From Underground to Mainstream Marketplaces: Measuring AI-Enabled NSFW Deepfakes on Fiverr

Mohamed Moustafa Dawoud
University of California, Santa Cruz
mdawoud@ucsc.edu

Alejandro Cuevas
Princeton University
cuevas@princeton.edu

Ram Sundara Raman
University of California, Santa Cruz
rsundar2@ucsc.edu

***Abstract*—Generative AI has enabled the large-scale production of photorealistic synthetic sexual imagery, yet prior work on non-consensual intimate imagery and deepfakes has focused mostly on underground forums and dedicated nudification tools. In this paper, we investigate whether these services have moved into *mainstream gig marketplaces*, where they benefit from larger user bases and higher trust.**

We present the first systematic study of sexually explicit AI generation services (often advertised as AI NSFW services) on a major freelance marketplace, Fiverr. We discover these listings by employing a range of sampling approaches, including keyword searches, sitemap analysis, and snowball sampling, and confirm that they are sexually explicit through an LLM classifier. Through this process we identify 593 AI-enabled NSFW gigs. We also collect a set of control groups from other AI and non-AI categories (n=1,028). We use an LLM to extract each gig's risk indicators, advertised tools, platform targets, pricing, and seller attributes.

Our results reveal a rapidly emerging market with new NSFW service freelancers joining at consistently higher rates than any other group we observed (74.9% of NSFW sellers joined in 2025). Within the NSFW segment, 82.8% expose deepfake-enabling features and 87.6% violate Fiverr's policies on pornography and deepfakes. We also uncover a new type of service, not previously documented: custom sexually explicit LoRA/model training. Sellers disproportionately target downstream platforms such as OnlyFans (54.2%), Instagram (29.5%), and Fanvue (24.1%). For the usable security and privacy community, our results reframe abuse-enabling generative AI as a mainstream problem rather than a dark corner of the Internet.

I. Introduction

Generative AI tools like Stable Diffusion [1], ComfyUI [2], and LoRA fine-tuning [3] have enabled the generation of hyper-realistic, highly personalized imagery at extremely low cost. For some users, this is a story about creative empowerment: AI-assisted illustration, avatar design, and content workflows. But the same tools also enable new avenues of online harm at scale: non-consensual synthetic intimate imagery (NSII), synthetic persona fraud, and cross-platform manipulation. Recent work identifies deepfake pornography and NSII as a distinct class of AI-driven privacy exposure and distortion harms [4].

Crucially, these capabilities no longer live only in underground forums or bespoke "nudification" apps. The landscape has evolved through multiple waves of enforcement and migration. Early standalone tools like DeepNude shut down in June 2019 after viral backlash [5]. Dedicated sexual-deepfake platforms like MrDeepFakes—which Han et al. characterize as operating a structured five-year marketplace with seller reputations and custom-order systems [6]—faced mounting pressure from service providers and legislators. In May 2025, just days after Congress passed the "Take It Down Act" [7] criminalizing non-consensual intimate imagery distribution, MrDeepFakes announced permanent shutdown after a critical service provider terminated support [8]. Yet experts feared that these communities will re-emerge somewhere else. Our paper investigates this question: *has enforcement against dedicated deepfake platforms pushed these services to reappear as commercial gigs on general-purpose freelance marketplaces?*

Most research on deepfakes and image-based sexual abuse has focused on *exceptional* spaces: dark web markets [9], [10], specialized nudification services like DeepNude [5] and the broader nudification application ecosystem [11], dedicated sexual-deepfake platforms like MrDeepFakes [6], and abuse communities operating through Telegram channels and semi-closed websites [12], [13]. In this prevailing research paradigm, harmful capability is "elsewhere": visible to investigators, perhaps, but not part of everyday online work.

However, we lack a systematic understanding of whether and how abuse-enabling AI services are *embedded directly* into the same gig marketplaces where benign AI services are offered. In this paper, we study how AI-enabled NSFW content services are commercialized on a popular mainstream gig platform, Fiverr. We focus on Fiverr because it combines (1) a supply-driven marketplace where sellers publicly advertise fixed-price "gigs"; (2) rich, publicly visible data about services, pricing, and seller attributes; and (3) Terms of Service that explicitly prohibit adult and sexually explicit content and separately forbid deepfakes, impersonation, and other forms of deceptive synthetic media [14], [15]. Fiverr's scale also makes it an important site for studying commercialized AI-enabled NSFW services: the platform reports 3.5 million active buyers, 700+ service categories, and over $1.1 billion in annual gross merchandise value [16]. Taken together, these features make

Symposium on Usable Security and Privacy (USEC) 2026
27 February 2026, San Diego, CA, USA
ISBN 978-1-970672-07-7
https://dx.doi.org/10.14722/usec.2026.23068
www.ndss-symposium.org, https://www.usablesecurity.net/USEC/

Fiverr an unusually revealing vantage point for examining how AI-enabled NSFW content emerges and persists within mainstream commercial platforms.

Our user-focused study asks the following key research questions about AI-enabled NSFW content on Fiverr:

- **RQ1 (Prevalence)**: How prevalent and economically successful are NSFW AI gigs on Fiverr?
- **RQ2 (Capabilities):** What technical and commercial capabilities do these AI-enabled NSFW services offer, and how do these relate to harms ranging from basic ToS violations to enabling non-consensual intimate imagery and cross-platform abuse?
- **RQ3 (Market Structure):** How do NSFW AI services differ from mainstream gigs in terms of growth, revenue concentration, technology choices, and geography?

To answer these questions, we perform a comprehensive dual-track measurement study of Fiverr. We collect and analyze 1,838 service listings using a sampling design that emulates user searches for AI and NSFW services. The first track uses NSFW-targeted keyword searches (21 query variants such as "AI NSFW," "AI OnlyFans," and euphemistic terms like "spicy AI content"), yielding 810 resulting gigs. After validation, we find that 73.2% of these gigs (n=593) contain confirmed NSFW content ("NSFW gigs"). The second track samples four mainstream categories surfaced through Fiverr's own taxonomy: AI Artists (n=266), AI Avatar Design (n=249), AI Image Editing (n=228), and a non-AI Logo Design control (n=285). Our dual-track approach lets us compare mainstream and NSFW AI services and assess whether AI-enabled adult content is ambient across AI offerings or concentrated in a distinct segment reached through targeted searches. We then apply a structured content analysis pipeline assisted by Large Language Models (LLMs) to extract 47 fields per gig (titles, descriptions, package details, and visible metadata). The schema captures features such as service type (e.g., AI influencer creation, LoRA training); technology stack (e.g., Stable Diffusion, Midjourney); target downstream platforms (e.g., OnlyFans, Instagram, Fanvue, TikTok); risk indicators (e.g., deepfake risk, platform policy violation); and commercial metrics (prices, reviews, seller level, join date, country). To assess reliability, we manually validate a stratified sample of 200 gigs (100 NSFW, 100 SFW).

Our findings show that Fiverr's AI content ecosystem is bifurcated, with sharp structural differences between mainstream and NSFW AI-enabled gigs. While 73.2% (593 of 810) of NSFW-targeted searches resulted in AI-enabled NSFW gigs, we also found that 3.1% of all gigs surfaced in mainstream AI categories are also AI-enabled NSFW gigs. This pattern indicates that AI-enabled adult content is *discoverable but contained*, concentrated in a well-defined NSFW segment and in specific "boundary" categories where avatar and influencer work blur into sexualized services.

Within the NSFW gigs segment, we document three escalating tiers of concern that matter directly for governance. At the first tier (*policy violations*), 87.6% of NSFW gigs contain explicit AI-enabled adult content that violates Fiverr's stated prohibitions. At the second tier (*abuse-enabling capabilities*), 82.8% of NSFW gigs expose deepfake-enabling features such as face swaps, reference-photo-based generation, or custom model creation; 20.7% explicitly offer custom LoRA or model training; and 1.4% require reference photos of real individuals. As demonstrated in our case study of an NSFW gig from our dataset (§IV), these configurations can be used to generate non-consensual synthetic intimate imagery from client-supplied reference photos. At the third tier (*downstream platform harms*), NSFW gigs are explicitly optimized for other downstream platforms: 54.2% mention OnlyFans, 29.5% Instagram, 24.1% Fanvue, and 8.6% TikTok, positioning Fiverr as an upstream supplier of synthetic persona content that may violate downstream community standards.

We also find that the NSFW segment is not just different in what it offers, but in *who* participates and *how* value is distributed. NSFW sellers are entering the market far more quickly than mainstream AI providers: 74.9% of NSFW sellers joined Fiverr in 2025, compared to 28.0% in mainstream categories, a 2.7× higher new-entrant rate. Revenue is highly unequal: NSFW gigs show a Gini coefficient of 0.83, with the top 10% of sellers capturing 60.7% of estimated revenue, compared to 0.72–0.76 in mainstream categories. Technologically, NSFW gigs invert mainstream preferences: while AI Artists heavily favor Midjourney (47.4%) with Stable Diffusion as secondary (21.1%), NSFW gigs favor Stable Diffusion (24.3%) and toolchains like ComfyUI (12.6%) and Runway ML (12.4%), with Midjourney mentioned in only 4.2% of NSFW listings and DALL-E in 0.5%. We note that the tools used heavily by NSFW gig sellers are primarily open and locally controllable, thereby being less constrained by platform-level content moderation. Geographically, mainstream categories match Fiverr's typical South-Asian-heavy distribution (e.g., Pakistan 26.7–36.5%, Bangladesh up to 23.7%) [16], whereas NSFW services are disproportionately Western: the United States (14.3%) and United Kingdom (12.8%) together account for 27.1% of NSFW sellers, nearly double the mainstream Western share.

