BaitBlock: Measuring and Mitigating Phishing Propagation on Kick

Andy Wu
cw4483@nyu.edu
B.S. Computer Engineering

1 Introduction

1.1 Problem Statement

Livestreaming platforms, such as Twitch, Kick, and YouTube Live, have become an increasingly popular vector for novel phishing and impersonation scams. Attackers exploit trust and parasocial relationships within communities surrounding various streamers, then target their viewers with phishing scams.

Livestreaming is a relatively new form of media. While the first public video stream occurred in 1995, Justin.tv, the first social media platform for video streaming, launched in 2007. Justin.tv was then acquired by Amazon in 2011 and rebranded as Twitch.tv. YouTube Live was launched later that year and has since become another giant in the industry. On the other hand, Kick only arrived in 2022, 11 years later than its comparable counterparts. Security for traditional phishing vectors has had twenty-plus years of academic scrutiny, standardized reporting, and maturing detection heuristics. The security surrounding livestreams lacks those advantages. This immaturity makes livestream environments extremely easy to exploit.

Kick stands out as especially vulnerable because it is by far the youngest and least moderated major livestreaming platform. Kick took a “swing fast, swing hard” approach to rapidly establish market share as a new player in the industry. Meaning, their growth strategy focused on creating massive disruptions in the livestream industry. Firstly, they aggressively subsidized creators with massive payouts. Most notably, Kick has publicly announced that they spent $100 million for a non-exclusive contract for popular streamer xQc to use their platform. Additionally, Kick offers a 95% to 5% revenue split with its creators. This massively outsizes other platforms like Twitch, which offers a 50% to 50% revenue split.

Secondly, Kick offered extreme creative freedom by minimally enforcing content policies. Kick launched with almost no guardrails. Content that included illegal activities and explicit content was not taken down from the platform. Most infamously, popular streamer Adin Ross demonstrated this creative freedom live by streaming content directly from pornhub.com and received zero moderation. "I’m gonna go on porn and nothing gonna happen", Ross stated during the livestream. Moreover, street racing streams that features creators recklessly cutting through traffic were a popular category on Kick. Creators such as Squeez Benz also did not face moderation after several hit-and-runs with day-to-day traffic and high-speed chases with law enforcement.

Needless to say, Kick’s unique growth strategy created an environment attractive to "shock value" creators seeking minimal restrictions. I believe the lack of moderation enforcement bred Kick to be a fertile platform for large-scale scam networks. At the very least, Kick’s growth strategy separates itself from other streaming platforms and deserves additional academic attention.

Today, Kick has begun to slowly legitimize itself to attract advertisers. Kick has refactored its platform policies with additional rules, formal definitions, and enforces them more consistently. For example, illicit and illegal content are now flagged and banned. Streamer Adin Ross, mentioned previously, has received a short-term ban after streaming illicit content for the second time. Streamer Squeez Benz has been permanently banned from the platform after his arrest. However, even as Kick attempts to clean up its image and operate as a legitimate competitor to Twitch and YouTube, residue from its early design choices remains. Scam bots still circulate through chats at scale, exploiting Kick’s comparatively weak security ecosystem. Gaming blog DuckyOBrien.com [7] anecdotally reported that many top Kick streams in mid-2023 had titles like “giveawaycrypto2023,” “money-day,” “charity-giveaway,” etc., each spiking suddenly with tens of thousands of viewers. These streams played a pre-recorded video or minimal content that spammed chat with links to “double your crypto” sites.

A recent joint study between UCSD, UCLA, and Google by Liu et al. [5] gathered data on cryptocurrency-based scams on X, YouTube Live, and Twitch. From July 24, 2023, to

January 21, 2024, the study found no identifiable cases on Twitch during the study window despite previous research indicating their abundance. But Liu et al. did find a significant amount on YouTube Live and X. During the research window, attackers netted $1.9 million in revenue from livestreams and $2.7 million from tweets from cryptocurrency-related scams.

Most likely due to how new Kick is, I could not find studies similar to Liu et al. [5] that provide insights into Kick. However, Liu et al. did provide a perspective on the enormity of the issue on other platforms. Notably, YouTube Live viewers suffered $1.9 million in losses during the study period. By contrast, Kick is a far more immature platform characterized by its lack of automated detection, enforcement history, and meaningful guardrails. Although no formal academic research has yet quantified the prevalence of scams on Kick, the structural vulnerabilities identified by Liu et al. in other platforms strongly suggest that Kick's exposure to fraudulent activity is likely far greater than that of Twitch or YouTube Live.

And unlike Twitch or YouTube Live, Kick lacks both the maturity and the institutional experience necessary to anticipate or mitigate these threats. Research on livestreaming safety has begun to catch up to traditional media, with established platforms like Twitch now producing annual transparency reports and openly collaborating with researchers. In contrast, Kick does not publicly offer any platform data regarding trust and safety. As stated previously, I could not find a single academic and empirical source regarding Kick's rate of exposure to scam activity. Kick's opaque moderation practices, limited oversight, and unique creator ecosystem make it difficult to quantify the scale or nature of malicious activity taking place. This is a major platform governance concern in the trust and safety of Kick.

