moderation [48, 52], we picked 12 words that were frequently used when talking about content moderation: 'moderate', 'censor', 'ban', 'block', 'suspend', 'restrict', 'warn', 'flag', 'appeal', 'violate', 'terminate', and 'remove'. Using these keywords, we first performed an open search in r/chatgpt and r/dalle2 through the Reddit website, where two researchers reviewed 25 posts per subreddit related to content moderation in AIGC creative tasks, to get a sense of the common discussion themes. Four additional keywords frequently used in those posts were identified: 'content policy', 'guardrail', 'filter', and 'refuse'. These collective keywords made up the final keyword list that we utilized for data collection.

Scraping Posts. We searched and recorded posts via PRAW with either a title or selftext (content of the original post) containing at least one of the keywords in our list. We used all forms and tenses of the keywords in this process. For example, when matching for the keyword 'censor', we used 'censor', 'censored', 'censoring', and 'censorship'. We did not apply time constraints in searching, indicating that all posts from when the subreddits were established to October 2024 (the time of data collection) were under the search scope. We collected a total of 5185 posts through this process.

Dataset Cleaning. We noticed a high false-positive rate on our collected posts, due to the broad scope and context of keywords used beyond content moderation in AIGC creative tasks. To reduce the high false-positive rate, we manually checked all posts and filtered out the irrelevant ones. To be specific, three researchers checked the selftext of each post, only keeping those that discussed a general opinion on content moderation in GAI online tools, or a user experience of being moderated when engaging in AIGC creative tasks using GAI online tools. After that, we utilized PRAW to get all corresponding comments on the remaining posts. We cleaned up the comments by removing those shown as '[deleted]' or '[removed]'. Through this process, we retained 1123 posts and 33465 corresponding comments, constructing our Reddit dataset.

5.2 Data Analysis

Due to the high volume of posts and comments, we analyzed a portion of the dataset rather than the whole dataset, following the common practice in prior works conducting qualitative analysis on Reddit (e.g., [48, 52, 78]). We randomly selected 130 posts from our dataset, which corresponded to 3839 comments, for our data analysis. All posts and corresponding comments were imported into MAXQDA for further analysis. The statistics of the final Reddit dataset and sampled posts for data analysis are shown in Table 2 in the Appendix.

We performed an iterative, deductive coding and thematic analysis. First, we ran another random sampling to split our sampled dataset into 30 posts (971 comments) and the other 100 posts (2868 comments). The first author then reviewed and performed an initial coding on the set of 30 posts and

[7] https://praw.readthedocs.io/en/stable/

[8] We excluded r/NovelAI since we found NovelAI has enforced almost no content moderation, based on Study 1 analysis and discussions in this subreddit.

corresponding comments to build the codebook. During this process, the research team held regular meetings to review sampled posts, discuss questions, and refine the codebook.

All 130 posts and corresponding comments were then coded iteratively by three researchers using the codebook. First, all posts were divided equally into three sets with a comparable number of comments and assigned to each coder for primary coding. These sets were later reassigned among the coders for secondary coding, ensuring that each post and comment was coded by at least two coders. Three coders met regularly to discuss the coding progress, compare each other's codes, and solve discrepancies during secondary coding. We did not calculate IRR, as the iterative analysis was informed by our data and all disagreements were resolved during the process [3, 57].

We reached thematic saturation midway through analyzing these 130 posts—where no new codes emerged—indicating that our analysis was comprehensive for extracting qualitative findings [67]. Thus, we did not sample additional posts.

5.3 Limitations and Ethics

We acknowledge a few limitations of our Reddit study. First, our data collection only covered discussions on six GAI online tools. We also observed an uneven distribution of Reddit discussions across different tools. For instance, while discussions on content moderation in ChatGPT are extensive, we found only seven related posts about Perplexity AI. Additionally, our analysis was limited to a sample set of user perceptions posted in public discussions. Therefore, without a broader analysis, our study may not have assessed all successes and failures around content moderation policy enforcement.

Although all posts and comments we collected and analyzed are publicly accessible, there are potential privacy violations for the users who post them, a widely acknowledged ethical concern of social media and Reddit research [20]. Therefore, when reporting any post or comment in the paper, we removed all identifying information and adjusted some wording into synonyms to prevent a direct search. Moreover, the Reddit study was reviewed by our Institutional Review Board (IRB) before the data collection.

6 Study 2: Reddit Study Findings

In this section, we present findings on how content moderation policy enforcement in GAI online tools succeeds and fails (RQ2), from the case of user experience of using these tools for AIGC creative tasks.

6.1 Successes of Moderation System: Blocking Malicious AIGC Creation Attempts

Based on user-shared examples in public Reddit discussions we analyzed, we found that moderation systems in GAI online tools are generally effective in detecting and blocking AIGC creations and requests that maliciously violate moderation criteria, such as requests for or generated content involving pornography, scams, or hate speech. We observed instances where users posted about how problematic AIGC creations were either denied at the request stage or immediately removed after generation across all six GAI online tools.

