Contextual Classification of Cybercriminal Posts Using Large Language Models: A Comprehensive Study on Tech Support Scam Marketplaces

Raghavendra Cherupalli
School of Cyber Studies
The University of Tulsa
Tulsa, USA
rac8609@utulsa.edu

Hawken Grubbs
School of Cyber Studies
The University of Tulsa
Tulsa, USA
hlg9644@utulsa.edu

Yi Ting Chua
School of Cyber Studies
The University of Tulsa
Tulsa, USA
ytc2805@utulsa.edu

Weiping Pei
School of Cyber Studies
The University of Tulsa
Tulsa, USA
weiping-pei@utulsa.edu

Tyler Moore
School of Cyber Studies
The University of Tulsa
Tulsa, USA
tyler-moore@utulsa.edu

Gary Warner
Department of Criminal Justice
Univ. of Alabama at Birmingham
Birmingham, USA
gar@uab.edu

***Abstract*—In a tech support scam (TSS), cybercriminals impersonate legitimate service providers by mimicking the interactions consumers routinely have with companies. We conduct a comprehensive analysis of the supply side of the TSS ecosystem on Facebook, where groups operate as informal marketplaces that lack traditional trust or reputation metrics. The study utilizes an AI-driven technique to classify posts into different categories, based on labels derived from manual classification, using Gemma original and Gemma-3-12B large language models. In total, we categorized 381,843 posts across 96 groups made between April 2015 and March 2024. The results highlight different user types and their characteristics. We analyze the resulting posts to shed light on the various types of products and services offered by the groups. We also investigate the extent of specialization and generalization among cybercriminal participants. It is hoped that the detailed study on such ecosystems can aid law enforcement and policy efforts to identify suitable intervention points and effective countermeasures against the TSS ecosystem.**

***Index Terms*—tech support scam, cybercrime measurement, large language models, in-context learning**

I. Introduction

Technical support scams (TSSs), also known as call center or impersonation scams, first appeared around 2012 [1]. Scammers utilized cold-calling operations and social media messages to contact victims about a virus by posing as trusted brands such as Microsoft or Dell. The scammers duped users into permitting them to take over their systems remotely as part of the ruse.

By 2025, TSS has proliferated, causing widepread societal harm while exhibiting no signs of decline. The FBI Internet Crime Complaint Center (IC3) recorded 36,002 complaints that totaled a loss of USD $1.46 billion [2]. These figures not only reflect the scale of victimization, but also signal the presence of a well-developed ecosystem that enables, coordinates, and monetizes these scams.

979-8-3315-8969-1/25/$31.00 ©2025 IEEE

Recent studies have revealed that the TSS ecosystem involves multiple layers of actors and services. At the center of this ecosystem is the existence of marketplaces. Liu et al. [3] uncovered not only scammers answering calls, but also webmasters who sell victim numbers to call centers, agents who engage with victims, and financial actors who assist in laundering proceeds. This ecosystem relies on public social media channels such as Facebook and WhatsApp, which are used for advertisements, job offerings, and dissemination of scam-related infrastructures. The availability of these public social media channels suggests a mature marketplace structure.

Despite growing public awareness and platform-level intervention, these scam-contributing groups continue to thrive on social media platforms. In response, this study seeks to examine the internal dynamics of TSS marketplaces with a specific focus on Facebook-based scam ecosystems. By examining the services and roles within these groups, the study lays the groundwork for more targeted and data-driven disruption strategies. This study is guided by the following research questions:

- *What are the various types of services/products offered in the groups?* This question aims to identify and categorize the range of services/products advertised within these groups using a mixed-method approach.
- *What service/product categories are dominant within the groups? What do these patterns suggest about the structure and priorities of the TSS ecosystem?* The distribution of services/products offer insight into market demands, actor and marketplace specialization, and potential intervention points within the ecosystem.

A. Key Contributions

This paper makes both empirical and methodological contributions to the study of TSS marketplaces:

- **Ecosystem Insights:** We analyze the prevalence and distribution of service/product categories to uncover dominant market segments. Money laundering services had the highest overall number of posts, with two of the top 10 groups specializing in the category. Contribution of specialized posters in the job offerings category is much higher when compared to the contribution of specialized users within other categories. In addition, we identify specialization at the group level, with 40 out of 96 groups with names that focus on a specific product or service pertaining to TSS.
- **Documentation of Generalist and Specialist Behavior:** We examine both marketplaces and vendor specialization within the ecosystem. The evidence indicates specialization at the marketplace level, even as most vendors themselves are non-specialists.
- **Large-Language Model (LLM) Evaluation for Post Classification:** We investigate the extent to which prompting strategy can replicate human-annotated categorization of scam-related services/products. This includes a comparative evaluation of different LLM architectures in performing multi-label classification tasks on noisy, real-world social media data.

II. Related Work

We now review three types of relevant work that has informed this research: measurement studies of TSS, studies of vendor specialization among cybercriminals, and relevant research utilizing LLMs to label forum posts.

A. Measurement Studies of Tech Support Scams

A number of researchers have documented the evolution of tactics employed in Initially, TSS predominantly utilized outbound phone calls, wherein criminals contacted prospective victims. Eventually, the technique transformed into an inbound model, in which victims are duped into initiating communication with the crooks. This shift was enabled by counterfeit websites and online pages crafted to resemble authentic services, frequently utilizing recognizable logos and user interfaces to bolster credibility. Miramirkhani et al. [4] documented the shift and developed an automated system for discovering the impersonations.

A subsequent study by Srinivasan et al. [5] further classified TSS domains into two types: aggressive and passive. Aggressive strategies evoke a sense of urgency or terror, frequently accompanied by repeating conversation boxes or persistent pop-up windows with frightening audio notifications. Passive strategies, by contrast, employ used branded images, official-looking certifications, trust seals, and other content to appear authentic. These sites often posed as approved support services for well-known technology companies. It was observed in both cases that scammers use black hat search engine optimization (SEO) techniques to alter search engine results, leading victims to find the scam websites rather than those belonging to the actual corporations whose help they are seeking.