Contributions. This work makes the following contributions to the usable security and privacy analysis of AI-enabled NSFW content in mainstream gig platforms:

1) We present the first systematic prevalence and characterization study of adult-oriented AI content services on a mainstream gig marketplace, Fiverr, showing that abuse-enabling capabilities exist despite potentially violating the platform's Terms of Service.
2) We characterize NSFW gigs along three escalating tiers of concern—ToS-violating content (87.6%), NSII-enabling capabilities (82.8% deepfake-enabling, 20.7% custom model training), and downstream platform harms (59.0% targeting platforms like OnlyFans and Instagram)—providing a more nuanced analysis than binary "harmful/not harmful" classifications.
3) We reveal systematic structural differences between NSFW and mainstream segments in technology adoption, highlighting how open-source and locally controllable

tools are preferentially used to circumvent content restrictions while leveraging the legitimacy of mainstream infrastructure.
4) We show that AI-enabled NSFW gigs are growing over time at a faster rate than mainstream AI gigs and that AI-enabled NSFW content are concentrated in revenue, with average pricing similar to that of mainstream AI-based gigs.

Together, these contributions recast adult AI services not only as an external "dark" ecosystem, but as a set of abuse-enabling capabilities woven into mainstream marketplaces that ordinary users and creators already trust. For usable security and privacy, this shifts the challenge from merely detecting harmful websites to designing discovery flows, moderation tools, and policy interventions that reflect how people actually encounter, commission, and govern AI-generated intimate imagery in practice. The success of our approach suggests that LLM-assisted classification may not only help with discovering AI-based NSFW content, but also in content moderation, as supported by previous work [17]. We hope our work encourages the continued monitoring of AI-enabled abuse on mainstream gig platforms.

II. Background & Related Work

In this section, we situate our work at the intersection of deepfakes and non-consensual synthetic intimate imagery as forms of image-based sexual abuse and the broader commercialization of deepfake and AI-enabled abuse-as-a-service.

A. Deepfakes, NSII, and Image-Based Sexual Abuse

Deepfakes—AI-generated synthetic media that transplant a person's face or likeness into fabricated images or videos—have become a central vector for image-based sexual abuse. Early investigations reported that roughly 96% of deepfake videos circulating online in 2019 were pornographic and overwhelmingly targeted women without consent [13]. Subsequent legal and socio-technical scholarship has framed non-consensual deepfake pornography as a form of image-based sexual abuse (IBSA), emphasizing the gendered, enduring, and networked nature of the harm [18], [19].

Recent empirical work has shifted from anecdotal case studies to population-level measurement. Umbach et al. propose the term *non-consensual synthetic intimate imagery* (NSII) and, using a survey of over 16,000 respondents in ten countries, estimate that 2.2% of respondents report NSII victimization and 1.8% report perpetration [20]. Their findings show that synthetic intimate abuse is already non-trivial in prevalence and persists even in jurisdictions that have enacted specific NSII legislation. Flynn et al. similarly document "sexualized deepfake abuse," analyzing perpetrator and victim perspectives on the motivations, targets, and dynamics of non-consensually created and shared sexualized deepfake imagery [21]. These studies situate deepfake pornography within broader patterns of technology-facilitated sexual violence and highlight that women, public figures, and marginalized groups are disproportionately targeted.

The regulatory landscape is evolving but fragmented. Many jurisdictions have updated IBSA frameworks or introduced deepfake-specific provisions, yet gaps remain. In the United States, nearly all states now have some form of non-consensual intimate image law, with a growing subset explicitly addressing sexual deepfakes. At the federal level, the 2025 "Take It Down Act" criminalizes the knowing distribution of non-consensual intimate images, including AI-generated deepfakes, and imposes removal obligations on platforms within strict time frames [7]. Policy analyses warn that such frameworks simultaneously expand remedies for victims and raise concerns about over-removal and reliance on automated filters [22]. Overall, this line of work establishes that deepfake pornography and NSII are recognized, gendered harms, but focuses primarily on *end-user* victimization and legal remedies rather than the commercial infrastructures that make abuse scalable.

B. Commercialization of Deepfake and AI Abuse-as-a-Service

A growing body of research examines the commodification of deepfake and AI-enabled abuse. Technical reports already note a shift from hobbyist experimentation to "marketplace services," in which individual deepfake creators advertise custom video fabrication on forums and semi-closed websites [13]. Subsequent industry and academic reports have traced the emergence of deepfake services in underground forums, Telegram channels, and dedicated sexual-deepfake platforms, documenting price structures, target demographics, and specialization (e.g., face-swaps, "nudification," and voice cloning) [9], [10], [12].

Most recently, Han et al. conduct a large-scale measurement of the MrDeepFakes ecosystem, combining forum data and video attributes to characterize buyers, sellers, and targets over five years [6]. They show that sexual deepfakes form a structured marketplace with repeat sellers, reputation systems, and explicit price lists. However, prior work largely focuses on underground or single-purpose sexual-deepfake communities. Established research on gig-economy platforms such as Fiverr, Upwork, and other labor markets has examined phenomena like crowdturfing, scams, and labor precarity [23]–[25]. Yet, apart from anecdotal observations, there is little systematic evidence on how mainstream freelance marketplaces host *abuse-enabling* generative AI services, particularly those that can facilitate NCII. Our work addresses this gap by treating Fiverr as a supply-driven marketplace where sellers openly advertise AI services, allowing a measurement of AI-enabled NSFW content in plain sight.

C. Platform Governance, Cross-Platform Harm, and Generative AI

The governance of generative AI abuse on mainstream platforms intersects with broader debates on online safety, integrity, and platform accountability. SoK-style work on hate, harassment, and online abuse highlights how platforms struggle to adapt legacy moderation tools to new modalities, including deepfakes and synthetic media [26], [27]. Policy and integrity organizations warn that generative AI accelerates

the *commodification* of harmful content, enabling scalable harassment, intimate privacy violations, and synthetic persona fraud that cut across platform boundaries [28].

Cross-platform harm is a central challenge: services on one platform are explicitly designed for deployment on another. NSFW and deepfake gigs commonly advertise content "for OnlyFans" or "for Instagram," making gig marketplaces upstream suppliers of downstream Terms-of-Service violations. This multi-platform, multi-actor pipeline sits in a regulatory gray zone, where synthetic adult content may be lawful but still enables non-consensual or abusive uses.

At the enforcement level, platforms rely heavily on automated moderation, typically keyword filters and nudity classifiers. These systems perform poorly when signals are euphemistic, concealed, or cross-modal (e.g., benign text paired with explicit images), limiting their ability to detect AI-enabled abuse.

Franco et al. argue that LLMs can be integrated into the moderation stack to provide context-aware classification, explainability, and improved communication with users [29]. Kumar et al. systematically evaluate GPT-3.5 and other models on multiple platform policies, showing that LLMs can approach or match human moderator performance on many text-only tasks, while also revealing inconsistencies and policy interpretation drift [17]. OpenAI reports promising results using GPT-4 to implement policy-as-prompt style moderation, enabling more consistent labeling and faster policy iteration than hand-engineered rules [30]. Beyond text, multimodal large language models (MLLMs) are increasingly evaluated for safety on images and text combined. Liu et al. review attacks and defenses for MLLM safety, highlighting that adversaries can exploit both modalities (e.g., benign text plus adversarial images) to bypass safeguards [31].

D. Summary

Overall, prior work establishes that (1) deepfake pornography and NSII are pervasive, gendered forms of image-based sexual abuse, increasingly recognized in law and policy [13], [20], [21]; (2) deepfake and AI-enabled abuse have evolved into a service economy, from underground forums to specialized sexual-deepfake platforms [6], [9], [12], [13]; and (3) LLM-based moderation offers promising but still contested tools for enforcing platform policies at scale [17], [29], [30]. Yet, we lack systematic measurement of how mainstream gig marketplaces host AI services that can facilitate NCII and other intimate privacy violations. This paper addresses this gap by providing the first large-scale prevalence study of adult-oriented AI services on a major freelance platform, analyzing how they function as infrastructure for NCII and cross-platform harm.

III. Methodology

We conduct our study by discovering NSFW-targeted AI gigs on Fiverr through keyword searches. We then compare this sample against broader AI service categories which are compiled by Fiverr.

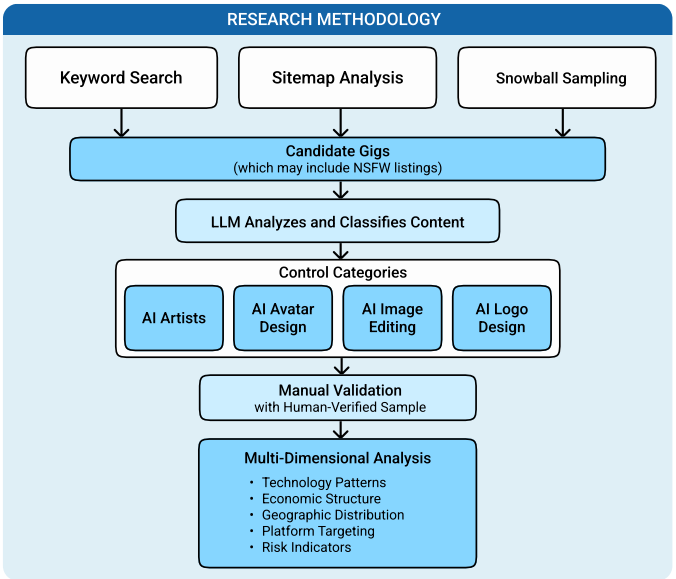

Fig. 1: Overview of our research methodology. We collect NSFW-targeted gigs through keyword search, sitemap analysis, and snowball sampling, then compare against mainstream AI categories using LLM-based classification.

A. The Fiverr Platform

Gig platforms such as Fiverr, Upwork, and Freelancer are common sites for digital labor research [32]–[34]. Fiverr, launched in 2010, is a platform that facilitates hiring freelancers for services across design, marketing, programming, and recently, AI-related offerings [35]. All gig pages are viewable without authentication. Fiverr uses fixed-price "gigs" containing service descriptions, examples, pricing tiers, delivery options, and reviews. This standardized, public structure enables consistent data collection.