We found users posted about their satisfaction and appreciation with current content moderation enforcement for 3/6 tools, regarding its success in blocking malicious attempts. Some users strongly recognized the potential harms of malicious AIGC and, therefore, appreciated the comprehensive moderation criteria and moderation systems that GAI online tools employ to successfully protect individuals and society from harmful content. This is exemplified by a user who commented on a request for porn to be moderated in ChatGPT's DALL-E: *"Basically no deep fake images that could be used to falsify evidence of adultery [are allowed]. I'm really glad this sort of thing wasn't available in my teens, and that [OpenAI] is protecting people from it now."*

6.2 Failures of Moderation Pipelines: Moderation System and User Support Failures

Despite evidence that moderation systems effectively block malicious AIGC creation, we found that many users discussed failures in the current moderation practices on GAI online tools when performing AIGC creative tasks. These cases span not only the moderation systems themselves that failed to make justified moderation decisions (§6.2.1), but also the post-moderation stages, where service providers failed to assist users in understanding and redressing moderation decisions (§6.2.2). Although some user-reported moderation failures here resemble those in online communities (e.g., biased and inconsistent moderation in social media [33, 51, 83]), their impact on user experience extends beyond the frustration recognized in moderation failures within online communities [59] – moderation failures in GAI online tools have hampered the usability and capability of GAI, *"making [the tool] completely useless"* for users. Next, we expand on how users perceived failure cases of the moderation system and user support, and how they blocked the usability of GAI online tools for AIGC creativity tasks.

6.2.1 Failures in Moderation System Decision-Making

According to public Reddit discussions we examined, users frequently talked about how the moderation systems, especially the algorithmic and automatic ones, behaved inaccurately, inconsistently, and unreasonably when GAI online tools were used for AIGC creative tasks. While similar failure cases have been shown by researchers through model testing and auditing (e.g., [54]), we observed that users developed their own 'folk theories' — collective understanding of sys-

tem operation based on personal knowledge [15] — on how and why moderation systems fail in real use cases.

Failure in Mitigating False-Positive Rate (5/6 tools). When discussing the moderation system inaccuracy, users frequently posted the pervasive false-positive moderation decisions – where users were moderated for using harmless prompts to generate AIGC that was not intended to violate any rules. Users discussed that some false positives were just random bans issued by moderation systems [31], or because their prompts resulted in arbitrary malicious outputs due to the randomness of GAI output [24, 73]. As a comment reasoning a random false-positive moderation case in Copilot's DALL-E: *"That [harmless input] still allows for random imagery that you aren't in control of. You've input token words, but whatever is composed might break the filters. It's just an RNG game with loosely curated results."*

Meanwhile, users also reported that some false positive decisions to stop a request from being processed might be owing to the over-sensitive input moderation. Users observed a high likelihood of requests blocked if they were using input with, or requesting content correlated to moderation criteria, even if their requests were not to create AIGC that violated policy. For example, requesting images of children could be falsely flagged as child abuse. Furthermore, many posts mentioned the over-sensitivity of banning inputs containing certain words. This input moderation strategy, although not explicitly detailed in any policy because of its ambiguity in disclosing details of moderation methods (§4.2.1), was widely recognized by users as being implemented in image generation tools such as Midjourney and DALL-E. A ChatGPT's DALL-E user wrote: *"When asking for a portrait of a person, 'headshot' is banned. Of course, it is for violent reasons. But double meanings in English abound."* Table 3 in the Appendix presents examples of false-positive moderation on input related to each moderation criterion.

Policies acknowledged that there could be false-positive moderation (§4.2.1), and users spoke of circumventing these cases through tricks like jailbreaking [43, 91]. Yet, many users argued that the frequency of false-positive moderation in practice was excessively high, which impaired normal tool usage and frustrated users. A former ChatGPT's DALL-E user shared their experience: *"I got too many warnings for nothing [...] I just find myself thinking of something to try, then becoming afraid of triggering arbitrary warnings, and then just not trying anything. It's resulted in me just not using it."*

Failure in Making Moderation Decisions Consistently (5/6 tools). Users also discussed the inconsistency of moderation systems – where moderation systems' behaviors varied across different users or responded differently to similar or identical requests that should uniformly be either moderated or allowed. For example, a Copilot's DALL-E user questioned the moderation system when two requests to create copyrighted content produced different outcomes, where one was processed while the other was moderated: *"I tried 'Batman and Catwoman getting married' and got caught by the filter. Somehow 'Nightwing and Starfire getting married' worked."*

Most inconsistencies could be attributed to the inherent bias of moderation algorithms [6, 54], as well as the randomness of algorithmic moderation systems and GAI output, as some users also realized. Some other users attributed this to intentional unfair enforcement in moderation, speculating that undisclosed implicit or non-transparent rules existed in GAI online tools. For example, a ChatGPT user speculated that the moderation systems work differently in different accounts based on prior user behavior: *"Sometimes I think ChatGPT works differently based on previous interactions. It always quit with my friends who constantly try to break it, but it let 95% of my attempts pass."* Some users, however, expressed distrust of service providers regarding potential implicit or non-transparent rules. A ChatGPT's DALL-E user criticized: *"I tried to use a prompt with Elon Musk's name in it. That's not allowed. Neither is Joe Rogan [...] But I'm allowed to use other high-profile people's names? Seems like the Developers are letting their personal bias get in the way."*

Failure in Enforcing Moderation Criteria Regarding the Context of Tasks (6/6 tools). Researchers have found that GAI content moderation guardrails overly-censored cultural content through algorithmic auditing [54, 62]. This finding aligns with real user experiences – many discussions we investigated highlighted frequent and extensive restrictions in doing AIGC creative tasks. Although many moderation cases broadly aligned with the moderation criteria (§4.1.2), users argued that these decisions should be reconsidered in the context of AIGC creative tasks, feeling they were being overly restricted.

Users frequently reported moderation examples when they were creating fictional, horror, and fantasy writing and art, with prompts and AIGC creations being flagged as harmful or violent. Another type of content users reported frequently being prohibited was romance and artistic content, which was classified as sexual. We also found various and scattered discussions on the requests being blocked or the AIGC creations being removed, spanning requests for historical materials, autobiographies, educational materials, propaganda, and creations involving jokes, sarcasm, or profanity. Users also perceived divergences from their requests in writing tasks as over-restrictive moderation practices. They reported that the generated writing was often rendered plain or positive, failing to write debatable topics even without policy violations. We note that this phenomenon mainly arose from value alignment of LLMs — a process of fine-tuning models to produce responses that adhere to widely accepted opinions [27, 34] — rather than content moderation mechanisms. Table 4 in the Appendix lists examples of AIGC creative tasks users tried to complete or completed with harmless intentions, when they felt they were over-restricted under each moderation criterion.