The behavior of scammers during real-time interactions in TSS was thoroughly examined qualitatively and exploratory in the work by Rauti et al. [6]. Through the use of participant observation and case study methodologies, the study demonstrated that although scammers' techniques have changed over time, the fundamental tactic has remained the same: posing as a trustworthy service provider or brand, winning the victim's trust, and then requesting payment for a fake good or service. According to the findings by Larson et al. [7], toll-free numbers are essential for enabling contact between fraudsters and their victims. These numbers work as the main entry point, making it easier for victims to get in touch with scammers while also allowing them to look authentic. Notably, this study is also the first to integrate artificial intelligence techniques for the detection of TSS webpages, a significant advancement in automated scam identification methods.

One of the techniques used by call center scammers to advertise themselves is through SMS messages, which instruct victims to either call a given phone number, visit the website link attached or pay based on the content of the message. As explored in the work of Choi et al. [8], smishing assaults usually commence with a "pre-crime phase" to facilitate the scam, such as procuring tools and supplies through illicit web marketplaces.

Liu et al. [3] undertook the first extensive research of the TSS ecosystem, using posts from real-time social media channels, primarily Facebook and WhatsApp. The study examined 13 WhatsApp groups and 10 Facebook groups that actively contribute to TSS. Based on the results, these groups are mostly used to solicit or advertise a variety of illegal services to people involved in TSS activity. This classification serves as a preliminary foundation for our understanding of the structure and functioning of these groups. Liu et al [3] presented their Topic-Agnostic Scam Recognizer (TASR) to differentiate TSS websites from legitimate service providers. Recently, Wood et al. [9] used automated machine learning techniques to analyze recordings of scam calls from YouTube and classified the scams into various categories. These prior studies, however, were mostly devoted to creating a system that could identify TSS websites or distinguish them from legitimate tech support websites. By contrast, the present study seeks to better understand the broader ecosystem facilitating TSS scams, using a trained LLM approach to classifying discussions from a much larger collection of criminal service provider Facebook groups.

B. Vendor Specialization in Underground Marketplaces

The online underground, whether on the dark web or surface web, is known to supply various types of services and products related to online crime and deviant behaviors. Market segmentation often occurs at the platform level. Some forums and marketplaces are dedicated to a narrow range of cybercrime commodities such as stolen data [10]–[13], while others operating on the dark web have historically hosted a

broader array of goods and services [14]–[17]. Within these marketplaces, some vendors adopt a specialized role, focusing on specific product or service to establish credibility and competitive advantage [17]–[19]. Soska et al. [17] found that nearly half of the vendors on multiple dark web markets sold exclusively one type of product. Similarly, Van et al. [19] demonstrated that the professionalism and focus of vendors significantly predicted sales. Thus, the overall ecosystem comprises both niche specialists and broader generalists, with each type being shaped by market norms and product demands.

Despite the presence of niche expertise, specialization is not universal. One critical factor is technical complexity. According to the findings of Holt et al. [20], vendors offering high-skill products like malware or counterfeit identity documents are more likely to be specialized due to the knowledge and tooling required. A detailed analysis by Haslebacher et al. [11] showed that only about 31% of vendors in major online stolen data marketplaces sold a single type of product, while other vendors offer a mix of products (e.g., email lists, PayPal credentials). Even among the most reputable vendors, specialization was not uniform. In one case, only seven of the top 20 vendors deal in just one product category [11]. These findings highlight that vendor specialization is contingent upon market structure, technicality of products or services, and individual business strategies.

C. LLMs and Forum Post Analysis

Large language models (LLMs) such as GPT-4 [21], LLaMA [22], and Gemma [23], [24] have demonstrated remarkable capabilities in a wide range of natural language understanding and generation tasks. These models are typically used in inference settings through in-context learning (ICL), where task-specific instructions and examples are provided as input without requiring parameter updates, as discussed in the research by Hao et al. [25]. Depending on the number of examples included in the prompt, ICL can be applied in zero-shot settings (as shown by Xian et al. [26] and and Pourpanah et al. [27]), few-shot settings (e.g., Brown et al. [28]), or many-shot settings (e.g., Agarwal et al. [29]). Zero-shot prompting relies solely on natural language instructions, while few-shot and many-shot approaches include one or more labeled examples to guide the model's reasoning and improve performance. To enhance interpretability, work by Zelikman et al. [30] have also developed Chain-of-Thought (CoT) prompting, which encourages the model to generate intermediate reasoning in natural language. While CoT has shown promise in tasks involving logical reasoning and classification, challenges remain in ensuring that the generated justifications reflect genuine reasoning rather than post hoc rationalization. These prompting strategies, combined with advances in model architecture and extended context windows, form the foundation for applying LLMs to more complex, domain-specific tasks.

Recent advances in LLMs have created new opportunities for the automated analysis of unstructured, large-scale data from online forums. One line of work has explored the use of LLMs to support qualitative analysis of large-scale forum-style text. For example, Rao et al. [31] proposed QuaLLM, an LLM-based pipeline to extract structured themes and representative quotes from over a million Reddit comments related to gig economy workers. Similarly, a recent study on a health support forum by Muasher-Kerwin et al. [32] evaluated the use of GPT-3.5, GPT-4, and LLaMA models to summarize and interpret posts from a brain tumor community.

Another emerging area involves automated quantitative categorization. A notable example is the study by Giannilias at al. [33], which evaluated the performance of four open-source LLMs on the task of classifying hacker forum posts into predefined categories. The study tested multiple training approaches, including zero-shot prompting, few-shot prompting, and full fine-tuning, using expert human annotations as a baseline. Distinct from prior work, our study applies LLMs to a novel domain: the automated categorization of user-generated posts in tech support scam–motivated Facebook groups that operate as informal marketplaces. Our approach integrates both quantitative analysis (querying LLMs to classify scam-related posts into predefined categories) and qualitative reasoning (eliciting explanations for the model's classification decisions). This hybrid approach enables a richer understanding of scam discourse and supports both empirical insights and interpretability in scam analysis.