We selected Fiverr because: (1) Fiverr is a mainstream marketplace and among the largest gig platforms in the world; (2) sellers generate publicly visible gig metadata; (3) profiles, pricing, and reviews are accessible without interaction. It is important to note that Fiverr prohibits adult and sexually explicit content—e.g., "nude or adult-related images or videos," "sex or sexual-partner oriented content," "sex chats," and "non-consensual content," including AI-generated material [14], [15]. These rules provide a clear ToS boundary for identifying prohibited listings.

B. Data Collection

Fiverr advertises broad categories of freelance services on their main page. For example, under the "Art & Illustration" section, users can browse options such as "Comic Illustration," "Storyboards," "AI Artists," and "AI Avatar Design". Services that may contain NSFW content are not openly advertised in these top-level categories. Thus, to discover NSFW AI services, we relied on various discovery mechanisms. We then constructed multiple control groups by sampling from top-

level AI categories (i.e., those advertised on Fiverr's homepage), as well as a non-AI category.

Identifying Potential NSFW Gigs. We employ three complementary approaches to discover and collect NSFW AI gigs:

Keyword Exploration. We first explored Fiverr's public search interface manually to seed an initial manually-constructed keyword list (e.g., "AI NSFW," "AI adult"). We then expanded this list by inspecting Fiverr's autocomplete suggestions, platform-specific terminology appearing in manually discovered gig descriptions ("OnlyFans AI," "Fanvue content"), and euphemisms identified during manual review ("spicy content AI"). This process yields 21 unique keyword combinations spanning five categories (Table I). We use these keywords to identify *candidate* AI-enabled NSFW gigs—that is, gigs that match NSFW-related search terms but require further validation to confirm they genuinely offer NSFW content (§III-C).

Sitemap Analysis. Guided by the keyword list, we crawl Fiverr's public XML sitemaps (e.g., `sitemap_gigs.xml.gz`, `sitemap_tags.xml.gz`) and filter for potential NSFW-associated terms in URL slugs or tags ("nsfw," "adult," "erotic," "onlyfans"). Because these sitemaps are explicitly listed in `robots.txt`, they provide a *robots-compliant* mechanism for enumerating gig URLs at scale.

Snowball Sampling. Finally, starting from gigs identified via sitemaps, we follow platform-provided navigation links such as "Related Gigs" to discover additional NSFW services. We restrict our crawler to gig pages and other paths allowed by `robots.txt`.

Our discovery process yields 810 unique NSFW candidate gigs after URL normalization and deduplication. We analyze and validate this set of gigs further in §III-C.

Creating Comparison Groups. To contextualize our measurements, we construct control groups that span two dimensions: (1) AI-based vs. non-AI services, and (2) generative vs. utility-focused applications. This design tests whether NSFW content is endemic to AI services broadly, concentrated in specific AI subcategories, or absent from non-AI creative work. We selected the following four categories from Fiverr's homepage:

1) **AI Artists** (n=266): Fiverr's popular AI category, representing general AI art and image generation services.
2) **AI Avatar Design** (n=249): Character and persona creation services.
3) **AI Image Editing** (n=228): Photo enhancement and manipulation, representing utility-focused AI applications distinct from generative services.
4) **Logo Design** (n=285): Non-AI creative control, establishing baseline NSFW prevalence in non-AI creative categories.

Fiverr's public search interface exposes up to 10 pages of results per query (up to 48 gigs per page, 480 maximum per category), which we confirmed through manual inspection of the UI. For automated collection, however, we rely on robots-compliant sources: category- and tag-specific entries in `sitemap_gigs.xml.gz` and other public sitemaps rather than search pagination. Sample sizes vary across categories (n=228–285) due to differences in the number of gigs linked from these category- and tag-level sitemap entries.

TABLE I: NSFW-Targeted Search Keywords List (n=21)

Category	Keywords
Explicit NSFW	"AI NSFW," "AI adult," "AI erotic," "NSFW AI art"
Platform-specific	"AI OnlyFans," "AI Fanvue," "OnlyFans content AI," "OF AI model," "Fansly AI content"
Service types	"AI influencer NSFW," "AI girlfriend," "AI girl NSFW," "AI companion"
Euphemistic	"spicy AI content," "mature AI art," "adult AI models," "AI 18+ content"
Hybrid	"NSFW image generator," "adult character AI," "18+ stable diffusion," "NSFW diffusion"

C. Classification Pipeline

Our primary goal is to quantify the number of listings that offer NSFW services. Beyond this binary NSFW classification, we also extract the technologies and services advertised, target downstream platforms for content distribution, pricing tiers, and geographic distribution of sellers.

We used Claude 3.5 Haiku to classify all 1,838 gigs. Manual coding at this scale would be time-intensive, especially for NSFW content that often appears through euphemistic phrases, indirect language, or portfolio-only signals. To ensure consistency, each gig's full HTML was parsed and fed into a structured prompt that required the model to return a nested JSON object. This approach standardized the output format and allowed downstream analysis to operate on uniform fields rather than free-form text. We developed the prompt iteratively through pilot testing on a small sample of gigs (n=50), refining instructions to reduce ambiguous outputs and ensure consistent JSON formatting. The prompt was designed to: (1) provide clear definitions of NSFW content aligned with Fiverr's ToS, (2) require explicit evidence extraction (e.g., specific phrases, platform mentions) rather than subjective judgments, and (3) use structured output fields to minimize post-processing errors. We validate our classification in §III-D.

The extracted fields fall into several groups, designed to answer our research questions: NSFW classification for RQ1 (prevalence), service and platform analysis for RQ2 (capabilities and harms), and seller metrics for RQ3 (market structure):

- **NSFW Classification**: a binary `explicit_nsfw` flag and a three-way categorical `sfw_nsfw_both` label with values `nsfw_only` (seller exclusively offers NSFW services), `both` (seller offers both SFW and NSFW options), or `sfw_only` (seller advertises SFW services only). For our prevalence analysis, we treat both `nsfw_only` and `both` gigs as NSFW, since offering

adult content—even alongside SFW options—constitutes a policy violation.

- **Service Analysis**: the primary service type along with all explicitly named technologies (e.g., Stable Diffusion, Midjourney, ComfyUI) and offerings (e.g., custom model training). This captures how sellers frame their technical capabilities.
- **Target Platforms**: explicit mentions of downstream platforms such as OnlyFans, Fanvue, Instagram, and TikTok, indicating intended use cases.
- **Seller Metrics**: pricing tiers, review counts, seller level, membership date, and country—fields already present in gig pages but normalized here for consistent analysis. We also estimate Revenue from pricing information as Revenue = (Reviews + Orders in Queue) × Basic Price.
- **Potentially Problematic Services**: We want to identify listings that explicitly advertise services that could facilitate misuse. In particular, we extract listings that advertise the creation of "deepfakes" or the usage of reference photos (e.g., the gig allows the customer to provide photos for which they want NSFW AI services). We refer to these features as *risk indicators* for the remainder of the paper.

Our full LLM-prompt is available in Appendix A. We apply the same prompt and schema to both NSFW-targeted and mainstream samples to avoid classification bias. This uniform approach is crucial for filtering false positives from keyword-based discovery. Although keyword search (§III-B) retrieved 810 NSFW candidates, LLM classification confirmed only 593 (73.2%) as genuinely NSFW; the remaining 217 (26.8%) were non-adult uses of NSFW-associated terms (e.g., disclaimers, SFW service targeting, metaphorical language) and were removed. Across both samples, the pipeline classified 1,838 unique gigs (810 NSFW candidates + 1,028 mainstream). After excluding the 217 false positives, 1,621 gigs remained (593 NSFW-targeted + 1,028 mainstream), forming the dataset primarily used in our results (§V).

D. Validation

We validated the accuracy of our LLM classification on a stratified sample of 200 gigs (100 NSFW, 100 SFW). A researcher manually reviewed each full gig page (text and portfolio images) using the same criteria as the prompt to identify false positives and false negatives. On this sample, the classifier achieved 100.0% precision (95% CI: 96.4–100.0%), 71.4% recall (95% CI: 61.4–80.1%), and an F1 score of 83.3%. Precision was perfect—no SFW gigs were misclassified—but recall indicates that 28.6% of NSFW gigs were missed. We deliberately prefer this conservative balance to accurately report a lower-bound of confirmed AI-enabled NSFW services on Fiverr.

Qualitative analysis of the 57 false negatives reveals five recurring patterns (Table II): visual-only NSFW signals (42.1%), euphemistic language (28.1%), dual-use framing—legitimate services with NSFW available "on request" (17.5%), platform dog-whistles—coded terms like "spicy content" or "OF" (OnlyFans) (8.8%), and character obfuscation via misspellings ("n$fw") or Unicode substitutions (3.5%). In 94.7% (54/57) of these gigs, there were no explicit textual NSFW markers, suggesting that most misses stem from text-only limitations rather than misinterpretation of the available text. We did not iterate further on the prompt after validation because the false negatives resulted from information that was simply not present in the text (e.g., portfolio images, implicit context). Prompt refinement cannot address these cases; instead, multimodal classification would be required in future work.

TABLE II: False Negative Analysis: NSFW Identification Evasion Strategies

Strategy	Count (%)	Example
Visual-only signals	24 (42.1%)	NSFW in images; generic text
Euphemistic language	16 (28.1%)	"spicy," "mature," etc.
Dual-use framing	10 (17.5%)	"SFW default, NSFW on request"
Platform dog-whistles	5 (8.8%)	"content creators" as coded term
Obfuscation	2 (3.5%)	Misspellings, Unicode variants

Keyword Baselines. We compared the LLM classifier to two keyword-based baselines: (1) a narrow list of explicit NSFW terms and (2) the full 21-term set in Table I. The explicit-term baseline achieved 92.3% precision but only 34.2% recall. The expanded 21-term list improved recall to 58.1% but reduced precision to 67.8%. By contrast, the LLM reached 71.4% recall with 100% precision, capturing 37.2% more NSFW content than the explicit-term baseline and 13.3% more than the expanded list, without the precision losses inherent to broader keyword matching (see Appendix C for detailed comparison). Most importantly, the 100% precision of our best-effort LLM-based approach allows us to present a lower bound of confirmed NSFW gigs in our results (§V).