To address this gap, BaitBlock will systematically measure chat data from Kick. BaitBlock aims to accurately quantify and compare the prevalence, types, and severity of scams present on each platform. The project's objective is to provide empirical evidence about Kick's vulnerability to scams, evaluate how moderation practices affect scam exposure, and highlight areas where trust and safety mechanisms may need improvement. BaitBlock hopes to reveal the effects of limited moderation on an emerging streaming platform.

1.2 Motivation and Significance

The motive and intended damage of livestream phishing is similar to traditional phishing scams, but the livestreams as an attack vector are far more scalable and immediate. Liu et al. [5] found that scams on X and YouTube achieved conversion rates of roughly 1 in 1,000 tweets and 4 in 100,000 livestream views, respectively. While the average success rate of these attacks is a minuscule 0.004%, the operation still netted millions of dollars in revenue through the sheer scale of the attacks alone. Further illustrating this reach, an article from *BleepingComputer* by Toulas [4] analyzed data from a cybersecurity firm, Group-IB, showing that cryptocurrency scam links averaged 15,000 visits each, exposing an estimated 30 million potential victims worldwide. The number of such fraudulent websites rose by 300% in 2022, signaling how easily scammers can now mass-produce fraudulent content. Toulas attributed the remarkably accessible scams to the wide reach of livestream platforms and deepfake video generation tools.

Additionally, a recent USENIX study by Nguyen et al. [6] tracked 181 impersonation attacks over three months, with 123 additional cases identified retrospectively across a year. They defined a new category of scam: PROSPER (Payment Re-routing on Social Media via Personal Impersonation). This type of attack impersonates everyday individuals instead of high-profile streamers and injects itself into a real-time conversation. The fake account would then attempt to persuade the payment sender to reroute their transfer to the attackers. Nguyen et al. [6] attributed the prevalence of PROSPER attacks to the fast-paced nature of online conversations and the lack of anti-fraud user interface cues. This suggests that the attack is potentially UI-enabled as well as a classification exercise. Most concerningly, Kick, YouTube Live, and Twitch do not support anti-fraud UI cues as a feature. This is a change that BaitBlock can bring.

Younger audiences are particularly vulnerable. Streaming demographics skew young, and many underage viewers are eager to interact with their favorite creators, often without the skepticism or technical literacy to recognize sophisticated impersonation attempts. A *The Independent* article by Cuthbertston analyzed a multi-year study by *RiskIQ* reported by [2] revealed that scammers impersonating major YouTubers like James Charles and Philip DeFranco successfully deceived more than 70,000 users with prize-related phishing links.

Nguyen et al. [6] also noted that impersonation and phishing through live chats remain critically understudied despite their rapid growth. Kick is the worst offender of this issue. As stated in section 1.1, while other platforms openly work with researchers and publish data regarding trust and safety, Kick provides zero public transparency. This research gap highlights a pressing need for empirical, platform-specific analysis. It also serves as the primary motivation for BaitBlock. BaitBlock seeks to fill in the gap by generating measurable data that can inform both future academic attention and practical moderation policy.

That said, it is an incredibly difficult challenge to detect and prevent this new form of phishing attack entirely. Firstly, Streamers do frequently organize real cryptocurrency-themed promotional events. They reach out to fans and winners in the same communication channels (e.g, chatroom messages, direct messages) where impersonators attempt scams. Occasionally, the official events sponsored by prominent streamers are scams themselves. Numerous high-profile NFT and meme coin projects promoted by very popular streamers have ended in pump-and-dump collapses. A joint research paper from

Huazhong University and Pecking University in China by Huang et al. identified 7,487 rug pulls out of 173,373 NFT projects, or 4.32%. Huang et al. detailed that this percentage is on the lower-bound, and the study uses a very strict detection for rug pulls, and the true number is most likely higher. This further blurs the boundaries between legitimate and malicious engagements on streaming platforms. It will thus be difficult to build effective filter solutions with only rule-based logic.

1.3 Scope and Project

This project focuses on the measurement and comparative analysis of scam activity within Kick livestream chat environments, with an emphasis on cryptocurrency-related phishing and impersonation attempts. The core unit of analysis is individual chat messages collected from public livestreams. There is an enormous amount of messages on Kick, and it will not be realistic to observe all the traffic during the length of this project. Thus, I will only collect messages from three categories notorious for inappropriate activity: *Crypto & Trading*, *Casino & Slots*, and *Just Chatting*. The collected messages will be evaluated using a detection pipeline that combines transformer-based NLP (natural language processing) machine-learning classification with custom rule-based heuristics. This project aims to examine the severity of visible scam probation in these three categories. To isolate effects from stream characteristics such as audience scale, channel size, and stream time, the analysis controls for these factors where possible, enabling comparisons between structurally distinct livestreaming environments.