III. Data Collection Methodology

A. Data Source

The study is conducted on Facebook groups whose primary objective is to support TSS. While there are many legitimate Facebook groups for call center services and recruiting, these groups were selected based on their demonstrated tolerance for criminal activity, or in some cases, their expressed preference for offering "crime as a service" tools and capabilities. The keywords utilized to search these groups comprise "blue screen of death (BSOD)", "Porn-popup", "email blasting", "toll-free number (TFN)", and "tech support x" (with x representing specific locations such as the USA, UK, Delhi, Lahore and Noida). ***BSOD*** refers to a tactic for eliciting calls by crashing a victim's browser and displaying a message indicating the machine is infected with a virus. ***Porn-popup*** refers to techniques that cause pornographic images to appear on the computer and claiming the victim must either call tech support or law enforcement about the situation. ***Email blasting*** refers to sending emails claiming a credit card charge has been made to pay a tech company for a service, with a phone number provided to cancel the service. ***TFN*** refers to toll-free number. While not criminal in nature, TFN is a common service offering since call center numbers are routinely disabled due to fraudulent activity. Because not all of these groups are implicitly criminal in nature, each group is manually validated using "About" statement from the group's bio, and a recent collection of posts made to the group. Only groups with clear "crime as a service" offerings were included in the dataset.

The dataset for this study consists of 96 groups, which had a total of 186,936 members. The average size of these groups

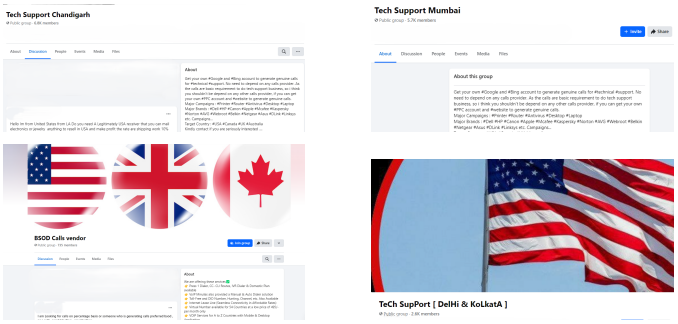

Fig. 1: Snapshots of few Facebook groups that contribute to criminal activities.

was 3,250 members. This is a longitudinal dataset that covers a total of 381,843 posts which were uploaded to these groups starting from the time period of April 2015 to March 2024.

The dataset is categorized into various sub-datasets, each containing a distinct type of metadata:

- The post metadata component includes a unique `post_id`, an `author_id` (corresponding to the `member_id` of the user who uploaded the post), the `group_id` representing the group where the post was published, the post text, and a timestamp indicating when the post was uploaded. The same `author_id` can appear multiple times, since a single member may create multiple posts.
- The member metadata component contains records of individual members, each identified by a unique `member_id`, along with the corresponding `group_id` indicating the group to which they belong. Since an individual can be a member of multiple groups, the same `member_id` may appear multiple times, once for each group affiliation. This structure represents a many-to-many relationship between members and groups.
- The group metadata component provides the `group_id` with each group having a unique identifier and creation date.

B. Data Preprocessing

For this paper, only the post and group subdatasets were considered from the entire dataset. To ensure that only relevant English-language posts were retained for downstream processing, a robust, hybrid language detection strategy was used as part of the data cleaning pipeline. This step was crucial because the cleaned posts were intended to be fed into large language models (LLMs) that operate exclusively on English input. Retaining non-English content would reduce the efficacy and consistency of the LLM's performance.

Three language detection libraries, LangID (Lui et al. [34]), langdetect, and FastText (Joulin et al. [35] and Sainte-Marie et al. [36]) were compared on both original and decoded versions of the text. A hybrid decision rule was applied: if any of the decoded classifiers (Langid, Langdetect) or the original FastText output predicted English ("en"), the final label was set to English; otherwise, the label defaulted to FastText's original prediction. This rule prioritized recall for English while ensuring high precision, and it leveraged FastText's robust performance in cases where no classifier confidently predicted English.

The hybrid model achieved an accuracy of 99.2%, significantly outperforming all individual language detection libraries when evaluated against a manually labeled subset of 133 diverse posts comprising English, null, and various foreign language content. Detailed evaluation metrics and comparative results for all models are provided in the Appendix A.

The resulting dataset consists of 321,319 posts after excluding 60,523 null or non-English posts as a part of data cleaning.

IV. Post Categorization

Each group in the dataset comprises a diverse range of service providers. To comprehend the ecosystem of TSS groups, it is crucial to break them into various service providers based on the intent presented in the post. These groups operate as informal, uncontrolled marketplaces where traditional systems of trust are missing. There is no confirmed vendor identity, no reputation or feedback system, and no assurance regarding the quality, legitimacy, or delivery of the advertised products or services.

A. Manual Categorization Process

Five manual annotation iterations were performed to categorize the posts, following a process inspired by grounded theory [37]. In each round, two annotators labeled randomly selected posts from the dataset. Any disagreement in the definition or understanding of category labels was resolved after discussions between both annotators. Consistency between annotators was measured using precision, recall, F1 score, Jaccard index, and Hamming loss. The table I lists the outcomes of each round. Each of these indicates the extent of agreement or disagreement among annotators when categorizing sample data.

Round 1 (n = 300): Beginning with manual labeling, the annotators started classifying posts using the categories in the study by Liu et al. [3] as a starting point. The categories include announcements, call centers, victim data sales, expandable attackers, money launderers, null, tech support web masters, and toll-free number providers. The category "expandable attackers" is defined to cover the posts that periodically refer to services such as pay-per-call, popup advertisements, email blasting, SMS blasting, and other intentions that assist promoting and advertising such campaigns.

Round 2 (n = 200): A number of new, more focused categories, such as "PPC/Popup Calls","Blasting Campaign Services" and "Fake/Illicit Documents Services" were introduced. The new categories were not precisely defined, leading to inconsistent labeling by annotators, which resulted in a significant decrease in precision and recall. The previous category "expandable attackers" is broadened into three categories: (1) "Advertising operators", (2) "Blasting Campaign Services" and

(3) "Fake/Illicit Documents Services", with each contributing to a unique offering in these marketplaces.

Round 3 (n = 100): To address Round 2's misinterpretation, preliminary definitions are added to each category based on the intention of the post. Additionally, "Criminal IT Infrastructure Operations" was defined. In order to handle few posts that never fit any predefined categories, "Other" as a category was proposed. The output made a tremendous difference, boosting the accuracy from 0.08 to 0.69.

Round 4 (n = 100): This round included further fine tuning of some category names and definitions. These minor adjustments substantially increased clarity without upsetting earlier advances. Metrics continued to improve marginally, suggesting stable and high agreement.

Round 5 (n = 100): Category labels were refined to improve both clarity and understanding. Despite these modifications, annotator agreement remained high, indicating that the revised definitions were effective. The agreement metric remained consistent with the previous round, demonstrating stability in annotation quality.