Users argued that the one-size-fits-all moderation enforce-

ment failed to consider user intentions and the intended use of AIGC, thereby risking normal creative processes. A ChatGPT user posted: *"Even if the generated text is about something bad, it could potentially still be used for good. A story set in medieval times will probably contain fighting and violence. The story can still be considered good if people like reading it."* Users spoke of how the broad definition of problematic content was, in some cases, essential for certain creations like fictional writing and art. As a ChatGPT's DALL-E user argued: *"I agree that depictions of violence and some nudity should be perfectly fine to generate as long as they're not photorealistic or otherwise problematic. After all, both are extremely prevalent in art."* Moreover, users stated that without awareness of the context of moderation enforcement, the GAI online tools were now more restrictive than most digital and physical platforms. For example, many moderated requests were intended to generate content that, according to users, was *"perfectly acceptable on network TV"*, or even *"permissive in books marketed at early teens"*.

Meanwhile, users complained about overly restrictive moderation enforcement on AIGC creative tasks because they limit GAI's creativity and generative capabilities. For example, a representative comment by a Gemini user spoke of getting moderated in writing political and historical topics: *"These experiences lead me to believe that Gemini's strict content filtering significantly limits its ability to engage in a wide range of topics, potentially hampering its usefulness."*

6.2.2 Failures in Supporting Users After Moderation

In public Reddit discussions we analyzed, users frequently expressed frustration over their experiences after being moderated. They discussed unclear explanations of moderation decisions, ambiguous policies that hamper the reasoning of moderation decisions, and limited assistance provided in user appeals. Failures in supporting users after moderation further undermine the usability of GAI online tools.

Failure in Providing Clear Explanations of Moderation Decisions (5/6 tools). Users expressed their confusion about the explanations provided when GAI online tools moderated their input or output. Similar to what happens after content removal in online communities [59], when users got moderated in a GAI online tool either resulting in system refusal or account ban, they sometimes received only a generic notice of policy violation or experience a plain system-generated refusal like *"I can't assist with that request"*. These responses typically lacked specific details about where, what, and how the policies were violated.

In text and multi-modal generation tools, even if users obtain a GAI-generated detailed explanation of moderation decisions, it might further confuse the users not because of the lack of response transparency, but due to the hallucination of GAI output [94]. As reported by users in ChatGPT, Claude, and Gemini, they could always ask the system to generate the reason their input or output got moderated. However, the explanations generated by GAI were nonsensical. A ChatGPT's DALL-E user shared: *"I had one image out of four hit the censorship today, was doing pencil sketches of foxes. I asked why they were censored, and ChatGPT said it didn't know why."* Another ChatGPT user reasoned this phenomenon as: *"ChatGPT doesn't actually know why the content policy kicked in. If you ask, it will just make up something based on the prompt. But the real reason isn't actually known by it."* Unreasonable justifications generated by GAI systems were also observed when the system returned a refusal with an explanation directly. As an example shared by a Gemini user: *"I asked it to write a script for an ASMR video featuring hypnosis and finger snaps and it refused because it could be dangerous and should be done by a professional."*

Failure in Presenting Clear Moderation Criteria in Policies (3/6 tools). Regardless of the widely acknowledged content moderation criteria in policies (§4.1.2), users argued that these criteria often lacked details or, in some cases, were not provided at all. Some users questioned the clarity of moderation criteria explanations. For example, a Midjourney user remarked: *"[The explanations are] all kinds of vague. 'Avoid nudity but also avoid fixation on the naked breast.' So is the male torso OK or not?"* The problem of unclear criteria was particularly severe in image generation services that blocked input containing certain words, as none of them provided a list of banned words to the public. Many users argued that service providers should address this lack of transparency by presenting the banned word, along with explanations for why each word was prohibited. As a ChatGPT's DALL-E user criticized: *"How as a user am I supposed to know all of the 'no' words when I'm trying to edit an image and each bad strike counts towards a hidden 'ban' counter? [...] A product should clearly define what is and what is not acceptable through its ToS. If a prompt is 'bad', the reason should clearly be spelled out so that a user knows not to pursue that. "*

The unclear moderation criteria, along with users receiving what they perceived as inadequate and unreasonable explanations after moderation, left users without a reliable framework to interpret moderation decisions. When triggering moderation, users relied on self-reasoning about how the moderation system made the decision, which policies they might have breached, and if the moderation decision was a system failure. This situation, users argued, severely compromised the tool's usability. A ChatGPT user wrote: *"Now imagine having to redo a message 10 times to find exactly what words trigger the filter. I'm not left with a lot of messages to do my thing."*

Failure in Effectively Supporting User Appeals (3/6 tools). User appeals were not widely acknowledged in policy (§4.3.2). Nevertheless, many conversations we examined mentioned how badly the existing appeal process functioned, echoing what happened in online communities [59]. When appealing decisions where their account were restricted on a GAI online

tool due to content moderation, users often faced long wait times for a final decision, or sometimes, never received a response at all. Users mentioned that this situation could occur when appealing to the moderation system failures as well. A user who was previously banned by OpenAI because of moderation from DALL-E wrote: *"It took a solid 5 months to get my account back. It's a massive problem that your account can be terminated for innocent mistakes, and it takes that long to recover it."* Meanwhile, users who appealed or submitted feedback regarding moderated content also reported that the process was sometimes ineffective, not only in having their creation requests reconsidered but also in obtaining an explanation for the moderation decision. A Midjourney user wrote: *"Now I try to appeal and it just gives me this: 'Sorry! Our AI moderator thinks this prompt is probably against our community standards. Please review our current community standards...'"* Because of the broken appeal pipeline, some users would not even consider appealing even when facing severe problems. As a Midjourney user who received massive arbitrary false positive moderation on their creations acknowledged: *"I know I can contact support, but that takes time and effort every time."*

7 Discussion

Our policy analysis shows content moderation policies in GAI online tools are fairly comprehensive but lack details (§4). Through investigating user perceptions, we also reveal that content moderation in GAI online tools succeeds in blocking the majority of malicious content generation but users feel that they fail in making justified moderation decisions and providing post-moderation support (§6). Based on our findings, we outline implications for improving GAI content moderation policies and enforcement below.