Twitch is used as a comparative control platform due to its relative maturity in trust and safety practices, including long-standing moderation policies, established enforcement mechanisms, and publicly available transparency reporting. As one of the earliest and most widely studied livestreaming platforms, Twitch provides a meaningful baseline against which to contextualize findings from Kick, particularly when evaluating how differences in platform governance and moderation infrastructure influence scam prevalence. However, the scope of Twitch data collection is constrained by platform access limitations. Unlike Kick, Twitch utilizes the combination of IRC and Websockets to efficiently update chat messages. This makes the chatroom inaccessible via direct DOM inspection and instead requires authenticated access through official APIs and developer credentials. Unfortunately, an approval for a Twitch developer account was not obtained within the project timeframe. As a result, direct measurement on Twitch is limited, and the analysis relies on publicly available transparency reports to contextualize and benchmark observed patterns.

Additionally, the scope of this work remains focused on detection and measurement, with limited end-user intervention in the form of passive visual UI cues. BaitBlock will not remove high-risk messages entirely from the UI. BaitBlock cannot perform automated enforcement actions, content removal, or account-level penalties. The authority for moderation penalties is exclusive to Kick's moderation strategies. BaitBlock does not share this authority.

Finally, this project will not capture and analyze downstream behavioral effects such as user compliance, scam conversion rates, or changes in user trust resulting from these interface cues. Moreover, it is unrealistic for this project to measure scam activity outside of publicly available livestream chats. Including but not limited to: direct messages, external platforms, and off-platform transactions. This project is also confined to phishing and impersonation scams rather than the full spectrum of online fraud.

1.4 Project Outcomes

This project delivers a fully functional Chrome extension that operates as both a measurement instrument and a user-facing trust and safety tool for Kick. The extension performs real-time collection and classification of public chat messages, assigning graded risk labels to messages identified as high-risk scams. As mentioned in section 1.2 by Nguyen et al. [6], the prevalence of PROSPER attacks can at least be partially attributed to a lack of protective UI cues. Thus, leveraging UI injections enabled by Chrome extensions, BaitBlock will visually highlight high-risk messages to protect users.

Additionally, BaitBlock's backend will log structured metadata of every message into a PostgreSQL database with Supabase. This enables systematic analysis of scam prevalence across a range of factors as mentioned in section 1.3, including audience size, channel scale, and content category. The resulting dataset supports a comparative evaluation of scam exposure on Kick, allowing this project to quantify how scam frequency, message risk distribution, and visibility vary across different stream contexts. Collectively, these outcomes provide both an empirical characterization of scam activity on an emerging livestreaming platform and a reusable methodological framework for auditing trust and safety conditions in real-time, user-generated communication systems.

2 Related Work

2.1 Core Sources

My approach is grounded in prior work that characterizes how attackers weaponize trust in online communities. And in response, how platforms counter with combined technical and organizational controls. Liu et al. [5] provide an end-to-end analysis of cryptocurrency "giveaway" scams across major social platforms, emphasizing that conversion rates as low as 0.004% can still yield substantial aggregate harm when scaled across high-volume distribution channels. Thus, we can conclude that attacks in the social media vector are driven by volume. Thus, the primary empirical task for BaitBlock's

mission is to quantify scam exposure rates and identify where the volume concentrates within our scope.

Complementing this, Nguyen et al. [6] document PROSPER attacks, a class of real-time payment re-routing scams that rely on interpersonal impersonation and the speed of live conversation rather than overt malware delivery. Although PROSPER is studied in the context of social media payments rather than livestream chat specifically, its operational logic generalizes to livestream environments. Attackers exploit attention constraints, time pressure, and weak interface cues to induce errors. This aligns with my decision to inject UI signals directly in the chat interface as passive, user-facing cues rather than treating detection as a purely backend classification problem.

I designed BaitBlock's detection pipeline in line with the moderation framework described by Gorwa et al. [3], who characterize content moderation as a layered system combining automated classification with heuristic filtering to operate under constraints such as low latency, high message volume, and ambiguous policy boundaries. In BaitBlock, the automated classification layer is instantiated using a transformer-based NLP model from Hugging Face: ealvaradob/bert-finetuned-phishing [1]. According to the model's documentation, it achieves approximately 98% accuracy and a weighted F1-score above 0.97 on benchmark phishing datasets, with particularly strong performance on short-form text containing URLs, call-to-action language, and impersonation cues.

Consistent with Gorwa et al.'s emphasis on multi-layered moderation systems, the model is not treated as a complete solution. Instead, I applied rule-based heuristics on top of model predictions to account for platform-specific patterns that are underrepresented in traditional phishing corpora. These heuristics accounted for most of the shortcomings of this model for my use case, which I will expand on in the section below. This hybrid approach allows BaitBlock's detection pipeline to preserve the classifier's generalization strengths while mitigating predictable failures in the fast-paced and adversarial livestream chat environment.