The final categories and definitions are provided below:

- ***Blasting Campaign Services***: Services that transmit mass email or SMS messages to prospective victims using the technique called email/SMS blasting, by which inbound calls are generated.
- ***Job Offerings***: Hiring workers that take phone calls at call centers.
- ***Criminal IT infrastructure Operations***: Scripts and tools required to operate criminal IT infrastructure, excluding toll free number providers, website development services and remote access services.
- ***Fake/Illicit Document Services***: Impersonating the documents of legitimate organizations.
- ***Money Launderers***: Services that cash out or transfer stolen payments.
- ***PPC/Popup Calls***: Services that assist scammers in running malware advertising campaigns, such as PPC and pop-ups to generate inbound calls.
- ***Remote Access Services***: Remote access software used to connect to victim devices.
- ***Scammer Warnings***: Warnings to the community about alleged dishonest scammers.
- ***Toll Free Number Provider***: Vendors that provide Toll-free number (TFN) numbers.
- ***Victim Data Sales***: Data of victim contact or personal details.
- ***Web Development Services***: Website services, including hosting and promotion.
- ***Not Related***: Posts that do not describe any component of a cyber crime, or posts written in foreign languages.
- ***Other***: Posts that do not fit with any existing categories.

There was a notable improvement in annotator agreement after the iterative refinement process over five manual annotation rounds, especially in Rounds 4 and 5. All assessment metrics (F1 score, Jaccard similarity, and Hamming loss) demonstrated consistent performance in these rounds, suggesting that all category definitions were now widely accepted and routinely used. Furthermore, neither of these rounds produced any new, discrete categories, indicating that the current taxonomy was adequate to account for the observed differences in the data. This led to the conclusion that the categories had conceptually reached saturation. As a result, the established taxonomy was used as the basis for the automated labeling pipeline.

TABLE I: Evaluation Metrics Over Multiple Rounds

Round	Precision	Recall	F1	Jaccard	Hamming Loss
Round_1	0.433	0.433	0.433	0.433	0.112
Round_2	0.081	0.151	0.102	0.075	0.160
Round_3	0.693	0.740	0.708	0.690	0.054
Round_4	0.738	0.858	0.771	0.712	0.049
Round_5	0.733	0.728	0.725	0.705	0.049

B. Automating Category Assignment

Following the completion of manual labeling, the process turned to automating the labeling process utilizing an In-Context Learning (ICL) approach proposed by Min et al. [38] with justified classification technique, moving beyond traditional keyword-based methodologies. The fundamental objective behind this transition was to mitigate human bias commonly introduced through heuristic keyword selection and to assess whether the proposed categories were both semantically distinct and readily interpretable by language models.

Our approach employs a structured many-shot ICL prompting strategy to guide an LLM in assigning one or more suitable category label to each post across the entire dataset. The prompt is carefully designed not only to generate the most appropriate category label(s), but also to provide a justification for each assigned label. These justifications offer a transparent view into the model's reasoning, aligning with Explainable AI (XAI) principles as discussed by Zhao et al. [39] and Danilevsky et al. [40] and supporting efforts to assess and refine the model's understanding of complex and subtle post intention.

To facilitate the automation process, we utilized 151 manually labeled posts. These posts were selected to ensure coverage of all 13 categories, including instances with multi-label annotations. For example, we manually assigned category labels and constructed justifications grounded in the corresponding category definitions, which served as the rationale to guide the model's classification process. This approach is consistent with explainability-oriented Natural Language Processing technique as highlighted by Zhao et al. [39] and Danilevsky et al. [40], where rationale generation precedes label assignment.

The final prompt template included eight key components: (1) task overview, (2) category definitions and intent, (3) structured many-shot examples, (4) overall distinction guide, (5) instructions, (6) final matching step, (7) the task input (i.e., the text to be labeled), and (8) a category alignment check. Each example in the prompt followed the template:

Post: {post} \n A: Let's think step by step. Pre-defined category(s): [{category}] Reason: This is the {reason}

This design ensures that LLMs not only learn to apply the correct labels, but also to verbalize its reasoning, improving both transparency and reliability. Full prompt template details are provided in the Appendix B.

We conducted multiple iterations of prompt development to refine the performance of LLM in categorizing posts. Initial experiments employed zero-shot and few-shot prompting techniques. However, manual inspection of the model outputs revealed that the produced labels often lacked contextual relevance and failed to align with the intended category definitions. To improve performance, several adjustments were introduced: (a) category definitions were clarified to enhance interpretability, and (b) some category names were revised to reduce ambiguity. Additionally, the categories "Not Related" and "Other" were merged, as the model struggled to distinguish between them due to their overlapping characteristics. Subsequent prompt refinements focused on improving both the consistency and accuracy of the model's output. These included standardizing the output formatting, incorporating category alignment checks to encourage more deliberate reasoning, and experimenting with the positioning of many-shot examples within the prompt. Given the relatively large number of posts (151), there was a risk that the model might lose focus on the defined category schema. To mitigate this, we introduce a high-level distinction guide summarizing the key differences between categories, which served as an anchor for the model's interpretation through the prompt.

To evaluate model performance and consistency, we tested two LLM versions: the original Gemma and the enhanced Gemma 3 (12 billion parameters). A random sample of 100 posts, distinct from those used in the many-shot prompt, was selected for evaluation. Each selected post was manually labeled by a human annotator to serve as ground truth for evaluating the model's accuracy. As shown in Table II, the original Gemma model achieves an accuracy of 94.8%, while Gemma 3 reaches 97.9%, demonstrating improved alignment with human annotations. This evaluation not only highlights the effectiveness of structured prompting with many-shot examples, but also provides insight into how well the LLM internalized the distinctiveness and clarity of our category schema. Based on these results, the remaining dataset was analyzed using the best-performing configuration, Gemma 3 combined with our refined structured many-shot prompt.