E. Ethical Considerations

We follow best practices established in prior studies of online marketplaces [6], [36] and web crawling and scanning research [37]. Although Fiverr's Terms of Service forbid any automatic, manual, or systematic collection of site content [38], we argue following the beneficence principle from the Menlo Report [39] that the benefits—specifically, identifying gigs that may facilitate AI abuse—from this work far outweigh the limited risks posed by our strictly observational, robots.txt-compliant methods. Furthermore, we note and follow recommendations from recent academic work using Fiverr data that adopt similar approaches [36], as well as ethical practices for academic work involving data collection that may violate platform ToS [40]. Below we outline how we sought to minimize risks and respect the integrity of Fiverr and its users.

Our study analyzes only publicly visible Fiverr gig pages that require no authentication and are linked from Fiverr's public sitemaps. Our measurements are strictly observational:

we only read publicly advertised listings accessible to any unregistered visitor and did not upload, solicit, or transact with any content. We did not create accounts, log in, contact sellers, purchase services, or access any private data. We reviewed Fiverr's `robots.txt` (November 2025) and restricted automated collection to explicitly permitted paths, including gig and tag sitemaps (e.g., `sitemap_gigs.xml.gz`, `sitemap_tags.xml.gz`). We did not request disallowed search endpoints or private-user areas. All gig URLs were obtained from these sitemaps or from links on allowed pages. We also established conservative rate limits for scraping (7–10 second randomized delays) and fetched only essential HTML, keeping traffic well below typical human browsing rates.

Analyses are reported in aggregate, and we avoid naming individual sellers, reproducing or saving sensitive imagery, or attempting deanonymization. Lastly, we shared our findings with Fiverr prior to publication, including a curated list of gigs that appear to explicitly violate their Terms of Service. Recent work has raised concerns about the nonconsensual use of nude imagery in machine learning research [41]. Our study avoids these issues: automated classification operates on textual content only, manual validation viewed but did not store portfolio images, and we do not collect, distribute, or publish any explicit imagery.

IV. Case Study: Deepfake-Enabled NSFW Service

A. Gig Description

To demonstrate how our LLM-based classification methodology identifies NSFW content with deepfake risk, we present a detailed analysis of a representative gig from the NSFW-targeted sample. This case illustrates the intersection of three concerning patterns: (1) explicit NSFW offering with dual-use framing, (2) deepfake technology infrastructure, and (3) cross-platform fraud targeting. A blurred screenshot of how such gigs appear on Fiverr is shown in Appendix C.

Gig: "I will create photorealistic nsfw or sfw images of a person using ai"

AI Platform: Stable Diffusion
Expertise: Illustration, Photography
Pricing: Basic tier $15

About this gig
I create photorealistic images using AI—perfect for social media, avatars, book covers, or personal creations.

You provide:

- Description
- Reference photo (optional, for accuracy)

I provide:

- High-quality, detailed images
- 4K resolution (scalable)

For adult or alternative versions, please send me a private message.

All images are 100% AI-generated + retouching in Photoshop is possible.

Basic Package Description
I create for platforms like OnlyFans, Fanvue, TikTok, Instagram, Twitter (X), YouTube, and Reddit. Whether you need a face swap, background edit, AI-enhanced video, or stunning reel I deliver high-quality, custom results.

What You Get:

- Custom-designed AI model or virtual influencer
- Realistic styles using SDXL technology
- Created by a professional AI artist
- Optimized for AI Inst@gram posts and profile branding
- NSFW or SFW your choice!
- High-resolution images of your avatar for personal or commercial use
- Ideal for Inst@gram models, creators, or brand influencers

B. Binary Classification and Risk Assessment

Claude 3.5 Haiku classified this gig as **NSFW (high confidence)** based on explicit textual cues: the title includes "nsfw," the description offers "adult or alternative versions," the package states "NSFW or SFW your choice," and the gig targets OnlyFans/Fanvue. The gig also triggers a deepfake-risk flag because its "face swap" feature and optional "reference photo" enable generation of synthetic intimate imagery of real individuals.

C. Evidence Analysis

a) Abuse Infrastructure: The gig advertises photorealistic SDXL-based generation, face-swap capabilities, and reference-photo alignment, enabling non-consensual deepfake creation. Custom models support consistent outputs suitable for repeated misuse.

b) Cross-Platform Targeting: The seller explicitly targets adult platforms (OnlyFans, Fanvue) and mainstream social media (Instagram, TikTok, Twitter, YouTube, Reddit), indicating intentional multi-platform deployment.

c) Evasion Techniques: The gig employs keyword obfuscation ("Inst@gram"), dual-use framing that masks NSFW offerings behind legitimate services, and redirection to private messages for adult versions. These strategies evade keyword filters but are detectable through contextual analysis.

D. Policy Violations and Methodology Validation

The gig violates Fiverr's rules prohibiting adult content and deepfake services through its explicit NSFW offerings, face-swap tools, and targeting of OnlyFans/Fanvue, despite being listed under "AI Services."

TABLE III: NSFW prevalence in keyword-discovered candidates and mainstream Fiverr categories.

Category	Total	NSFW	% (95% CI)
NSFW-Targeted	810	593	73.2 (70.1–76.2)
AI Avatar Design	249	18	7.2 (4.6–11.1)
AI Artists	266	11	4.1 (2.3–7.3)
AI Image Editing	228	3	1.3 (0.4–3.8)
Logo Design (control)	285	0	0.0 (0.0–1.3)
Mainstream Total	1,028	32	3.1 (2.2–4.4)

E. Implications

This case highlights: (1) the integration of face-swap tools with NSFW generation, enabling NCII; (2) systematic evasion methods requiring contextual rather than keyword-based detection; and (3) multi-platform packaging that lowers barriers for deploying synthetic content across social and adult platforms.

V. Results

We analyzed 1,621 gigs across five samples: NSFW-targeted (n=593), AI Artists (n=266), AI Avatar Design (n=249), AI Image Editing (n=228), and a non-AI Logo Design control (n=285). We first summarize basic market characteristics within each category, then show how NSFW services form a structurally distinct segment in terms of prevalence, growth, capabilities, revenue concentration, geography, and platform targeting. We note that the NSFW-targeted sample represents 593 gigs confirmed as genuinely offering NSFW content after LLM classification of 810 keyword-discovered candidates (Section III-C). The remaining 217 gigs (26.8%) were false positives, and were excluded from most analysis in this section except in §V-A where we talk about the prevalence of NSFW content in searches.

A. NSFW Discoverability and Category Spillover

As shown in Table III, NSFW-targeted searches surfaced 810 candidate gigs, of which 593 (73.2%) were confirmed as NSFW after LLM classification. Only these 593 listings are used in subsequent analysis. The large number of NSFW candidates illustrate that keyword-based discovery exposes a concentrated cluster of explicit and deepfake-enabling services. NSFW content appear to a limited extent in Fiverr's mainstream categories: AI Avatar Design (7.2%), AI Artists (4.1%), AI Image Editing (1.3%), and Logo Design (0.0%). These cases mostly involve sexualized character work or NSFW variants of fictional personas, in contrast to the explicit offerings in the targeted segment (e.g., nudification, face swapping, and services accepting user-submitted photographs). Overall, NSFW content on Fiverr is highly clustered within the targeted sample, with limited spillover into mainstream categories.

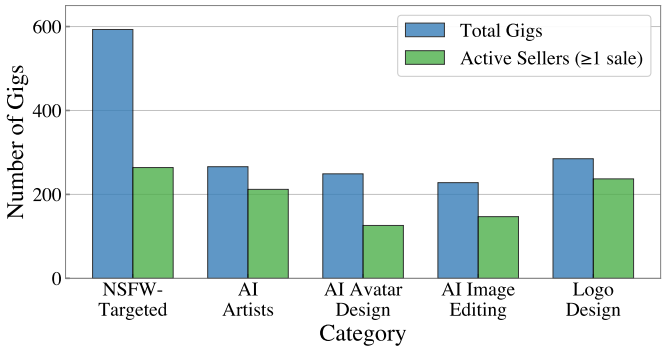

Fig. 2: Dataset overview by sampling category. Bars show total gigs and the subset of active sellers ($\geq$ 1 sale) in each category.

B. Dataset Overview

Figure 2 summarizes the total number of gigs and the share with at least one sale (active gigs) across all categories.[1] The Logo Design control shows the highest activity rate (83.2%), reflecting its status as a long-established Fiverr category. AI Artists and AI Image Editing are also relatively mature, with 79.7% and 64.5% active gigs. In contrast, the NSFW-targeted sample shows the lowest activity rate (44.5%), consistent with it being the youngest segment (74.9% joined in 2025 vs. 28.0% in mainstream AI categories).

To compare structural market characteristics across segments, Figure 3 plots the cumulative distributions for platforms targeted, basic price, total reviews, and estimated revenue. While the maximum number of platforms targeted is comparable across categories, NSFW-targeted gigs show greater concentration around multi-platform delivery, consistent with specialization toward downstream creator ecosystems. Price distributions show comparable medians but lower upper tails for NSFW gigs, suggesting a supply-heavy, competitively priced segment. The review and revenue distributions sit noticeably below those of mainstream categories, indicating that NSFW sellers have accumulated fewer sales and lower earnings overall—patterns consistent with a younger, less mature market.