Because Twitch is not the primary measurement target in this study, our use of Twitch is anchored in transparency reporting rather than full-scale message collection. In particular, Twitch's Australian Code of Practice submission [8] provides a detailed description of its enforcement pipeline, explicitly framing moderation as a combination of automated detection, proactive human review, and user reporting. The report also discloses data on operational responsiveness. In H2 2024, 93% of user reports were responded to in under one hour, and 96% in under 24 hours. Additionally, the report discloses Twitch's large-scale enforcement against spam, scams, and fraud. In the report, Twitch issued 34.1 million account enforcements globally in H2 2024, including 66,414 for accounts based in Australia. These disclosures provide an institutionally mature benchmark for how a major platform operationalizes scam mitigation, and they serve as a contextual baseline for interpreting Kick-side measurements in the absence of analogous reporting by Kick.

2.2 Identify Gaps

Existing work has established that phishing scams can generate substantial aggregate harm by scaling through social media, even when individual conversion rates remain extremely low [5]. However, much of this measurement literature concentrates on platforms that provide either stable access pathways for researchers or mature reporting practices that enable independent verification of enforcement outcomes. In contrast, Kick does not offer any empirical trust and safety research. Additionally, BaitBlock's scope is relatively small compared to formal platform-backed research like Liu et al. [5]. Thus, Liu et al. [5] cannot provide insights directly into what I should expect to gather from Kick.

Additionally, prior work such as Nguyen et al. [6] highlights the role of UI cues in enabling impersonation and payment redirection attacks, but the literature stops short of operationalizing this insight into deployable features. While [6] hypothesizes that UI warning cues may help to prevent user harm, Nguyen et al. do not offer empirical data on the efficacy of UI-layer prevention methods. Thus, this leaves a gap in the implementation of BaitBlock's UI warning cue feature. Although the feature does address Kick's lack of warning cues, as mentioned in section 1.3, I cannot measure the downstream impact of this feature nor calculate it through existing empirical work.

Moreover, while Gorwa et al. [3] accurately describe real-world moderation systems as layered pipelines, a gap emerges when I attempt to build this pipeline to suit a livestream environment. All publicly available phishing detection models, including the one used in this project [1], are trained on traditional vectors such as emails, SMS messages, or static social media posts. These training distributions differ substantially from livestream chats, where messages are extremely short, repetitive, and often do not contain any real substance.

As a result, directly deploying a pretrained phishing classifier in a livestream setting is insufficient on its own. While NLP models are effective at identifying the intent of texts, they fail in a livestream environment when they are reading short, spam-like messages that do not contain any real intent. BaitBlock addresses this gap by operationalizing Gorwa et al.'s layered moderation framework in a livestream-specific context. I applied rule-based heuristics to clean the data that goes into the model to compensate for mismatches between the model's training distribution and my use conditions.

3 Methodology

3.1 Research Questions and Objectives

This project is guided by two primary research questions.

Firstly, where and how has Kick failed to provide adequate trust and safety protections for its users? This question focuses on identifying observable gaps in Kick's moderation ecosystem by measuring scam exposure in public livestream chats and examining how that exposure varies across stream categories and audience sizes.

Secondly, how can BaitBlock protect users in contexts where Kick's trust and safety mechanisms fall short? This question evaluates whether an external, user-level system can meaningfully mitigate risk by detecting and surfacing scam activity in real time, without relying on platform enforcement or privileged access.

These are the two main objectives that guide this project. Together, they frame BaitBlock as both an audit of Kick's existing trust and safety conditions and an exploration of what protections are possible when platform-level safeguards are weak, opaque, or absent.

3.2 Methodological Approach

To answer the first of the two main questions framed in section 3.1, BaitBlock adopts a quantitative, measurement-driven approach based on primary data collection from public Kick livestream chats. The data pipeline starts with a Chrome extension built with Vite, React, and TypeScript that collects chat messages in real-time. This approach was chosen in favor of using other scraping methodologies to ensure a seamless user experience while preserving BaitBlocks' empirical and research capabilities. Alternatively, using Python libraries such as BeautifulSoup would enable a simpler data pipeline that is also potentially more capable. However, BaitBlock would then be unable to be integrated with the browser as an extension. And consequently, UI warning for high-risk messages would also become extremely difficult to implement. This was not the vision I had for BaitBlock.

The Chrome extension detects incoming messages through TypeScript's MutationObserver class, which can detect DOM changes in the document body. I then identified the HTML structure of chat messages to isolate them from other DOM changes and parsed the required information. If a DOM change led to the disappearance of a single message but not a batch of messages, it would be considered taken down by moderators. A background service worker then enters the information into my database. The service worker also batches the messages and passes the payload into a local server every second. I decided to batch messages because both the local server and the classification model it hosts can achieve higher throughput when it's given an array of objects. The time frame of 1 second was arbitrarily chosen. It was kept because it resulted in an acceptable level of latency.

Although I chose not to use Python to scrape the chatroom, I chose to build BaitBlock's backend server primarily in Python. This decision was driven by two main factors. Firstly, I wished to leverage Python's mature ecosystem for machine learning inference. And secondly, I had prior experience working with Python-based data pipelines. Using Python allowed the detection pipeline to integrate directly with a pretrained transformer model without additional serialization layers or cross-language overhead, reducing both implementation complexity and inference latency.