TABLE II: Category-wise Accuracy Comparison Between GEMMA and GEMMA3

Category	**Accuracy**	
	GEMMA	**GEMMA3**
Remote Access Services	0.94	0.98
Job Offerings	0.99	0.98
Criminal IT Infrastructure Operations	0.91	0.97
Fake/Illicit Document Services	0.98	0.97
Victim Data Sales	0.95	0.99
Blasting Campaign Services	0.91	0.98
Money Laundering Services	0.88	0.95
PPC/Popups Calls	0.94	0.98
Toll-Free Number Providers	0.94	0.99
Other	0.94	0.97
Scammer Warnings	1.00	0.99
Web Development Services	1.00	1.00
Average Accuracy	**0.948**	**0.979**

V. Group Categorization

In addition to categorizing posts, we observe that the names of the groups themselves often indicate a particular emphasis. Hence, two of the authors manually categorized all 96 groups into four distinct categories based on the group names:

- **Specialized tech support scam service/product groups**: where group name aligns mostly focusing on a specific product or service but restricting to TSS.
- **Specialized scam service/product groups**: where group name aligns mostly focusing on a specific product or service but not restricting to TSS.
- **Geo-localized tech support scam groups**: where group name aligns mostly focusing on a specific city or location that they can operate, restricting to TSS.
- **General tech support scam groups**: where group name aligns mostly not focusing any product or service rather created to contribute to TSS.

VI. Analysis

A. Post and author summary statistics

A total of 321,319 posts were labeled by LLM as a part of the automated classification process. The LLM successfully labeled 97.2% of posts using the defined categories. The remaining 8,733 posts either encountered an error during label generation, introduced a category not included in the predefined set, or produced output from which the label could not be reliably extracted due to formatting inconsistencies.

The *Overall* columns in Table III indicate the number and percentages of authors and posts in each category. The largest number of posts and authors was found in the Money Laundering Services category, followed by Victim Data Sales. Call generating techniques such as Blasting Campaigns, PPC/Popups Calls, and Criminal IT Infrastructure Operations are also popular.

The top 10 groups, ranked by post count from highest to lowest, account for 111,009 posts out of a total of 312,586 posts, representing 35.5%. The top 18 groups accounted for over fifty percent of the total posts in the dataset.

Although the total number of distinct members across all 96 groups is 186,936, only 30,764 individuals have actively contributed posts. Among users who contributed posts, 78.1% posted exclusively in a single group. An additional 11.36% participated in two groups, while 3.73%, 1.76%, and 1.13% of users posted in three, four, and five groups, respectively. Less than 1% of users posted in more than six groups. The

TABLE III: Category-wise Percentages of authors and post

Category	Overall		Specialists		Generalists		One-Time Posters	
	Authors	Posts	Authors	Posts	Authors	Posts	Authors	Posts
Money Laundering Services	16.7%	22.4%	22.3%	30.3%	49.0%	66.7%	28.7%	3.0%
Victim Data Sales	14.3	21.9	12.9	16.4	60.1	81.2	27.0	2.5
Blasting Campaign Services	8.3	11.8	8.7	3.8	70.2	94.1	21.1	2.1
PPC/Popups Calls	6.3	9.7	9.1	4.9	63.6	92.7	27.3	2.5
Other	21.5	8.5	14.3	31.0	41.7	53.5	44.1	15.5
Criminal IT Infrastructure Operations	8.0	7.9	7.7	7.1	67.2	89.4	25.1	3.6
Job Offerings	12.2	6.8	29.5	46.0	29.2	43.6	41.3	10.4
Toll-Free Number Providers	2.8	5.1	4.2	2.7	85.9	96.5	9.9	0.8
Fake/Illicit Document Services	5.1	2.3	4.5	10.7	76.0	83.3	19.4	6.0
Remote Access Services	1.3	1.8	9.7	25.4	80.5	73.7	9.8	1.0
Web Development Services	2.0	1.3	4.6	8.4	64.8	85.0	30.6	6.6
Scammer Warnings	1.5	0.5	3.2	3.8	79.1	88.0	17.7	8.1
Overall	–	–	23.0	18.7	28.9	77.0	48.1	4.3

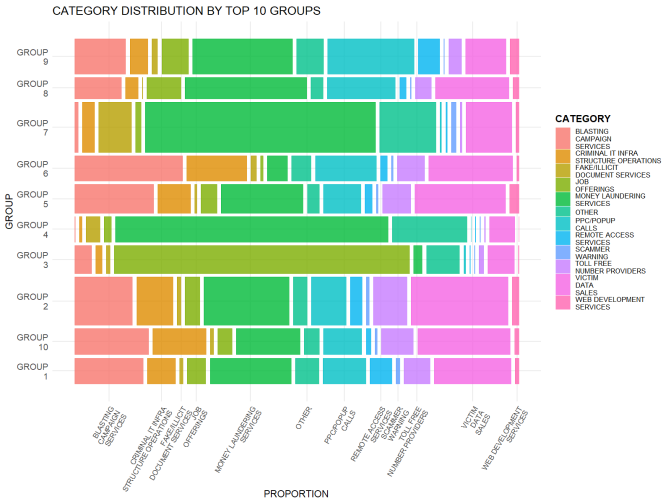

Fig. 2: Mosaic Plot of category distribution by top 10 groups

seven most prominent users were active participants in more than 50 groups.

A significant concentration of activity is observed, with the top 1% of users accounting for 44.5% of all posts. This concentration increases substantially, as the top 10% of users collectively contribute 77.6% of the total posts. The high level of posting inequality is further reflected by a Gini coefficient of 0.81, indicating a pronounced imbalance in user participation.

By examining the distribution of post categories by group, we can investigate whether specialization appears within the groups themselves. Fig 2 shows the distribution of categories among the top 10 groups. We see evidence of specialization in groups 3 (job offerings), 4 (money laundering services), and 7 (money laundering services). However, in other groups, the distribution is more evenly spread across post categories. The distributions of Groups 1, 2, and 5 closely mirror the overall post distribution shown in Table III.

TABLE IV: Results of Group Categorization

Group Category	Total Posts	# Groups
Specialized Tech Support Scam Service/Product Group	128797	40
Specialized Scam Service/Product Group	113563	29
General Tech Support Scam Groups	44832	16
Geo-Localized Tech Support Scam Groups	25394	11

B. Group categories

Table IV breaks down the group prevalence according to these categories. Overall, groups with specialized TSS products and services is the most common, comprising of 40 groups with 128,797 total posts. The second most common groups focused on particular scam service, but is not restricted to TSS alone. General TSS groups are less common, with those geolocalized to particular city (typically in India) occurring in 11 groups.

Between 2008 and 2023, new groups are established at a steady, but not remarkable pace (See Figure 4). The exception was in 2015, where 37 groups were created. Among the 37 groups, the specialized TSS service groups were most common, although eight of the 11 geo-localized groups were also created. In addition, we observed groups that remained in operation over time. For example, the 37 groups established in 2015 remained active when data collection occurred.