C. Deepfakes and Terms of Service Violations

We examine the distribution of deepfakes and other risk indicators across categories (Table IV). In the NSFW segment, 82.8% of gigs advertise deepfake-enabling services, and 87.6% contain content that violates Fiverr's Terms of Service. Through qualitative analysis of the LLM-extracted service descriptions, we identified three recurring patterns of ToS violations: (1) explicit AI-enabled adult content targeting mainstream platforms such as OnlyFans or Instagram (e.g., "create hyper-realistic and sexy AI influencer for OnlyFans"),

[1]Activity rate reflects whether a gig has made at least one sale over its lifetime. Because our dataset is a cross-sectional snapshot from November 2025, newer gigs—especially in the NSFW segment, where 74.9% joined in 2025—have had less time to accumulate sales.

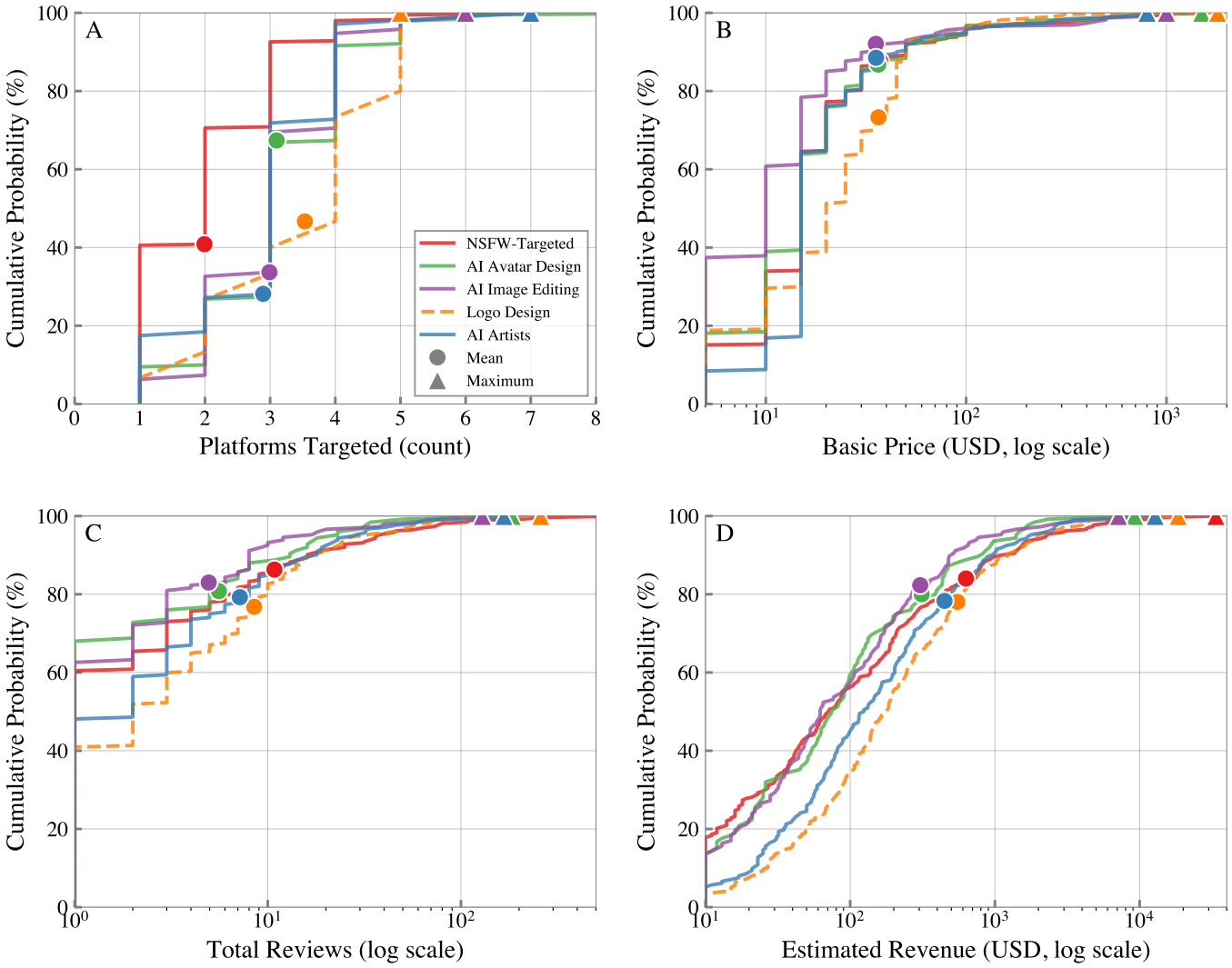

Fig. 3: Cumulative distributions across categories for (A) platforms targeted, (B) basic price, (C) total reviews, and (D) estimated revenue.

TABLE IV: Risk Indicators by Category

Indicator	NSFW	AI Artists	AI Avatar	AI Editing
Deepfake Risk	82.8%	0.8%	1.2%	0.9%
Policy Violation	87.6%	4.1%	7.2%	1.3%
Reference Photo Required	1.4%	0.0%	0.0%	0.0%
Custom Model Training	20.7%	1.1%	0.9%	1.8%

(2) custom model-training services that enable personalized deepfakes (e.g., "train NSFW AI influencer models with Flux LoRA"), and (3) workflows requiring user-submitted photographs (e.g., "set up AI OnlyFans influencer model").

Overall, 20.7% of gigs offer custom model-training capabilities, 1.4% explicitly state that they allow user-submitted reference photos, and 22.0% provide at least one of these features.

Mainstream categories (i.e., our control samples) show substantially lower prevalence across all risk indicators. Gigs advertising deepfake services remain below 1.2%, and reference-photo requirements are essentially absent. Policy violations are highest in AI Avatar Design (7.2%), likely reflecting occasional sexualized character work, while AI Image Editing shows the lowest rate among categories with any violations (1.3%), and Logo Design shows none (0.0%). Overall, these findings reveal a small but growing corner of Fiverr that uses AI to produce abuse-enabling content.

D. Market Growth

Figure 4 summarizes how quickly each segment is expanding. The NSFW category is by far the youngest and fastest-growing: 74.9% of its sellers joined in 2025, compared to 28.0% across mainstream AI categories and 21.4% in Logo

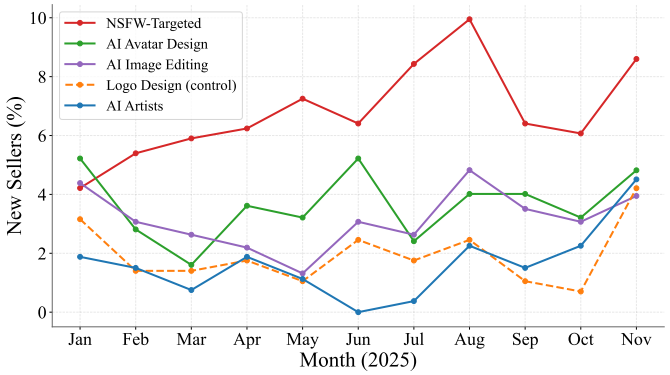

Fig. 4: Monthly new-seller rates in 2025. NSFW-Targeted growth spikes in July–August and November, whereas mainstream categories remain comparatively steady.

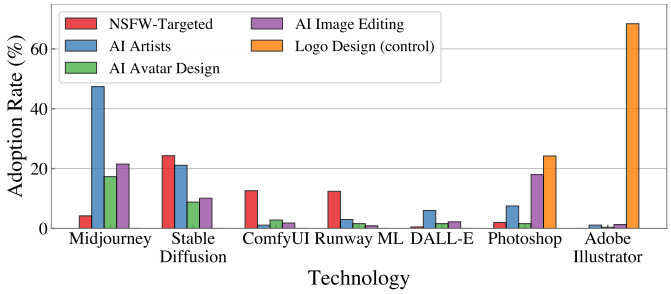

Fig. 5: AI technology adoption rates across categories. NSFW gigs predominantly use open-source tools that be locally run.

Design. NSFW onboarding accelerates over the year, with pronounced surges in mid- and late-2025 (especially July, August, and November), possibly due to the shutdown of the popular deepfake marketplace, Mr. Deepfakes [6]–[8]. On the other hand, mainstream categories exhibit flatter, steadier growth. Within the mainstream group, AI Avatar Design shows the fastest recent growth (40.2% new entrants), while AI Artists remains the most established (18.0% new entrants). Our findings show that NSFW AI content in marketplaces like Fiverr is steadily growing.

E. Technology Adoption Patterns

Technology use differs sharply across categories (Figure 5). AI Artists rely primarily on Midjourney (47.4%) and Stable Diffusion (21.1%), while AI Image Editing blends AI and traditional tools, with comparable use of Midjourney (21.5%) and Photoshop (18.0%). Expectedly, Logo Design gigs advertise almost entirely non-AI software, such as Adobe Illustrator (68.4%).

The NSFW segment shows the opposite pattern: Stable Diffusion dominates (24.3%) while Midjourney is rarely mentioned (4.2%). ComfyUI (12.6%) and Runway ML (12.4%) appear far more frequently than in any mainstream category, and tools with strict content policies (DALL-E 0.5%, Photoshop 2.0%) are nearly absent. This pattern reflects strategic selection of tools that evade content moderation, particularly

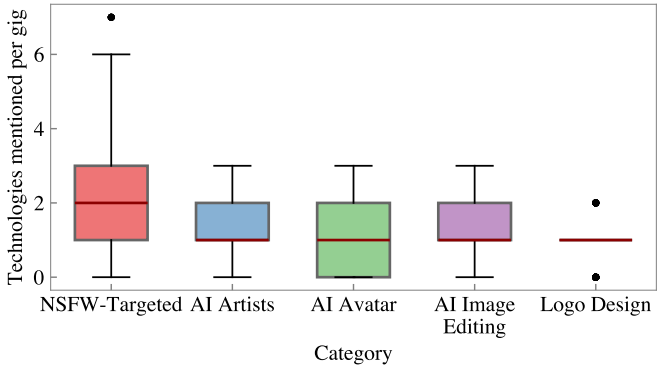

Fig. 6: Technology stack complexity. NSFW sellers use more tools per gig compared to the single-tool workflows of mainstream categories.