The server itself was implemented with Python's FastAPI, which provides an asynchronous, lightweight framework well-suited for handling high-frequency, low-payload requests, such as batched chat messages. FastAPI's request validation and automatic schema enforcement ensure that message batches sent from the Chrome extension are well-formed, while its async execution model allows the server to process incoming data continuously without blocking on model inference. This design enables the backend to act as a real-time classification service rather than a traditional request–response API.

Within the extension's background service worker, incoming chat messages are first normalized and filtered to remove malformed or low-information inputs before being passed to the NLP phishing classifier [1]. Model outputs are then combined with rule-based heuristics to produce final risk labels. This includes a custom confidence score of 0.999997 to remove false positives. This confidence score is tuned by raising the value by 0.000001 from 0.99999 every 30 minutes until the accuracy of its classification is consistently higher than 98% when compared to the same data labeled by ChatGPT 5.2. While this method is not a formal evaluation procedure such as k-fold cross-validation, it provides a practical, deployment-oriented mechanism for calibrating confidence thresholds under real-world operating conditions.

Finally, the classifier's results are returned to the extension for UI rendering and simultaneously logged to persistent storage for offline analysis. By consolidating classification, heuristic filtering, and data logging into a single FastAPI service, BaitBlock maintains a coherent and auditable detection pipeline that supports both real-time user protection and longitudinal measurement of scam activity.

All classified messages and metadata are logged to another table in Supabase for offline aggregation and analysis of the results. I attempted to use Twitch as a control group for this project's data, but because Twitch chat delivery relies on IRC and WebSockets and requires authenticated API access that was not obtained, I could not collect data comparable with Kicks. Instead, Twitch is used as a comparative reference point, but not as a symmetric data source. Twitch's transparency reporting is used to contextualize Kick's observed scam exposure against a platform with established moderation practices and published enforcement metrics.

4 Findings

4.1 Reconnection to Research Questions and Objectives

The findings presented in this section directly address the project's two guiding research questions. One: Where and how does Kick fail to provide trust and safety protections for its users? And two: how does BaitBlock mitigate these failures at the user level?

Analysis of the collected chat data reveals two distinct and extremely prevalent scam attack vectors operating on Kick. In the first vector, obvious bot accounts would target small streamers through chat messages, offering promotional or networking opportunities. The messages are always in one of two different templates: "Would you like many viewers? Let's make you popular - ...", and the second template: "Hey homie, I checked your videos seems like you are new in kick not yet affiliated and verified. I got a proposal for you by sharing your stream to a kick community for you to gain 2000 active audience and earn nothing less than $2000 while me on discord if you're interested..."

Both templates follow traditional phishing characteristics, and the message content itself does not constitute a novel form of attack. In both templates, attackers explicitly mention the streamer by username and attempt to move the interaction off-platform, a common precursor to downstream phishing scams. However, the use of either hyperlinks or Discord usernames, without overlap between templates, suggests the presence of at least two distinct attacker groups operating parallel campaigns with similar objectives.

For clarity, I will be referring to this attack vector as promoter bait. Promoter bait is fully observable in chat data and is quantified through both aggregate bucket-level analysis and per-channel trend analysis. Extremely concerningly, 100% of channels with less than 50 subscribers have received at least one promoter bait message. And around 64.5% of channels with 51-150 subscribers have received a promoter bait message. Of the 868 channels that received a Pomoter Bait, only 12 were taken down from the platform. Upon manual inspection, all 12 of the take-downs were from human moderators isolated per stream, and none were taken down by the platform itself.

The second attack vector targets viewers rather than streamers through continuously running, bot-operated streams in the Crypto & Trading category. These streams broadcast static or minimally changing content, such as live cryptocurrency price charts, while aggressively promoting fraudulent cryptocurrency giveaways, most commonly framed around Solana. Viewers are repeatedly prompted through on-screen messages and chat to redeem "free crypto" codes via external links. For clarity, I will be referring to this vector as bait crypto streams.

As noted in Section 1.3, this project does not have downstream visibility into viewer behavior, link clicks, or financial outcomes, which unfortunately prevents quantitative measurement of harm for this vector. However, qualitative observation during the data collection period showed that a small set of clearly bot-operated streams using near-identical naming conventions such as SOLANA247, solpengu247, and SOLCODES247, consistently ranked among the top streams in the category. During the observation window, these bot streams outperformed legitimate creators in the category 100% of the time. This persistence highlights a distinct failure of platform-level enforcement and discovery controls, separate from chat-based scam mitigation.

To address the first research question: Where and how Kick fails to provide trust and safety protections for its users? The findings point to two clear and observable failures. First, Kick leaves small streamers largely unprotected against direct scam targeting in chat, allowing promoter bait messages to dominate early-stage channels without warning or intervention. Second, Kick permits bot-operated cryptocurrency scam streams to persist within its discovery and ranking systems, effectively displacing legitimate creators and exposing viewers to high-risk content. Although the attack is obvious and individual scam conversion rates are most likely very low, prior work by Liu et al. [5] demonstrates that such attacks can produce substantial harm when scaled. Given Kick's user demographics and platform structure, these failures likely result in meaningful user exposure and loss.