Figure 3 illustrates how the content of posts compares with the group category. In most cases, the distribution of posts adheres to the general population. There are some exceptions, however. For example, job offerings tend to appear more often in general TSS groups, while fake/illicit document services happen most often in specialized scam service groups not restricted to TSS.

C. Generalists versus Specialists

Most posts are assigned only one label, with only 5.3% of posts being categorized with multiple labels. Hence, generalization does not appear to happened at the post-level, but rather at the user-level. To that end, we divide users into categories based on post count and their contribution to each category.

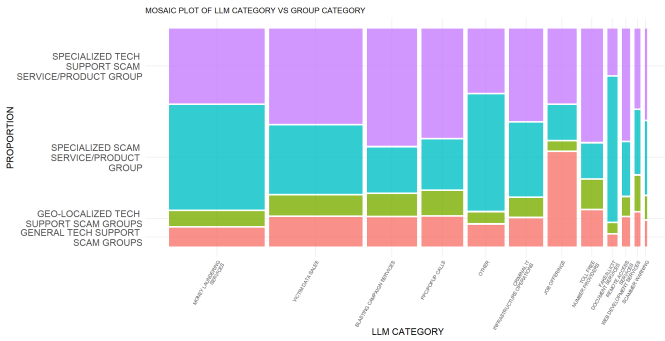

Fig. 3: Mosaic Plot of category distribution by Group categories

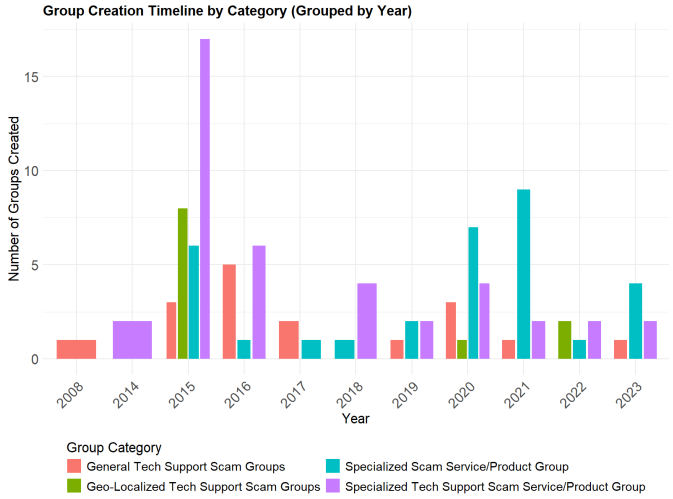

Fig. 4: Time graph Group creation by group category

- **Generalist**: user who posts at least twice and covers multiple categories.
- **Specialist**: user who posts at least twice and focuses on posting only in one category.
- **One-time poster**: user who posted only once across the dataset.

There are 14,789 one-time posters out of 30,764 users (48% of the total), 7,076 (23.0%) specialists, and 8,891 generalists (28.9%). On average, each specialist user made 8.97 posts and generalist users post more than three times as often at 26.4 posts.

1) By Post Categories: Table III shows the breakdown of specialists and generalists across post categories. Posts on job offerings tend to be made primarily by specialist authors. Similarly, posts related to money laundering and remote access services are also made more frequently by specialists, despite specialists representing a smaller share within these product/service categories. For generalist authors, money laundering remains a category where they contributed the most, accounting for roughly two-thirds of all posts. The least active categories – such as TFN providers, fake or illicit document services, remote access services, and web

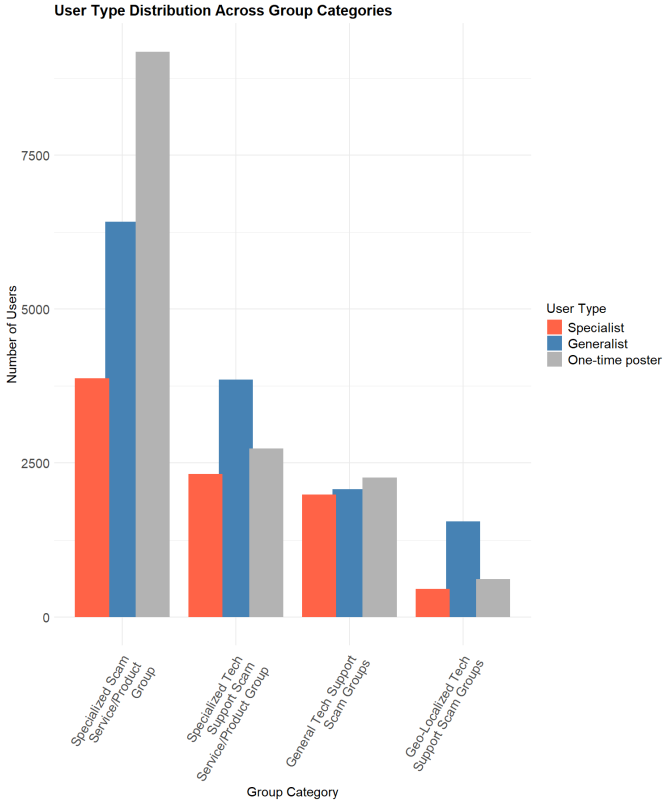

Fig. 5: Generalists versus specialists in group categories.

development services – are largely dominated by generalist authors as well.

2) By Facebook Groups: Across all groups, the *specialized scam service/product groups* had the highest numbers of all three user types, and the *geo-localized TSS groups* had the lowest numbers, as shown in Figure 5. Unlike specialized scam service/product group, specialized TSS service/product groups have a larger number of generalists than one-time posters and specialist users. Across all groups, the number of specialist users is low compared to the other two types.

In addition to user types, we observed variations in product category distribution between groups. Within both specialized Facebook groups, posts on *money launderers*, *blasting campaign services*, and *victim data sales* (see Figure 3). These product categories highlight the essential services and products necessary to initiate any TSS campaign. *Victim data sales* especially dominated the number of posts for specialized TSS service/product categories. On the other hand, there is a lack of product specialization in *geo-localized TSS groups*.