TABLE V: Revenue Distribution by Category (sellers with $\geq$1 sale). NSFW gig revenue is more concentrated than mainstream gigs.

Category	Mean	Median	Max	Gini
NSFW-Targeted	$575	$60	$33,660	0.83
Logo Design	$593	$185	$18,450	0.72
AI Artists	$452	$120	$12,800	0.74
AI Avatar	$387	$80	$9,240	0.76
AI Image Editing	$312	$65	$7,150	0.71

open-source tools like Stable Diffusion and ComfyUI that can be run locally without restrictions, rather than commercial services like Midjourney and DALL-E that explicitly prohibit NSFW content in their terms of service.

Figure 6 shows that NSFW sellers rely on more complex toolchains than all mainstream categories. Their distribution centers around two technologies per gig, with a wider interquartile range (IQR) and outliers reaching six or more tools, indicating multi-stage workflows (a common characteristic of ComfyUI pipelines). 86.8% of NSFW gigs mention at least one tool (mean 1.94), and 30.9% use three or more. In contrast, mainstream categories cluster around a median of one tool, with narrower IQRs and far fewer high-end outliers, reflecting generally simpler production pipelines.

F. Economic Comparison Across Categories

Revenue patterns differ substantially across categories (Table V and Figure 2 (D)). The NSFW segment shows the highest inequality (Gini 0.83), with the top 10% of sellers (ranked by estimated revenue) capturing 60.7% of total segment revenue—greater concentration than in any mainstream category.

Median prices are similar across AI categories ($10–$25), but maximum prices differ substantially. NSFW packages top out at $600 for services like custom LoRA training and ongoing content generation. In contrast, mainstream categories command higher premiums: AI Artists $800 (licensing tiers), AI Avatar $1,500 (influencer bundles), AI Image Editing $1,000 (bulk processing), and Logo Design $1,800 (brand

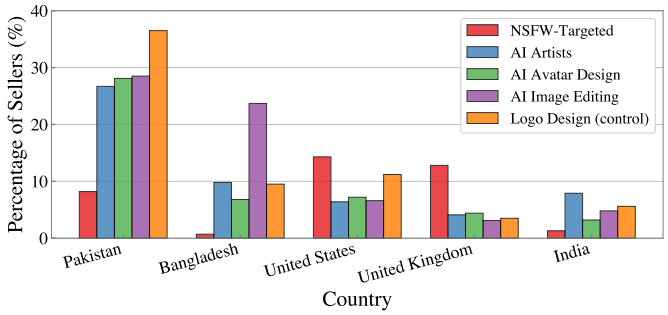

Fig. 7: Geographic distribution of sellers by category. NSFW content sellers are concentrated in the USA and UK, while mainstream categories are concentrated in South Asia.

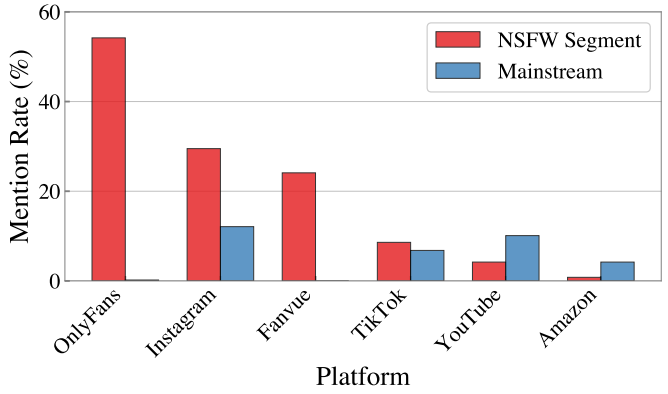

Fig. 8: Platform mentions by segment. NSFW gigs heavily target adult platforms (OnlyFans, Fanvue) while mainstream gigs reference YouTube and Amazon.

suites). These lower NSFW price caps, combined with higher revenue concentration (Gini 0.83), indicate that top NSFW sellers achieve market dominance through transaction volume rather than premium pricing.

G. Geographic Distribution

Seller geography differs sharply across categories (Figure 7). Mainstream categories follow Fiverr's typical pattern, with Pakistan (26–37%) and Bangladesh (7–24%) as the dominant supply regions. In contrast, the NSFW segment is more Western: the United States (14.3%) and United Kingdom (12.8%) together account for 27.1% of NSFW sellers, while South Asian representation is far lower. This indicates that NSFW services attract a different supplier base, concentrated in Western markets with higher AI-enabled adult-content monetization.

H. Target Platform Distribution

Downstream platform targeting differs sharply between segments (Figure 8). NSFW services prominently target adult platforms (OnlyFans, Fanvue). In contrast, mainstream gigs primarily reference commercial platforms such as YouTube and Amazon. NSFW sellers frequently co-target multiple platforms. The most common pairing is Instagram + OnlyFans (22.6%), indicating a funnel strategy that uses Instagram for discovery and OnlyFans for monetization–despite Instagram's prohibition of sexually explicit content and adult-service solicitation [42]. Fanvue + OnlyFans appears in 18.2% of NSFW gigs, and Fanvue + Instagram in 15.7%, reflecting diversification across adult and mainstream platforms.

Smaller patterns include Instagram + TikTok (2.7%) and OnlyFans + TikTok (2.5%), both involving platforms with strict anti-adult content rules. Overall, 59.0% of NSFW gigs mention at least one platform, and 35.1% target multiple platforms, averaging 1.18 platforms per gig. Our findings show that Fiverr's NSFW content sellers are not just producing explicit AI content but actively positioning it for deployment on platforms where it can bypass policies, monetize synthetic personas, and enable cross-platform abuse.

VI. Discussion

Our analyses show a rapidly emerging category of freelance labor dedicated to the creation of NSFW material using AI, despite many (if not all) of them violating Fiverr's ToS. s

Economic Insights. Many freelancers began offering AI-enabled NSFW services on Fiverr this year (∼75%), making it the fastest growing category we observed. This trend indicates that creating this type of content is profitable and potentially has a low barrier to entry. In fact, we observed several freelancers who offered completely different services before their incursion into this genre. This emerging market—i.e., the market for AI-enabled sexually explicit services— gives us a glimpse into new labor dynamics, tool adoption, and target platforms. Most importantly, this market (and eventually future ones) serves as a window through which we can study the services that customers seek and the increasing sophistication of sellers and the services that they offer.

Legislative Pressure. We found that freelancers offering NSFW AI services were mostly located in the US and the UK. This is a surprising result given the increasing pressure that legislators and payment processors are placing on platforms that host NCII (e.g., the Take it Down Act in the US [7]) and platforms that host adult content in general (e.g., the UK's Online Safety Act [43], [44]). As a result of these bills, various platforms took action to curb both NSII and NCII. Earlier this year, CivitAI and HuggingFace removed models and images that used celebrities' likeness [45]–[47], while the platform Tensor.Art prohibited all adult content [48]. Similarly, the most notorious AI-enabled NCII forum and marketplace, MrDeepfakes, shut down in May just a week after the Take it Down Act was passed by the US Congress [49]. Against this backdrop, we would expect that AI-enabled sexually explicit services would be pushed to less visible places (e.g., private Telegram and Discord channels), relying on off-shore and more resilient hosting (e.g., bulletproof hosts, Tor hidden services), as well as harder to track payment options (e.g., privacy-preserving cryptocurrencies). The emergence of freelancers who advertise services capable of facilitating NCII—including deepfake generation, face swapping, and reference-

photo-based model training—on mainstream platforms like Fiverr defies this expectation.

Potential for Abuse. Creating synthetic NSFW material using AI that do not resemble real people is not illegal. Nonetheless, these gigs are still problematic because, as they stand, they may allow people to easily access nudification, face swapping, or other AI-enabled non-consensual NSFW services. In fact, we found various listings that offer some of these capabilities. Offering these services are forbidden by Fiverr's Community Standards [15], whereby freelancers are not allowed to use AI to create deepfakes nor AI-generated "imposters" for sexual purposes. Nonetheless, they are still accessible and it is unclear what measures there exist to prevent people who hire freelancers to request and obtain access to those services.

Many of the AI NSFW gigs we studied explicitly mentioned adult platforms like OnlyFans and Fanvue as targets, as well as social media platforms like Instagram. All these platforms require creators to disclose AI-usage. However, it is unclear whether creators are abiding to these rules. Prior to the widespread usage of AI tools, academics had already documented online communities dedicated to creating fake personas—for influencing and sex work—either through manipulated images and videos, often of unsuspecting individuals (a practice they dubbed "eWhoring") [50]. Naturally, advances in AI tooling should facilitate this practice. Furthermore, there is little incentive for creators to disclose AI usage when the intention is to trick viewers into believing they are interacting with a real person.

On the other hand, using AI to create consensual sexually explicit material may be justified in some cases. For example, adult content creators may want to contract services that leverage their likeness to create content that their customers want but that they may not be able to produce. This would not be the first time that adult content creators have employed AI to support their work. In 2024, various outlets reported that OnlyFans creators were using AI chatbots—finetuned on their data to sound like them— to have sexual conversations with customers [51]. Using AI on their photos would simply constitute a change of modality and seems analogous to using AI on their texts. The key difference in both scenarios is consent. To determine whether the person who is requesting a service is able to consent to their usage, we would need a system to attribute and verify that the text, images, or audio, do indeed belong to requester. Unfortunately, such systems do not currently exist. Existing approaches instead rely on deterrents, such as laws and their enforcement [7], [52], [53], as well as notices and warnings [54], and education [55]. We encourage researchers to continue developing potential mitigations that account for benign uses.