Finally, to address the second research question: How does BaitBlock mitigate these failures at the user level? The findings demonstrate that meaningful protection is possible even in the absence of platform enforcement. Prior work by Nguyen et al. [6] shows that interface-level cues can reduce the effectiveness of real-time impersonation and payment redirection attacks by disrupting attacker momentum and increasing user awareness. BaitBlock applies this same principle to livestream chat environments. By detecting and visually highlighting high-risk messages in real time, the system introduces protective signals precisely where Kick provides none. Although BaitBlock cannot remove scam content or impose penalties, the explicit surfacing of promoter bait addresses the platform's failure to warn users—particularly small streamers—at the moment of interaction, reducing the likelihood that scams are mistaken for legitimate outreach.

4.2 Data Overview and Scope of Analysis

I chose to host my database with Supabase to leverage its built-in APIs and libraries for both Typescript and Python, as well as its PostgreSQL schema. The schema I required for each row was: id, streamer_name, username, text, emote_id, and timestamp. Supabase automatically generates a unique ID for each entry and the current time as a timestamp. The rest of the columns are filled in by the request body from the API request.

As previously mentioned in section 1.3, due to the enor-

mous quantity of messages on Kick, this project will only monitor 3 notorious categories: *Crypto & Trading*, *Casino & Slots*, and *Just Chatting*. These categories were selected due to their consistent association with spam, scams, and low moderation presence. I created a separate table for each, enabling comparisons across categories and analysis for each unique category. Messages were further grouped into subscriber-count buckets to examine how scam exposure varies as channels scale.

Five subscriber buckets were defined: 0–50, 51–150, 151–500, 501–1500, and 1501–5000 subscribers. For each bucket, messages were collected across many distinct channels rather than concentrating on a small number of high-traffic streams. This design choice prioritizes breadth over depth and allows scam prevalence to be measured as a function of channel size rather than individual streamer behavior. I observed Kick daily at arbitrary times for 50 days.

In total, the dataset includes:

- 0 - 50 subs: 5002 messages from 613 channels
- 51 - 150 subs: 5129 messages from 395 channels
- 151 - 500 subs: 8041 messages from 112 channels
- 501 - 1500 subs: 20145 messages from 124 channels
- 1501 - 5000 subs: 21006 messages from 4 channels

Additional data was briefly collected from two extremely large channels: xQc and Togi, totaling 7,140 messages. However, this collection was discontinued early due to disproportionately high message volume quickly filling up my Supabase polling limits. Additionally, a preliminary observation was that scam prevalence beyond 500 subscribers was already statistically negligible. These channels were therefore excluded from further analysis.

Data collection within each subscriber bucket continued until the incoming 200 messages no longer altered the observed scam prevalence by 1%. Rather than relying on a fixed sample size, collection was stopped once new data points ceased to meaningfully skew aggregate results. While this convergence was assessed heuristically rather than through formal techniques like the power analysis, I believe it is enough for the scope of this project due to the consistency of results across hundreds of channels.

4.3 Key Themes, Patterns, and Results

The promoter bait attack vector demonstrates that scam targeting on Kick is systematic, strategic, and nearly universal below a modest channel-size threshold. In channels with fewer than 50 subscribers, promoter bait messages account for over 96% of all observed chat traffic, and every monitored small streamer received at least one such message. These scams remain visible in chat without warnings, filtering, or removal, indicating an absence of effective automated detection or proactive moderation in early-stage channels. The sharp decline in scam prevalence beyond approximately 150 subscribers suggests that protection emerges only once informal safeguards such as active moderation or streamer vigilance become viable, rather than through platform-provided mechanisms. By contrast, Twitch's 2025 Transparency Report [8] documents 34.1 million account enforcements related to spam, scams, and fraud. While these figures are not directly comparable to chat-level prevalence measurements, they provide a useful reference point for platform enforcement capacity. During the data collection period for this project, I observed no instances of promoter bait messages being removed or moderated on Kick, highlighting a stark disparity in proactive trust and safety enforcement. I believe it is safe to conclude that Kick provides little to no trust and safety coverage at the stage where creators are most vulnerable.

Second, Kick fails to protect viewers from scam exposure at the platform level. The crypto stream bait vector shows that bot-operated scam streams can run continuously in the Crypto & Trading category, promote fraudulent giveaways, and consistently outperform legitimate creators in platform rankings. These streams exhibit clear indicators of inauthentic behavior—24/7 uptime, repeated naming conventions, static content, and aggressive off-platform promotion—yet remain discoverable and highly visible. This persistence points to a breakdown not just in moderation, but in discovery and enforcement controls, where scam content is amplified rather than suppressed. Although downstream harm could not be measured within this project's scope, the platform's failure to intervene at the discovery level exposes large audiences to high-risk content by default.

Taken together, these two vectors reveal a broader structural failure. Kick's trust and safety protections are inconsistent at best. Small streamers are left exposed, while large streams benefit from informal defenses, and scam streams can exploit ranking systems in the absence of meaningful enforcement. The absence of user-facing warning cues, transparent enforcement data, or measurable mitigation mechanisms further compounds these failures.