These findings suggest that the Facebook marketplace ecology is characterized by a predominance of non-specialist users operating within specialized group structures. While specialized scam service/product groups attracted the highest number of users across all groups, user-level specialization in services/products remain relatively uncommon across groups. Notably, even within the *specialized TSS service/product groups*, generalists outnumbered specialists, suggesting that

users engage opportunistically across product categories rather than cultivating domain expertise. Moreover, *geo-localized TSS groups* exhibited both lower user activity and a lack of product specialization, which is indicative of a more fragmented operational structure. In contrast, both specialized groups play a central role in facilitating core illicit infrastructures necessary for TSS. These dynamics underscore a marketplace structure in which generalist users rely heavily on specialized groups for essential resources.

VII. Discussion

This study advances the existing knowledge of TSS ecosystem by closely examining service offerings, vendor specialization, and the suitability of large language models for automated categorization. First, the findings showcase a diverse yet structured TSS marketplace ecosystem in terms of services and products offered. Specifically, frequent posts on money laundering services highlight the importance of financial facilitation in the operation of TSS. In other words, the ecosystem relies on intermediary services to transfer and convert money received from scams, instead of actors who execute TSSs. Second, the findings suggest the occurrence of group-level specialization instead of user-level specialization. The creation of specialized scam service/product groups and specialized TSS service/product groups indicates the functional specialization in the TSS marketplace ecosystem. Forty out of the 96 groups focused on specific services. However, among these groups, most vendors were considered generalists. This suggest a calculated adoption where vendors diversify their products and services to maintain competitiveness and limit risks, while groups were created to serve consumer demands. The contrast between the group-level and user-level specialization indicates a marketplace ecosystem with flexible structure and dynamic role distributions.

A primary methodological contribution of this research is to explore the role of LLMs in the classification of complex, unstructured, and real-time data from online discussions among cyber criminals. The LLM annotation achieved a high accuracy of 97.9%, illustrating the value of adopting LLMs in cyber threat intelligence. Few instances of false negative and labeling refusals occurred, indicating need for more context aware prompt engineering and refinement techniques. Overall, the use of LLMs provides a scalable and reproducible approach for examining huge volumes of dynamic cybercrime forum data without significant human intervention.

We ultimately arrived as 12 distinct products and services being discussed in posts. The most commonly occuring categories are money laundering services and victim data sales. We observed that the total number of authors were evenly distributed across generalists and specialists. However, in terms of activity and participation, generalists are more active than specialists. Finally, at the group level, the distributions of messages mirrors the overall population. There are some notable exceptions, however. For instance, the messages in a handful of groups are dominated by one category, notably groups dedicated to money laundering and job offerings.

VIII. Limitations

The study aimed to understand the products and actors participating in the informal marketplaces organized around TSS. However, a number of limitations should be acknowledged. First, the current analysis utilizes data from Facebook groups. Although similar groups exist on multiple platforms, Facebook is one of the largest and most active platforms of its kind, and its rich raw data provided a strong foundation for this work. Nonetheless, scammers discussions taking place on other platforms have not been included. Second, the classification of posts was performed using LLMs with categories derived from five rounds of manual annotation. The LLM achieved an accuracy of 97.9% though a small number of false negatives may exist. Additionally , some posts could not be labeled because LLM sometimes interprets certain requests posts as direct prompt rather than content to classify. Finally, the LLMs used for this study were Gemma3:12B and Gemma original. While this model performed reliably and yielded high accuracy, there exist more powerful and tested models such as Gemma3:30B, Qwen:30B or DeepSeekR1. Due to hardware constraints, these models were not accessible, but future work could leverage such models for higher accuracy, efficiency, and improved performance.

IX. Future Work

Future research could extend the current research in multiple ways. First, the posts can be further categorized, such as by sentiment or whether the post is advertising or requesting a service. These categorization approaches could provide a deeper understanding of the dynamics of the marketplace. Second, the dataset can be incorporated with data from other platforms, which would allow for a more comprehensive view on the TSS marketplace ecosystem. Third, longitudinal analysis at both author and population levels should provide insight not only on trends across the time but also into the process of scammer entry, exit, and overall trajectory in the ecosystem. Finally, researchers can leverage methods such as crime script analysis to understand the operational procedure within these marketplaces and TSS. This would help identify intervention points for targeted disruption measures.

X. Concluding Remarks

This study utilizes a multi-faceted approach to examine the TSS ecosystems. Our work has developed a data-driven, scalable process to categorize the high volume of social media posts. We employed an iterative modified grounded theory approach to generate reliable categories. We then automated the classification process using a structured prompting technique. We fully expect that this approach could be utilized in other cyber crime fora datasets.

Our analysis of product and service distribution reveals key trends in the TSS-relevant Facebook groups: (1) the prominence of money laundering services, (2) the presence of specialization at the group-level and (3) recruitment postings by specialized users across all the groups. While certain groups display domain expertise, user behavior tends to be more

diffuse, with a majority of users being generalists across the ecosystem.

Together, these findings advance the empirical foundation for understanding TSS ecosystems and evaluating computational tools for online crime monitoring.

REFERENCES

[1] D. Harley, M. Grooten, S. Burn, C. Johnston *et al.*, "My pc has 32,539 errors: how telephone support scams really work," *Virus Bulletin*, 2012.

[2] Federal Bureau of Investigation, Internet Crime Complaint Center (IC3), "2024 Internet Crime Report," Internet Crime Complaint Center (IC3), Federal Bureau of Investigation, Tech. Rep., Apr. 2025, iC3's annual report combining data from 859,532 complaints; reported losses reached $16.6billion in 2024—up 33% from 2023. [Online]. Available: https://www.ic3.gov/AnnualReport/Reports/2024_IC3Report.pdf

[3] J. Liu, P. Pun, P. Vadrevu, and R. Perdisci, "Understanding, measuring, and detecting modern technical support scams," in *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2023, pp. 18–38.

[4] N. Miramirkhani, O. Starov, and N. Nikiforakis, "Dial one for scam: A large-scale analysis of technical support scams," in *Proceedings 2017 Network and Distributed System Security Symposium*. Internet Society, 2017.

[5] B. Srinivasan, A. Kountouras, N. Miramirkhani, M. Alam, N. Nikiforakis, M. Antonakakis, and M. Ahamad, "By hook or by crook: Exposing the diverse abuse tactics of technical support scammers," *arXiv preprint arXiv:1709.08331*, 2017.