Implications for Platform Governance. Our findings suggest several intervention points for platforms seeking to mitigate abuse-enabling services while preserving legitimate AI creativity:

- **Search-path enforcement:** Because NSFW services cluster around predictable keywords, platforms can apply stricter review thresholds and mandatory image checks within these query paths without affecting mainstream categories.
- **Reference-photo safeguards:** Services offering custom model training or reference-photo-based generation pose the highest risk for NCII. Platforms could require sellers offering these capabilities to implement consent verification workflows or prohibit reference photos of identifiable individuals.
- **Downstream platform coordination:** Given that 54.2% of NSFW gigs target OnlyFans and 29.5% target Instagram, coordinated enforcement between gig marketplaces and downstream platforms could disrupt the supply chain for synthetic persona fraud.
- **LLM-assisted moderation:** Our results show that off-the-shelf LLMs achieve 100% precision in identifying NSFW gigs, suggesting that automated pre-screening could substantially reduce moderator workload while maintaining accuracy.

Limitations. Our findings should be interpreted in light of several caveats. First, while we perform a comprehensive evaluation of a mainstream marketplace, we observe only the portion of Fiverr that is publicly visible and ethically collectible; private transaction and removed listings fall outside our view. Second, our analysis focuses exclusively on Fiverr as a representative mainstream marketplace. While Fiverr is large and influential, understanding whether similar NSFW or deepfake-enabling ecosystems exist on other gig platforms remains an important direction for future work. Third, although many listings contain explicit indicators of abuse-enabling capabilities, we cannot determine whether sellers actually produce non-consensual or otherwise prohibited content in practice. Our conclusions therefore reflect what sellers advertise rather than verified abusive behavior.

VII. Conclusion

Our empirical study shows that AI-enabled NSFW services constitute a fast-growing, structurally distinct segment on Fiverr. This segment exhibits high levels of ToS-violating content (87.6%), deepfake-enabling capabilities (82.8%), and emerging offerings such as custom model training (20.7%), while relying heavily on open-source tools that evade commercial content restrictions. The key takeaway is this: **abuse-enabling AI services are no longer confined to underground forums or dedicated deepfake platforms, rather they are now embedded within mainstream commercial infrastructure.** This migration grants these services legitimacy, discoverability, and transactional convenience that underground alternatives cannot match. For researchers, this means that studying AI-enabled abuse requires attention to mainstream platforms, not just "dark" corners of the internet. For platforms, our findings demonstrate that targeted enforcement is feasible. For policymakers, the concentration of sellers in the US and UK despite recent legislative action suggests that enforcement mechanisms require strengthening.

VIII. Acknowledgments

We thank the anonymous reviewers for their thoughtful and constructive feedback.

References

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[2] comfyanonymous (pseudonym), "ComfyUI," https://github.com/comfyanonymous/ComfyUI, 2023.

[3] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models." *ICLR*, vol. 1, no. 2, p. 3, 2022.

[4] H.-P. Lee, Y.-J. Yang, T. S. Von Davier, J. Forlizzi, and S. Das, "Deepfakes, phrenology, surveillance, and more! a taxonomy of ai privacy risks," in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–19.

[5] J. Kastrenakes, "Controversial deepfake app DeepNude shuts down hours after being exposed," The Verge. https://www.theverge.com/2019/6/27/18761496/deepnude-shuts-down-deepfake-nude-ai-app-women, Jun. 2019.

[6] C. Han, A. Li, D. Kumar, and Z. Durumeric, "Characterizing the MrDeepFakes sexual deepfake marketplace," in *34th USENIX Security Symposium (USENIX Security 25)*, 2025, pp. 5169–5188.

[7] United States Congress, "Take it down act — tools to address known exploitation by immobilizing technological deepfakes on websites and networks act," S.146, 119th Congress, 2025, https://www.congress.gov/bill/119th-congress/senate-bill/146.

[8] CBS News, "AI-generated porn site Mr. Deepfakes shuts down," https://www.cbsnews.com/news/ai-generated-porn-site-mr-deepfakes-shuts-down/, May 2025.

[9] Kaspersky Team, "How real is deepfake threat? understanding the mechanics of the darknet deepfake industry," https://www.kaspersky.com/blog/deepfake-darknet-market/48112/, May 2023.

[10] T. Starks, "Deepfakes advertised on underground markets, signaling possible shift, recorded future says," https://cyberscoop.com/deepfakes-doctored-video-audio-future/, Apr. 2021.

[11] C. Gibson, D. Olszewski, N. G. Brigham, A. Crowder, K. R. Butler, P. Traynor, E. M. Redmiles, and T. Kohno, "Analyzing the AI nudification application ecosystem," in *34th USENIX Security Symposium (USENIX Security 25)*, 2025, pp. 1–20.

[12] D. Sancho and V. Ciancaglini, "Surging hype: An update on the rising abuse of genai," https://www.trendmicro.com/vinfo/us/security/news/cybercrime-and-digital-threats/surging-hype-an-update-on-the-rising-abuse-of-genai, Jul. 2024.

[13] H. Ajder, G. Patrini, F. Cavalli, and L. Cullen, "The state of deepfakes: Landscape, threats, and impact," Deeptrace, Tech. Rep., Sep. 2019, https://regmedia.co.uk/2019/10/08/deepfake_report.pdf.

[14] Fiverr International Ltd., "Community standards: Objectionable content and user safety," https://help.fiverr.com/hc/en-us/articles/32243276170385-Community-Standards-Objectionable-content-and-user-safety.

[15] ——, "Community standards: Illegal and prohibited services." [Online]. Available: https://help.fiverr.com/hc/en-us/articles/32243195699985-Community-Standards-Illegal-and-prohibited-services

[16] Backlinko Team, "Fiverr usage and growth statistics: How many people use Fiverr?" https://backlinko.com/fiverr-users#fiverr-revenue-by-geography, 2025.

[17] D. Kumar, Y. A. AbuHashem, and Z. Durumeric, "Watch your language: Investigating content moderation with large language models," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 18, 2024, pp. 865–878.

[18] N. Henry, C. McGlynn, A. Flynn, K. Johnson, A. Powell, and A. J. Scott, *Image-based sexual abuse: A study on the causes and consequences of non-consensual nude or sexual imagery*. Routledge, 2020.

[19] N. Henry and G. Beard, "Image-based sexual abuse perpetration: A scoping review," *Trauma, Violence, & Abuse*, vol. 25, no. 5, pp. 3981–3998, 2024.

[20] R. Umbach, N. Henry, G. F. Beard, and C. M. Berryessa, "Non-consensual synthetic intimate imagery: Prevalence, attitudes, and knowledge in 10 countries," in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–20.

[21] A. Flynn, A. Powell, A. Eaton, and A. J. Scott, "Sexualized deepfake abuse: Perpetrator and victim perspectives on the motivations and forms of non-consensually created and shared sexualized deepfake imagery," *Journal of Interpersonal Violence*, p. 08862605251368834, 2025.

[22] RAINN, "Image-based sexual abuse laws: Combat nonconsensual ai deepfakes," https://rainn.org/rainns-recommendations-for-legislators/image-based-sexual-abuse-laws-combat-nonconsensual-ai-deepfakes/, 2025.

[23] K. Lee, S. Webb, and H. Ge, "The dark side of micro-task marketplaces: Characterizing fiverr and automatically detecting crowdturfing," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, no. 1, 2014, pp. 275–284.

[24] K. Huang, J. Yao, and M. Yin, "Understanding the skill provision in gig economy from a network perspective: A case study of fiverr," *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, pp. 1–23, 2019.

[25] M. Dedema and H. Rosenbaum, "Socio-technical issues in the platform-mediated gig economy: A systematic literature review: An annual review of information science and technology (arist) paper," *Journal of the Association for Information Science and Technology*, vol. 75, no. 3, pp. 344–374, 2024.

[26] R. Gorwa, R. Binns, and C. Katzenbach, "Algorithmic content moderation: Technical and political challenges in the automation of platform governance," *Big Data & Society*, vol. 7, no. 1, p. 2053951719897945, 2020.

[27] S. Jhaver, A. Q. Zhang, Q. Z. Chen, N. Natarajan, R. Wang, and A. 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*, vol. 7, no. CSCW2, pp. 1–33, 2023.

[28] S. Mehta, S. Amos, R. Thein, and S. Feliz, "Why ai may make integrity jobs harder," https://www.integrityinstitute.org/blog/why-ai-may-make-integrity-jobs-harder/, Jun. 2023.

[29] M. Franco, O. Gaggi, and C. E. Palazzi, "Integrating content moderation systems with large language models," *ACM Transactions on the Web*, vol. 19, no. 2, pp. 1–21, 2025.

[30] OpenAI, "Using gpt-4 for content moderation," https://www.openai.com/index/using-gpt-4-for-content-moderation/, Aug. 2023.

[31] X. Liu, Y. Zhu, Y. Lan, C. Yang, and Y. Qiao, "Safety of multimodal large language models on images and texts," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24), Survey Track*, 2024.

[32] M. Sargeant, "The gig economy and the future of work," *E-Journal of International and Comparative Labour Studies*, 2017.

[33] A. J. Wood, M. Graham, V. Lehdonvirta, and I. Hjorth, "Good gig, bad gig: autonomy and algorithmic control in the global gig economy," *Work, employment and society*, vol. 33, no. 1, pp. 56–75, 2019.

[34] S. Vallas and J. B. Schor, "What do platforms do? understanding the gig economy," *Annual review of sociology*, vol. 46, no. 1, pp. 273–294, 2020.

[35] Fiverr International Ltd., "About us," https://www.fiverr.com/about-us.

[36] A. Hannák, C. Wagner, D. Garcia, A. Mislove, M. Strohmaier, and C. Wilson, "Bias in online freelance marketplaces: Evidence from taskrabbit and fiverr," in *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, 2017, pp. 1914–1933.

[37] Z. Durumeric, E. Wustrow, and J. A. Halderman, "ZMap: Fast internet-wide scanning and its security applications," in *USENIX Security Symposium*, 2013.