7.1 Improve Moderation Policy Outline

Similar to content moderation policies in online communities, we found content moderation policies in GAI online tools were scattered on multiple pages (§3.3) and that most tools lack a dedicated section outlining all rules for prohibited inputs and outputs (§4.1.1). We also found that content moderation policies for GAI online tools and their related online communities (e.g., image generation tools that are built with or integrated into online communities) are mixed up together (§4.1.1), making it difficult to distinguish policies for the GAI from policies on posting to a corresponding online community. Given the trend of integrating GAI services into online communities (e.g., [79]), this could deepen the problem for both external regulators to check legal compliance at scale and users to seek guidance on content generation using a GAI product, similar to the case of online communities with unstructured policies [68]. We, therefore, recommend that GAI products **establish a unified and clear policy structure** for presenting content moderation rules. This should include establishing dedicated policy pages for all GAI content moderation policies, and separating moderation rules for GAI input and output from those on GAI-associated online communities.

Moreover, we found content moderation policies in GAI online tools commonly lack detailed user-driven moderation methods or provide information on user appeal when they are moderated (§4.2.2, §4.3.2). Online communities heavily rely on collective user action for detecting problematic content to overcome challenges of platform-driven moderation, such as inaccuracies and biases in algorithmic detection, the sheer volume of user-generated content that burdens human moderators, and socially contextual boundaries of issues like misinformation and discrimination [68, 72, 84]. In GAI products, similar issues exist in moderation systems, with the challenges further amplified by the unpredictability [24, 73] and biases [27] of GAI outputs. Researchers have been advocating for collective user involvement in GAI safety, e.g., through user-driven GAI output auditing [14] and user-driven value alignment [16]. Therefore, we strongly recommend GAI products **set up detailed procedures with clear steps for user feedback on problematic GAI output** that are not detected by moderation systems. **A robust user appeal pipeline** can also provide a feedback mechanism to collect information on questionable moderation decisions, beyond supporting users after moderation.

7.2 Balance Usability and Safety in Moderation

We found users spoke of how moderation systems successfully blocked malicious AIGC creative tasks in GAI online tools in public discussions (§6.1). Simultaneously, we perceived that user-experienced failures of moderation systems in AIGC creative tasks were often linked to the censorship of normal requests (§6.2.1). In online communities, the trade-offs between free speech and maintaining community safety as well as the boundaries for content moderation have been long debated (e.g., [49]). For instance, artists argued that censorship of nudity in artwork prevents them from contributing to the creative community and engaging with other artists [63]. Our findings highlighted a similar tension between GAI tool usability, creativity, and safety in content moderation enforcement.

One direct approach to promote user safety while ensuring GAI usability is to **improve the accuracy and context awareness of GAI moderation systems**, as stakeholders have constantly worked on (e.g., [93]). We found user frustrations about outright user request denials and GAI output blocks that completely disrupt normal use, especially when caused by moderation system decision-making failures (§6.2.1). Therefore, GAI products could **employ soft moderation for user**

input and GAI output. In online communities, soft moderation is when a platform issues a warning of content or decreases the content visibility in recommendation feeds, instead of direct content removals or account bans [74, 92]. Similar strategies have already been adopted in real-world GAI products, with evidence indicating that they can improve both usability and user satisfaction. A prior work found that users are most frustrated with direct denials without explanations when making LLM requests, but are more satisfied when a diverted task from the original request is fulfilled [88]. We, furthermore, recommend a primary use of soft moderation if there is a chance of moderation system failures. That is, when there is a high likelihood of falsely moderating input or output, the system should mask the content, issue a warning, or generate a modified output, instead of completely blocking or removing the content.

Our findings also highlight user-reported issues of moderation systems making decisions without sufficient awareness of task context (§6.2.1). In response, GAI products could also **deploy personalized content moderation guardrails**, where users can adjust contextualized input and output filters for different tasks. This could be similar to the approach used in online communities for users to configure their safety preferences and personalized filters to customize if, and to what extent, they are exposed to disturbing content [39]. We note that personalized content moderation guardrails carry potential risks, including abuse by malicious users and misuse by minors. Therefore, deploying this feature requires careful consideration of the extent to which guardrails can be personalized, limiting access to users with a positive usage history (i.e., no malicious use or jailbreaking attempts), and incorporating features like parental controls.

7.3 Make Moderation Pipeline Transparent

When using GAI online tools for AIGC creative tasks, users not only encounter moderation system failures that result in unreasonable decisions but also experience frustration due to the lack of explanations for these decisions, as well as the absence of an effective user appeal process (§6.2.2). The lack of transparency and user support in the moderation pipeline has already been widely discussed in the context of online communities [59, 83]. We, however, consider this situation to be more detrimental to the user experience of GAI products, given how moderation systems in GAI products function at every stage from user input through output generation progress, to output endpoint. This is unlike online communities that solely host and moderate user-generated content.