4.4 Illustrative Evidence

Figures 1 and 2 below provide complementary visual evidence for the promoter bait attack vector identified in this study. Figure 1 aggregates scam prevalence by subscriber bucket defined in section 4.2, showing that promoter bait messages dominate chat traffic in the smallest channels and rapidly diminish as channel size increases. In channels with fewer than 50 subscribers, promoter bait accounts for over 96% of all observed messages, confirming that early-stage streamers experience near-zero organic chat traffic in their streams. Although prevalence decreases in the 51–150 subscriber range, promoter bait still constitutes the majority of messages, indicating that modest growth alone does not mean-

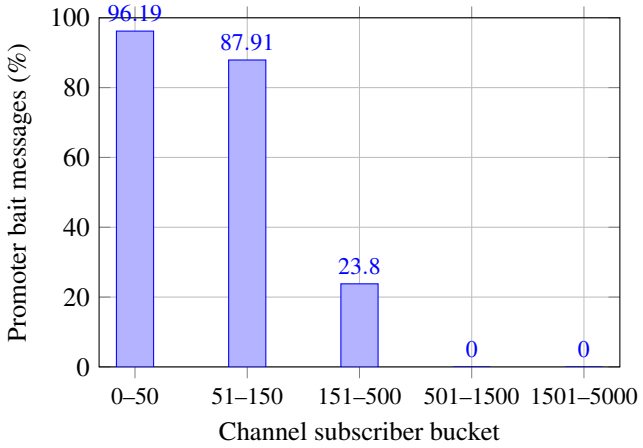

Figure 1: Promoter bait prevalence by bucket.

ingfully reduce exposure.

Figure 2 presents a continuous line graph view of the same phenomenon. Each point represents the average of ten channels, revealing a sharp inflection point around 150–160 subscribers. Below this subscriber count threshold, scam prevalence remains consistently high. Above 150-160 subscribers, prevalence collapses by several orders of magnitude almost immediately. This pattern reinforces the conclusion that promoter bait attacks are strategic rather than incidental. Attackers concentrate effort where enforcement is weakest and disengage once informal defenses such as active moderators or streamer awareness become the default.

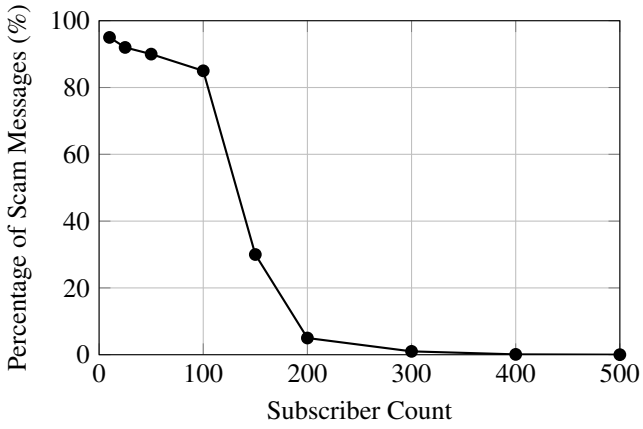

Figure 2: Promoter bait prevalence by subscriber count, one point equals 10 accounts

Together, these figures visually demonstrate that Kick's failure to protect users is not evenly distributed across the platform. Instead, it is concentrated at the lowest levels of visibility and power, where creators lack the tools, experience, or platform support needed to defend themselves. The harm is localized, systematic, and sharply bounded by channel scale.

5 Discussion

5.1 Interpretation and Contribution

This project makes three primary contributions to the study of trust and safety in livestreaming environments.

First, it provides the first empirical anchor for scam exposure on Kick, using primary chat data in the absence of transparency reporting or prior academic cooperation from Kick itself. This work establishes that scam activity on Kick is not sporadic or user-driven, but instead highly structured and strategically deployed. The promoter bait attack vector demonstrates that scam targeting is nearly universal below a modest subscriber threshold and collapses rapidly above it. This indicates that Kick does not provide platform-wide protections for creators, and early-stage creators are extremely vulnerable without their own moderation systems.

Second, the findings expose failures that extend beyond chat moderation and into stream moderation. The persistence and dominance of bot-operated crypto scam streams within the Crypto & Trading category shows that Kick not only fails to remove scams, but in some cases algorithmically amplifies them. These streams exhibit clear indicators of inauthentic behavior, such as continuous uptime, static content, repeated naming conventions, and aggressive off-platform promotion. Yet, they consistently outrank legitimate creators and are often recommended by Kick's algorithm for their dominance in the category. This suggests that Kick's ranking and discovery mechanisms lack effective safeguards against abuse, allowing scam content to displace authentic participation at scale.

Third, this work demonstrates that meaningful mitigation is possible even without platform cooperation. By applying a layered detection pipeline and surfacing risk through interface-level cues, BaitBlock operationalizes prior theoretical insights from Nguyen et al. [6] and Gorwa et al. [3] in a real-world livestream setting. While BaitBlock cannot remove content or enforce penalties, it directly addresses Kick's lack of UI level protections and warnings. For small streamers in particular, the ability to distinguish legitimate outreach from scam attempts in real time represents a concrete improvement in user safety.