[38] Fiverr, "Fiverr terms of service," https://www.fiverr.com/legal-portal/legal-terms/terms-of-service.

[39] M. Bailey, D. Dittrich, E. Kenneally, and D. Maughan, "The menlo report," *IEEE Security & Privacy*, vol. 10, no. 2, pp. 71–75, 2012.

[40] C. Fiesler, N. Beard, and B. C. Keegan, "No robots, spiders, or scrapers: Legal and ethical regulation of data collection methods in social media terms of service," in *Proceedings of the international AAAI conference on web and social media*, vol. 14, 2020, pp. 187–196.

[41] P. Cintaqia, A. Arya, E. M. Redmiles, D. Kumar, A. McDonald, and L. Qin, "Stop the nonconsensual use of nude images in research," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, vol. 8, no. 1, 2025, pp. 628–629.

[42] Meta Platforms, Inc., "Adult sexual solicitation and sexually explicit language," https://transparency.meta.com/policies/community-standards/sexual-solicitation/, 2025.

[43] UK Government, "Online safety act explainer," https://www.gov.uk/government/publications/online-safety-act-explainer/online-safety-act-explainer, 2023.
[44] European Parliament, "Eu ai act — first regulation on artificial intelligence," https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence, 2024.
[45] CivitAI, "Policy update: Removal of real-person likeness content," https://civitai.com/articles/15022/policy-update-removal-of-real-person-likeness-content, 2025.
[46] E. Maiberg, "Hugging face is hosting 5,000 nonconsensual ai models of real people," https://www.404media.co/hugging-face-is-hosting-5-000-nonconsensual-ai-models-of-real-people/, 2025.
[47] GIGAZINE, "5,000 AI models banned by CivitAI for generating sexual content without consent have been re-uploaded to Hugging Face," https://www.gigazine.net/gsc_news/en/20250716-nonconsensual-ai-model-on-hugging-face/, 2025.
[48] Vocal Media, "Tensor Art no longer allowing nudity or celebrity — here's what to do," https://vocal.media/art/tensor-art-no-longer-allowing-nudity-or-celebrity-here-s-what-to-do, 2025.
[49] A. Wise, "Major deepfake porn site shuts down," https://www.npr.org/2025/05/06/nx-s1-5388422/mr-deepfakes-porn-site-ai-shut-down, 2025.
[50] A. Hutchings and S. Pastrana, "Understanding ewhoring," in *2019 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2019, pp. 201–214.
[51] L. Prada, "OnlyFans models are using AI chatbots to talk dirty for them," VICE. https://www.vice.com/en/article/onlyfans-models-are-usuing-ai-chatbots-to-talk-dirty-for-them/, Dec. 2024.
[52] E. Meskys, J. Kalpokiene, P. Jurcys, and A. Liaudanskas, "Regulating deep fakes: legal and ethical considerations," *Journal of Intellectual Property Law & Practice*, vol. 15, no. 1, pp. 24–31, 2020.
[53] B. Van der Sloot and Y. Wagensveld, "Deepfakes: regulatory challenges for the synthetic society," *Computer Law & Security Review*, vol. 46, p. 105716, 2022.
[54] A. Lewis, P. Vu, R. M. Duch, and A. Chowdhury, "Deepfake detection with and without content warnings," *Royal Society Open Science*, vol. 10, no. 11, p. 231214, 2023.
[55] N. Naffi, M. Charest, S. Danis, L. Pique, A.-L. Davidson, N. Brault, M.-C. Bernard, and S. Barma, "Empowering youth to combat malicious deepfakes and disinformation: An experiential and reflective learning experience informed by personal construct theory," *Journal of Constructivist Psychology*, vol. 38, no. 1, pp. 119–140, 2025.

APPENDIX

This appendix documents the full classification prompt and the structured JSON output used in our LLM-based analysis of 1,838 Fiverr gigs.

A. Classification Prompt

This subsection documents the LLM-based classification pipeline used to analyze 1,621 Fiverr gigs. All classifications were performed using Claude 3.5 Haiku via the Anthropic API.

Each gig description was analyzed using the following prompt template. Variable fields (gig title, description, pricing, etc.) were populated from scraped data.

LLM Classification Prompt Template

You are analyzing Fiverr gig listings to identify services that offer NSFW (Not Safe For Work) content generation using AI tools. Your task is to classify each gig and extract structured information.

NSFW Definition:
A gig is considered NSFW if it explicitly offers adult/-sexual content, pornographic imagery, nude images, or services intended for adult platforms (OnlyFans, Fanvue, etc.). Evidence must be present in the text (title, description, package descriptions, seller bio).

Classification Criteria:

- ONLY classify as NSFW if explicit textual indicators are present.
- Explicit indicators include: "nsfw", "adult content", "nude", "18+", "OnlyFans", "Fanvue", "sexy", "erotic", references to pornography.
- Do NOT infer NSFW from artistic nude photography, medical imagery, generic "AI influencer" services, or portfolio images alone.
- Dual-use gigs ("NSFW or SFW your choice") are classified as NSFW.
- Private message redirects for "adult versions" count as explicit indicators.

Input Data Fields:

- Title
- Description
- Basic, Standard, Premium package descriptions
- Seller biography
- Pricing: Basic, Standard, Premium
- Rating and review count

Required Output (Nested JSON Schema):

```
{
  "service_analysis": {
    "primary_service":
        "ai_generated_video_content" |
                "ai_influencer_creation" |
                "image_generation" |
                "deepfake_face_swap" |
                "content_optimization" |
                "other",
    "explicit_nsfw": true | false,
    "sfw_nsfw_both": "nsfw_only" | "both" |
        "sfw_only",
    "technologies_mentioned": [
      "Stable Diffusion", "Runway ML",
      "DALL-E", "Midjourney", ...
    ],
    "target_platforms": [
      "OnlyFans", "Fanvue", "Instagram",
      "TikTok", "Twitter", ...
    ],
    "requires_user_photos": true | false |
        null,
    "custom_model_training_offered": true |
        false | null
  },
  "ethical_concerns": {
    "deepfake_risk": true | false,
    "platform_policy_violation": true | false,
    "concerns_notes": "Brief explanation of
        risks"
  }
}
```

Note: Seller, pricing, and review metadata are extracted via HTML parsing and not inferred by the LLM.

Important Guidelines:

- Prioritize precision over recall: when uncertain, classify `explicit_nsfw` as `false`.
- Set `deepfake_risk` to `true` when face swap, reference photos, or identity manipulation are offered.
- Set `platform_policy_violation` to `true` if the gig likely violates Instagram/TikTok policies.
- Use `null` for `requires_user_photos` and `custom_model_training_offered` when unspecified.
- Classify `sfw_nsfw_both = "both"` only when explicitly stated.
- Extract all explicitly named tools into `technologies_mentioned`.

B. JSON Output Schema

The classifier returns a nested JSON object containing service characteristics, NSFW labels, seller metadata, pricing, and risk indicators. Table VI summarizes the top-level structure.

The full nested JSON schema used in prompting is shown below.

```
{
  "service_analysis": {
    "primary_service": "...",
    "explicit_nsfw": true | false,
    "sfw_nsfw_both": "nsfw_only" | "both" |
        "sfw_only",
    "technologies_mentioned": [...],
    "target_platforms": [...],
    "requires_user_photos": true | false | null,
    "custom_model_training_offered": true | false |
        null
  },
```

TABLE VI: Summary of JSON Output Schema

Field	Type	Description
`gig_url`	string	Full Fiverr URL of the gig
`service_analysis`	object	NSFW label, service type, mentioned technologies, target platforms
`seller`	object	Username, level, country, membership date
`pricing`	object	Basic/standard/premium package information
`reviews_and_orders`	object	Ratings, review count, orders in queue
`ethical_concerns`	object	Deepfake indicators and policy-violation flags
`metadata`	object	Model identifier, timestamp, run status

```
  "ethical_concerns": {
    "deepfake_risk": true | false,
    "platform_policy_violation": true | false,
    "concerns_notes": "..."
  },
  "seller": {...},
  "pricing": {...},
  "reviews_and_orders": {...},
  "metadata": {...}
}
```

C. LLM vs. Keyword Baseline Comparison

Table VII compares the performance of our LLM classifier against two keyword-based baselines. The LLM achieves substantially higher recall while maintaining perfect precision.

Table VII compares the LLM to explicit and expanded keyword baselines.

TABLE VII: LLM vs. Keyword-Based Moderation

Approach	Precision	Recall
Keyword (explicit terms)	92.3%	34.2%
Keyword (expanded 21-term list)	67.8%	58.1%
LLM (Claude 3.5 Haiku)	100.0%	71.4%

Figure 9 shows a blurred example of how NSFW-oriented AI services are presented on Fiverr.

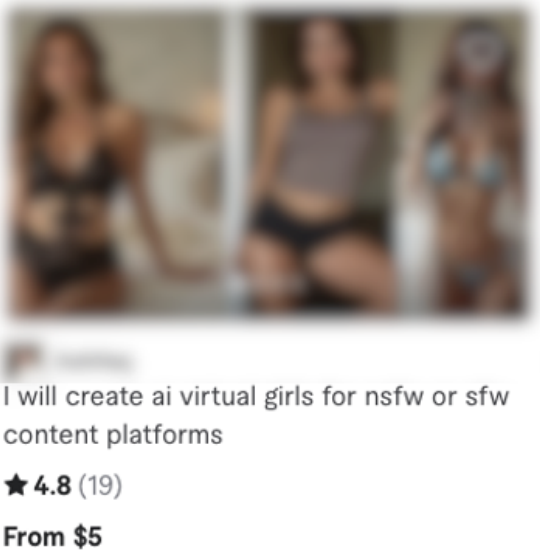

Fig. 9: Blurred example of a Fiverr gig offering AI-generated NSFW virtual models. Image and details are distorted for safety.