Therefore, we suggest that GAI products implement a supportive pipeline for users after they are moderated. They should **provide clear explanations and customer support after moderation** about which stage their generation request was moderated and which moderation criteria were possibly breached. This would help users to interpret at which stage they were moderated and reason about whether a moderation decision results from randomly generated malicious output or stems from inherent biases and inaccuracies within the moderation algorithm. This will also help users understand why their content or accounts were restricted, offering transparency and clarity in the decision-making process.

Additionally, we found user complaints about the insufficient details of moderation methods in policies, which aligns with our findings from the policy analysis (§4.2.1, §6.2.2). To address this issue, GAI product service providers should **further elaborate on how the moderation system operates in each stage**. This is especially critical for automatic methods, as content moderation criteria in policies and metrics used by algorithmic moderation are sometimes misaligned [4]. If an approach using implicit rules, like blocking input with specific words, is adopted, policies should also present these implicit rules, such as making the complete ban word list visible to users.

8 Conclusion

We analyzed GAI online tool content moderation policies, finding that these policies resemble those of online communities but place emphasis on governing inputs and outputs, employing unique detection methods and response strategies for problematic content. While policies in GAI online tools comprehensively outline content moderation practices from criteria to enforcement, they lack provisions for user-driven moderation methods and appeal pipelines. We also analyzed public discussions about the GAI online tool moderation for AIGC creative tasks on Reddit. We found that while moderation systems effectively block malicious AIGC creation, users frequently discuss instances where these systems fail to make justified decisions and do not provide enough information about moderation decisions. We suggest that the GAI product policy structure can be improved, with more information on users' roles in content moderation and user appeals, and providing better explanations for why and when moderation occurs. Future work can study additional GAI products and conduct user studies to build on our findings.

Ethical Considerations

We identified no ethical concerns in the policy analysis study, as all the web pages we collected were publicly available, not linked to any individuals, and gathered without the use of scraping tools. We also did not require ethical reviews for the policy analysis study. The Reddit study was reviewed and approved by our institutional IRB before data collection. To protect the privacy of users who made Reddit posts and comments we collected, we removed all identifiable information and slightly adjusted the wording of each quote we reported in the paper.

Open Science

We have made the following outcomes from our paper publicly available: the policy dataset (screenshots of pages collected for Study 1: Policy Analysis) with the analysis outcome (the codebook and annotated policy segments); and the Reddit dataset (Reddit posts collected for Study 2: Reddit Study) with the codebook for analysis. This artifact can be found at https://doi.org/10.6084/m9.figshare.29257187.

We have not made comments corresponding to the collected Reddit posts publicly available due to the ethical considerations stated above.

References

[1] Soyun Ahn, Jeeyun Baik, and Clara Sol Krause. Splintering and centralizing platform governance: how facebook adapted its content moderation practices to the political and legal contexts in the united states, germany, and south korea. *Information, Communication & Society*, 26(14):2843–2862, 2023.

[2] Ruth Elisabeth Appel. Generative ai regulation can learn from social media regulation. *arXiv preprint arXiv:2412.11335*, 2024.

[3] David Armstrong, Ann Gosling, John Weinman, and Theresa Marteau. The place of inter-rater reliability in qualitative research: An empirical study. *Sociology*, 31(3):597–606, 1997.

[4] Arnav Arora, Preslav Nakov, Momchil Hardalov, Sheikh Muhammad Sarwar, Vibha Nayak, Yoan Dinkov, Dimitrina Zlatkova, Kyle Dent, Ameya Bhatawdekar, Guillaume Bouchard, et al. Detecting harmful content on online platforms: what platforms need vs. where research efforts go. *ACM Computing Surveys*, 56(3):1–17, 2023.

[5] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[6] Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. Like trainer, like bot? inheritance of bias in algorithmic content moderation. In *Social Informatics: 9th International Conference, SocInfo 2017, Oxford, UK, September 13-15, 2017, Proceedings, Part II 9*, pages 405–415. Springer, 2017.

[7] Nicole Buckley and Joseph S Schafer. 'censorship-free'platforms: Evaluating content moderation policies and practices of alternative social media. 2022.

[8] Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S Yu, and Lichao Sun. A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. *arXiv preprint arXiv:2303.04226*, 2023.

[9] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proc. ACM Hum.-Comput. Interact.*, 1(CSCW), December 2017.

[10] Chen Chen, Jie Fu, and Lingjuan Lyu. A pathway towards responsible ai generated content. *arXiv preprint arXiv:2303.01325*, 2023.

[11] CNBC. Chatgpt blocked 250,000 image generations of presidential candidates. https://www.cnbc.com/2024/11/08/chatgpt-blocked-250000-image-generations-of-presidential-candidates.html, 2024. Accessed: 2025-01-08.

[12] Josef Dai, Tianle Chen, Xuyao Wang, Ziran Yang, Taiye Chen, Jiaming Ji, and Yaodong Yang. Safesora: Towards safety alignment of text2video generation via a human preference dataset. *arXiv preprint arXiv:2406.14477*, 2024.

[13] Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*, 2023.

[14] Wesley Hanwen Deng, Claire Wang, Howard Ziyu Han, Jason I Hong, Kenneth Holstein, and Motahhare Eslami. Weaudit: Scaffolding user auditors and ai practitioners in auditing generative ai. *arXiv preprint arXiv:2501.01397*, 2025.

[15] Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. First i" like" it, then i hide it: Folk theories of social feeds. In *Proceedings of the 2016 cHI conference on human factors in computing systems*, pages 2371–2382, 2016.

[16] Xianzhe Fan, Qing Xiao, Xuhui Zhou, Jiaxin Pei, Maarten Sap, Zhicong Lu, and Hong Shen. User-driven value alignment: Understanding users' perceptions and strategies for addressing biased and discriminatory statements in ai companions. *arXiv preprint arXiv:2409.00862*, 2024.