Taken together, these contributions show that Kick's trust and safety failures are structural rather than incidental. Protection emerges only through informal mechanisms once creators reach sufficient scale, leaving smaller users systematically exposed. BaitBlock demonstrates that this gap is not inevitable, but the result of design and enforcement choices.

5.2 Limitations

This study has several important limitations.

First, it does not measure downstream outcomes such as scam conversion rates, financial loss, or behavioral change in response to warnings. As a result, claims about harm re-

duction are limited to exposure and visibility rather than confirmed prevention. The impact of both this project's contributions and Kick's failures cannot be properly quantified. While prior work suggests that UI cues can meaningfully disrupt scams [6], this project cannot empirically validate that effect within its scope.

Second, data collection is restricted to three high-risk categories. Scam activity occurring through direct messages, external platforms, or less-monitored categories is not captured. Additionally, while the dataset spans hundreds of channels, it represents a snapshot in time rather than a longitudinal view of platform evolution.

Third, cross-platform comparison is constrained. Twitch is compared through transparency reporting rather than direct measurement, which limits conclusions to differences in enforcement posture rather than quantifiable scam prevalence. While this comparison is informative, it cannot establish causal relationships between moderation practices and observed outcomes.

Finally, some methodological choices, such as heuristics and confidence threshold tuning, were driven by deployment constraints and personal estimations rather than formal statistical guarantees. These choices were appropriate for an applied, system-oriented project, but they limit the precision and academic integrity of this project.

5.3 Future Work and Unresolved Questions

A natural next step is to evaluate behavioral impact. Controlled studies could measure whether UI-level warnings meaningfully reduce engagement with scam content, alter streamer responses, or decrease successful off-platform redirection. Such work would move beyond exposure measurement toward direct harm mitigation. It would be nice to see my work spark real academic attention in Kick as a platform.

Expanded data collection would also strengthen the analysis. There are hundreds of categories and millions of chat messages on Kick every moment. Monitoring additional categories, extending observation windows, and incorporating temporal trends could reveal how scam strategies adapt over time or respond to platform policy changes. A deeper analysis of the attackers' coordination across channels could further clarify the organizational structure of scam networks.

Real and meaningful progress in this space depends on platform accountability and awareness. And regulatory requirements for transparency reporting or structured researcher access are needed to enable more rigorous research into Kick. Unfortunately, I lack the resources, credentials, and skillset to present a truly academically rigorous research paper for this new and rapidly growing livestreaming platform. However, I do believe that BaitBlock does offer a clear and engaging, albeit incomplete and estimated insight into the state of phishing propagation on Kick.

I am personally interested in continuing this project and training a livestream-specific detection model rather than relying on pretrained phishing classifiers built for email, SMS, or static social media posts. The broader livestream environment includes spam-like text that often lacks clear semantic intent. A model trained directly on Kick chat distributions could better capture platform-specific linguistic cues, attacker templates, and adversarial phrasing patterns, reducing both false positives on low-information chat and false negatives on evolving scam variants. This would require building a labeled dataset from BaitBlock's collected messages and evaluating performance using held-out streams and time-separated splits to test generalization against new channels and attacker drift.

References

[1] E. Alvarado. Bert fine-tuned for phishing detection. https://huggingface.co/ealvaradob/bert-finetuned-phishing, 2023. Hugging Face model repository.

[2] A. Cuthbertson. Youtube impersonation scam has tricked 70,000 people, study reveals. *The Independent*, Jan. 2019. Accessed: 2025-11-02.

[3] R. Gorwa, R. Binns, and C. Katzenbach. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1):1–15, 2020. Creative Commons Attribution–NonCommercial License.

[4] J. Huang, N. He, K. Ma, J. Xiao, and H. Wang. A deep dive into nft rug pulls. *arXiv preprint*, arXiv:2305.06108, 2023. Accessed: 2025-11-11.

[5] E. Liu, G. Kappos, E. Mugnier, L. Invernizzi, S. Savage, D. Tao, K. Thomas, G. M. Voelker, and S. Meiklejohn. Give and take: An end-to-end investigation of giveaway scam conversion rates. *arXiv preprint*, arXiv:2405.09757v1, 2024. Accessed: 2025-11-11.

[6] H. D. Nguyen, S. Dhungana, M. Itha, and P. Vadrevu. "please don't send that bot anything": A mixed-methods study of personal impersonation attacks targeting digital payments on social media. In *Proceedings of the 34th USENIX Security Symposium (USENIX Security '25)*, pages 4859–4878, 2025. Open access.

[7] D. O'Brien. Does kick have a bot problem? Blog post, June 2023. Accessed: 2025-11-11.

[8] I. Twitch Interactive. Twitch transparency report: Australian code of practice on disinformation and misinformation — may 2025. Technical report, Twitch Interactive, Inc., May 2025. Report submitted under the Australian Code of Practice on Disinformation and Misinformation.