[17] Casey Fiesler, Jessica L. Feuston, and Amy S. Bruckman. Understanding copyright law in online creative

communities. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, page 116–129, New York, NY, USA, 2015. Association for Computing Machinery.

[18] Casey Fiesler, Jialun Jiang, Joshua McCann, Kyle Frye, and Jed Brubaker. Reddit rules! characterizing an ecosystem of governance. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12, 2018.

[19] Casey Fiesler, Cliff Lampe, and Amy S. Bruckman. Reality and perception of copyright terms of service for online content creation. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, page 1450–1461, New York, NY, USA, 2016. Association for Computing Machinery.

[20] Casey Fiesler, Michael Zimmer, Nicholas Proferes, Sarah Gilbert, and Naiyan Jones. Remember the human: A systematic review of ethical considerations in reddit research. *Proceedings of the ACM on Human-Computer Interaction*, 8(GROUP):1–33, 2024.

[21] FlexOS. [report] generative ai top 150: The world's most used ai tools (feb 2024). https://www.flexos.work/learn/generative-ai-top-150, 2025. Accessed: 2025-01-08.

[22] Forbes. New research shows chatgpt reigns supreme in ai tool sector. https://www.forbes.com/sites/chriswestfall/2023/11/16/new-research-shows-chatgpt-reigns-supreme-in-ai-tool-sector/, 2023. Accessed: 2025-01-08.

[23] Francisco José García-Peñalvo. Generative artificial intelligence and education: An analysis from multiple perspectives. *Education in the Knowledge Society*, 25:e31942, 2024.

[24] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online, November 2020. Association for Computational Linguistics.

[25] Tarleton Gillespie. Governance of and by platforms. *SAGE handbook of social media*, pages 254–278, 2017.

[26] Tarleton Gillespie. *Custodians of the Internet: Platforms, Content Moder ation, and the Hidden Decisions that Shape Social Media*. Yale University Press, 2018.

[27] Tarleton Gillespie. Generative ai and the politics of visibility. *Big Data & Society*, 11(2):20539517241252131, 2024.

[28] Tarleton Gillespie, Patricia Aufderheide, Elinor Carmi, Ysabel Gerrard, Robert Gorwa, Ariadna Matamoros-Fernández, Sarah T Roberts, Aram Sinnreich, and Sarah Myers West. Expanding the debate about content moderation: Scholarly research agendas for the coming policy debates. *Internet Policy Review*, 9(4):1–29, 2020.

[29] Github. Awesome generative ai. https://github.com/steven2358/awesome-generative-ai, 2025. Accessed: 2025-01-08.

[30] Eric Goldman. Content moderation remedies. *Mich. Tech. L. Rev.*, 28:1, 2021.

[31] Juan Felipe Gomez, Caio Machado, Lucas Monteiro Paes, and Flavio Calmon. Algorithmic arbitrariness in content moderation. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 2234–2253, 2024.

[32] Philipp Hacker, Andreas Engel, and Marco Mauer. Regulating chatgpt and other large generative ai models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1112–1123, 2023.

[33] Oliver L Haimson, Daniel Delmonaco, Peipei Nie, and Andrea Wegner. Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–35, 2021.

[34] Shengnan Han, Eugene Kelly, Shahrokh Nikou, and Eric-Oluf Svee. Aligning artificial intelligence with human values: reflections from a phenomenological perspective. *AI & SOCIETY*, pages 1–13, 2022.

[35] Yiqing Hua, Shuo Niu, Jie Cai, Lydia B Chilton, Hendrik Heuer, and Donghee Yvette Wohn. Generative ai in user-generated content. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2024.

[36] Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, et al. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*, 2023.

[37] Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. "did you suspect the post would be removed?": Understanding user reactions to content removals on reddit. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), November 2019.

[38] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. Does transparency in moderation really matter? user behavior after content removal explanations on reddit. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), November 2019.

[39] Shagun Jhaver, Alice Qian Zhang, Quan Ze Chen, Nikhila Natarajan, Ruotong Wang, and Amy X Zhang. Personalizing content moderation on social media: User perspectives on moderation choices, interface design, and labor. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–33, 2023.

[40] Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36, 2024.

[41] Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*, 2023.

[42] Jialun 'Aaron' Jiang, Skyler Middler, Jed R. Brubaker, and Casey Fiesler. Characterizing community guidelines on social media platforms. In *Companion Publication of the 2020 Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '20 Companion, page 287–291, New York, NY, USA, 2020. Association for Computing Machinery.

[43] Haibo Jin, Andy Zhou, Joe D Menke, and Haohan Wang. Jailbreaking large language models against moderation guardrails via cipher characters. *arXiv preprint arXiv:2405.20413*, 2024.

[44] Prerna Juneja, Deepika Rama Subramanian, and Tanushree Mitra. Through the looking glass: Study of transparency in reddit's moderation practices. *Proceedings of the ACM on Human-Computer Interaction*, 4(GROUP):1–35, 2020.

[45] Sara Kiesler, Robert Kraut, Paul Resnick, and Aniket Kittur. Regulating behavior in online communities. *Building successful online communities: Evidence-based social design*, 1:4–2, 2012.

[46] Kate Klonick. The new governors: The people, rules, and processes governing online speech. *Harv. L. Rev.*, 131:1598, 2017.

[47] Kevin Klyman. Acceptable use policies for foundation models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 752–767, 2024.

[48] Yubo Kou, Renkai Ma, Zinan Zhang, Yingfan Zhou, and Xinning Gui. Community begins where moderation ends: Peer support and its implications for community-based rehabilitation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2024.

[49] Kyle Langvardt. Regulating online content moderation. *Geo. LJ*, 106:1353, 2017.

[50] Runtao Liu, Chen I Chieh, Jindong Gu, Jipeng Zhang, Renjie Pi, Qifeng Chen, Philip Torr, Ashkan Khakzar, and Fabio Pizzati. Safetydpo: Scalable safety alignment for text-to-image generation. *arXiv preprint arXiv:2412.10493*, 2024.

[51] Henrietta Lyons, Senuri Wijenayake, Tim Miller, and Eduardo Velloso. What's the appeal? perceptions of review processes for algorithmic decisions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2022.

[52] Renkai Ma and Yubo Kou. " how advertiser-friendly is my video?": Youtuber's socioeconomic interactions with algorithmic content moderation. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–25, 2021.

[53] Renkai Ma, Yue You, Xinning Gui, and Yubo Kou. How do users experience moderation?: A systematic literature review. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–30, 2023.

[54] Yaaseen Mahomed, Charlie M Crawford, Sanjana Gautam, Sorelle A Friedler, and Danaë Metaxa. Auditing gpt's content moderation guardrails: Can chatgpt write your favorite tv show? In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 660–686, 2024.

[55] Todor Markov, Chong Zhang, Sandhini Agarwal, Florentine Eloundou Nekoul, Theodore Lee, Steven Adler, Angela Jiang, and Lilian Weng. A holistic approach to undesired content detection in the real world. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15009–15018, 2023.

[56] J. Nathan Matias, Austin Hounsel, and Nick Feamster. Software-supported audits of decision-making systems: Testing google and facebook's political advertising policies. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW1), April 2022.

[57] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for cscw and hci practice. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), November 2019.

[58] Rachel E Moran, Izzi Grasso, and Kolina Koltai. Folk theories of avoiding content moderation: How vaccine-opposed influencers amplify vaccine opposition on instagram. *Social Media+ Society*, 8(4):20563051221144252, 2022.

[59] Sarah Myers West. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, 20(11):4366–4383, 2018.

[60] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[61] Javier Rando, Daniel Paleka, David Lindner, Lennart Heim, and Florian Tramèr. Red-teaming the stable diffusion safety filter. *arXiv preprint arXiv:2210.04610*, 2022.

[62] Piera Riccio, Georgina Curto, and Nuria Oliver. Exploring the boundaries of content moderation in text-to-image generation. *arXiv preprint arXiv:2409.17155*, 2024.

[63] Piera Riccio, Thomas Hofmann, and Nuria Oliver. Exposed or erased: Algorithmic censorship of nudity in art. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2024.

[64] Sarah T Roberts. *Behind the screen: Content Moderation in the Shadows of Social Media*. Yale University Press, 2019.

[65] Pamela Samuelson. Generative ai meets copyright. *Science*, 381(6654):158–161, 2023.

[66] Sujan Sarkar. Ai industry analysis: 50 most visited ai tools and their 24b+ traffic behavior. https://writerbuddy.ai/blog/ai-industry-analysis/, 2023. Accessed: 2025-01-08.

[67] Benjamin Saunders, Julius Sim, Tom Kingstone, Shula Baker, Jackie Waterfield, Bernadette Bartlam, Heather Burroughs, and Clare Jinks. Saturation in qualitative research: exploring its conceptualization and operationalization. *Quality & quantity*, 52:1893–1907, 2018.

[68] Brennan Schaffner, Arjun Nitin Bhagoji, Siyuan Cheng, Jacqueline Mei, Jay L Shen, Grace Wang, Marshini Chetty, Nick Feamster, Genevieve Lakier, and Chenhao Tan. " community guidelines make this the best party on the internet": An in-depth study of online platforms' content moderation policies. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2024.

[69] Vera Schmitt, Jakob Tesch, Eva Lopez, Tim Polzehl, Aljoscha Burchardt, Konstanze Neumann, Salar Mohtaj, and Sebastian Möller. Implications of regulations on large generative ai models in the super-election year and the impact on disinformation. In *Proceedings of the Workshop on Legal and Ethical Issues in Human Language Technologies@ LREC-COLING 2024*, pages 28–38, 2024.

[70] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22522–22531, 2023.

[71] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[72] Joseph Seering. Reconsidering self-moderation: the role of research in supporting community-based models for online content moderation. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–28, 2020.

[73] Wai Man Si, Michael Backes, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, Savvas Zannettou, and Yang Zhang. Why so toxic? measuring and triggering toxic behavior in open-domain chatbots. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, CCS '22, page 2659–2673, New York, NY, USA, 2022. Association for Computing Machinery.

[74] Mohit Singhal, Chen Ling, Pujan Paudel, Poojitha Thota, Nihal Kumarswamy, Gianluca Stringhini, and Shirin Nilizadeh. Sok: Content moderation in social media, from guidelines to enforcement, and research to practice. In *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)*, pages 868–895. IEEE, 2023.

[75] Statista. Intention of generative artificial intelligence (gai) usage by adults in the united states as of august 2023, by type. https://www.statista.com/statistics/1461998/usa-generative-ai-usage-intention-by-type/, 2024. Accessed: 2025-01-08.

[76] Statista. Use of generative artificial intelligence (ai) programs in the united states in 2023, by use case. https://www.statista.com/statistics/1413836/use-of-generative-ai-us/, 2024. Accessed: 2025-01-08.

[77] Nicolas P Suzor. *Lawless: The secret rules that govern our digital lives*. Cambridge University Press, 2019.

[78] Madiha Tabassum, Alana Mackey, Ashley Schuett, and Ada Lerner. Investigating moderation challenges to combating hate and harassment: The case of Mod-Admin power dynamics and feature misuse on reddit. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 37–54, Philadelphia, PA, August 2024. USENIX Association.

[79] Social Media Today. Tiktok adds more generative ai features. https://www.socialmediatoday.com/news/tiktok-adds-gen-ai-image-tools-caption-suggestions/736361/, 2025. Accessed: 2025-01-19.

[80] Ioanna Tourkochoriti. The digital services act and the eu as the global regulator of the internet. *Chi. J. Int'l L.*, 24:129, 2023.

[81] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[82] U.S. House of Representatives. United states code: Title 15, section 9401 (preliminary edition). https://uscode.house.gov/view.xhtml?req=(title:15%20section:9401%20edition:prelim), 2025. Accessed: 2025-01-08.

[83] Kristen Vaccaro, Christian Sandvig, and Karrie Karahalios. " at the end of the day facebook does what itwants" how users experience contesting algorithmic content moderation. *Proceedings of the ACM on human-computer interaction*, 4(CSCW2):1–22, 2020.

[84] Kristen Vaccaro, Ziang Xiao, Kevin Hamilton, and Karrie Karahalios. Contestability for content moderation. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2), October 2021.

[85] Peiran Wang, Qiyu Li, Longxuan Yu, Ziyao Wang, Ang Li, and Haojian Jin. Moderator: Moderating text-to-image diffusion models through fine-grained context-based policies. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 1181–1195, 2024.

[86] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

[87] Yiluo Wei and Gareth Tyson. Understanding the impact of ai-generated content on social media: The pixiv case. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6813–6822, 2024.

[88] Joel Wester, Tim Schrills, Henning Pohl, and Niels van Berkel. "as an ai language model, i cannot": Investigating llm denials of user requests. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2024.

[89] Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Hong Lin. Ai-generated content (aigc): A survey. *arXiv preprint arXiv:2304.06632*, 2023.

[90] Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. Fine-grained human feedback gives better rewards for language model training. *Advances in Neural Information Processing Systems*, 36:59008–59033, 2023.

[91] Zhiyuan Yu, Xiaogeng Liu, Shunning Liang, Zach Cameron, Chaowei Xiao, and Ning Zhang. Don't listen to me: Understanding and exploring jailbreak prompts of large language models. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 4675–4692, Philadelphia, PA, August 2024. USENIX Association.

[92] Savvas Zannettou. " i won the election!": an empirical analysis of soft moderation interventions on twitter. In *Proceedings of the international AAAI conference on web and social media*, volume 15, pages 865–876, 2021.

[93] Jingyu Zhang, Ahmed Elgohary, Ahmed Magooda, Daniel Khashabi, and Benjamin Van Durme. Controllable safety alignment: Inference-time adaptation to diverse safety requirements. *arXiv preprint arXiv:2410.08968*, 2024.

[94] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023.

[95] Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–20, 2023.

Appendix

See the page below for appendix tables.

	Dataset		Sample for Analysis	
Subreddit	Post	Comment	Post	Comment
r/chatgpt	335	17090	38	2069
r/midjourney	216	4383	24	473
r/dalle2	240	5423	33	686
r/claudeAI	169	4230	17	280
r/Bard	142	2223	16	322
r/dalle	14	75	1	6
r/perplexity_ai	7	41	1	3
Sum	1123	33465	130	3839

Table 2: Statistics of Reddit dataset and random sample for analysis

Restrictions	**Examples of Over-sensitive Input Moderation**
Harmful Content	*I tried earlier to generate an image of a mother and child but it refused to do so. I took the child out of the request and it worked great. Not being able to generate images of minors in compromising situations is one thing, but to filter it out entirely is too restricting.* (Gemini, r/Bard)
Content that Violates Other's Rights	*I'm just trying to brainstorm some lyric ideas for AI music and half the time it works fine other half Claude tells me they can't write music. I'm not even telling it to copy someone's style or anything. I usually give it a rough few lines I wrote and then it spits nothing out.* (Claude, r/ClaudeAI)
Sexual	*I had issues simply with "the 2 characters' foreheads are touching in a display of tenderness and affection."* (ChatGPT's DALL-E, r/chatgpt)
Violence	*I tried to make a kitchen image. Cutting board was banned due to cutting.* (Midjourney, r/midjourney)

Table 3: Content moderation criteria and corresponding examples of false-positive input moderation. Note that the criteria 'content that is not appropriate for everyone' is split into 'sexual' and 'violence' when elaborating examples.

Restrictions	**Examples of Over-restrictions**
Harmful Content	*Then you mention the bank robber likes to kick puppies. Or that they're actively targeting a bank in a spot with a vulnerable population, which is a very logical thing for a bank robber to do [...] "Oh no, I can't write anything condoning kicking puppies or assaulting vulnerable people."* (Claude, r/claudeAI)
Content that Violates Other's Rights	*I tried to put a photo of myself in the editor and it said 'This violates our terms of service.'* (Midjourney, r/midjourney)
Misleading Content	*I asked it to generate an image of the 6th Army sieging Kingslanding and it refused because the 6th Army is real and Kingslanding is fantasy.* (Gemini, r/Bard)
Sexual	*All of my generated articles were just removed after I included a passage mentioning an adult work on sexuality and gender studies in my prompt.* (ChatGPT, r/chatgpt)
Violence	*The censorship has gotten so out of hand that I can't generate pictures with a halloween theme because they have 'blood' in the prompt.* (Midjourney, r/midjourney)
Value Alignment	*I have a fight scene and if the hero ever even remotely has anything bad happen, they instantly rebound with great courage and become a role model to everyone, vanquishing the baddie for all time.* (ChatGPT, r/chatgpt)

Table 4: Content moderation criteria and corresponding examples of over-restrictions. Note that the criteria 'content that is not appropriate for everyone' is split into 'sexual' and 'violence' when elaborating examples. Value alignment is also included here.