[6] S. Rauti and V. Leppänen, ""you have a potential hacker's infection": A study on technical support scams," in *2017 IEEE International Conference on Computer and Information Technology (CIT)*. IEEE, 2017, pp. 197–203.

[7] J. Larson, B. Tower, D. Hadfield, D. Edge, and C. White, "Using web-scale graph analytics to counter technical support scams," in *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018, pp. 3968–3971.

[8] Y.-J. Choi and J. Lee, "The change in the methods of smishing in south-korea after the onsent of covid-19," *Pt. 2 J. Legal Ethical & Regul. Isses*, vol. 24, p. 1, 2021.

[9] I. Wood, M. Kepkowski, L. Zinatullin, T. Darnley, and M. A. Kaafar, "An analysis of scam baiting calls: Identifying and extracting scam stages and scripts," *arXiv preprint arXiv:2307.01965*, 2023.

[10] Y. T. Chua, "Sale of private, confidential, and personal data," in *Handbook on Crime and Technology*. Edward Elgar Publishing, 2023, pp. 138–155.

[11] A. Haslebacher, J. Onaolapo, and G. Stringhini, "All your cards are belong to us: Understanding online carding forums," in *2017 APWG symposium on electronic crime research (eCrime)*. IEEE, 2017, pp. 41–51.

[12] T. J. Holt and E. Lampke, "Exploring stolen data markets online: products and market forces," *Criminal Justice Studies*, vol. 23, no. 1, pp. 33–50, 2010.

[13] M. Motoyama, D. McCoy, K. Levchenko, S. Savage, and G. M. Voelker, "An analysis of underground forums," in *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, 2011, pp. 71–80.

[14] N. Christin, "Traveling the silk road: A measurement analysis of a large anonymous online marketplace," in *Proceedings of the 22nd international conference on World Wide Web*, 2013, pp. 213–224.

[15] J. Broséus, D. Rhumorbarbe, C. Mireault, V. Ouellette, F. Crispino, and D. Décary-Hétu, "Studying illicit drug trafficking on darknet markets: structure and organisation from a canadian perspective," *Forensic science international*, vol. 264, pp. 7–14, 2016.

[16] M. Paquet-Clouston, D. Décary-Hétu, and C. Morselli, "Assessing market competition and vendors' size and scope on alphabay," *International Journal of Drug Policy*, vol. 54, pp. 87–98, 2018.

[17] K. Soska and N. Christin, "Measuring the longitudinal evolution of the online anonymous marketplace ecosystem," in *24th USENIX security symposium (USENIX security 15)*, 2015, pp. 33–48.

[18] J. Lusthaus, "Honour among (cyber) thieves?" *European Journal of Sociology/Archives Européennes de Sociologie*, vol. 59, no. 2, pp. 191–223, 2018.

[19] R. Van Wegberg, F. Miedema, U. Akyazi, A. Noroozian, B. Klievink, and M. van Eeten, "Go see a specialist? predicting cybercrime sales on online anonymous markets from vendor and product characteristics," in *Proceedings of the web conference 2020*, 2020, pp. 816–826.

[20] T. J. Holt and J. R. Lee, "A crime script analysis of counterfeit identity document procurement online," *Deviant Behavior*, vol. 43, no. 3, pp. 285–302, 2022.

[21] OpenAI, "Gpt-4 technical report," https://openai.com/research/gpt-4, 2023, accessed: 2025-07-27.

[22] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, S. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," https://ai.meta.com/llama, 2023, accessed: 2025-07-27.

[23] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love *et al.*, "Gemma: Open models based on gemini research and technology," *arXiv preprint arXiv:2403.08295*, 2024.

[24] G. Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière *et al.*, "Gemma 3 technical report," *arXiv preprint arXiv:2503.19786*, 2025.

[25] Y. Hao, Y. Sun, L. Dong, Z. Han, Y. Gu, and F. Wei, "Structured prompting: Scaling in-context learning to 1,000 examples," *arXiv preprint arXiv:2212.06713*, 2022.

[26] Y. Xian, B. Schiele, and Z. Akata, "Zero-shot learning-the good, the bad and the ugly," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4582–4591.

[27] F. Pourpanah, M. Abdar, Y. Luo, X. Zhou, R. Wang, C. P. Lim, X.-Z. Wang, and Q. J. Wu, "A review of generalized zero-shot learning methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 4, pp. 4051–4070, 2022.

[28] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[29] R. Agarwal, A. Singh, L. Zhang, B. Bohnet, L. Rosias, S. Chan, B. Zhang, A. Anand, Z. Abbas, A. Nova *et al.*, "Many-shot in-context learning," *Advances in Neural Information Processing Systems*, vol. 37, pp. 76 930–76 966, 2024.

[30] E. Zelikman, Y. Wu, J. Mu, and N. D. Goodman, "Star: Self-taught reasoner bootstrapping reasoning with reasoning," in *Proc. the 36th International Conference on Neural Information Processing Systems*, vol. 1126, 2024.

[31] V. N. Rao, E. Agarwal, S. Dalal, D. Calacci, and A. Monroy-Hernández, "Quallm: An llm-based framework to extract quantitative insights from online forums," in *Findings of the Association for Computational Linguistics: NAACL 2025*, 2025, pp. 1355–1369.

[32] C. Muasher-Kerwin, M. C. Hughes, M. L. Foster, I. Al Azher, and H. Alhoori, "Exploring large language models for summarizing and interpreting an online brain tumor support forum," *Digital Health*, vol. 11, p. 20552076251337345, 2025.

[33] T. Giannilias, A. Papadakis, N. Nikolaou, and T. Zahariadis, "Classification of hacker's posts based on zero-shot, few-shot, and fine-tuned llms in environments with constrained resources," *Future Internet*, vol. 17, no. 5, p. 207, 2025.

[34] M. Lui and T. Baldwin, "langid. py: An off-the-shelf language identification tool," in *Proceedings of the ACL 2012 system demonstrations*, 2012, pp. 25–30.

[35] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics, 2017.

[36] M. H. Sainte-Marie, D. Kozlowski, L. Céspedes, and V. Larivière, "Sorting the babble in babel: Assessing the performance of language detection algorithms on the openalex database," *Journal of the Association for Information Science and Technology*, 2025.

[37] J. M. Corbin and A. Strauss, "Grounded theory research: Procedures, canons, and evaluative criteria," *Qualitative sociology*, vol. 13, no. 1, pp. 3–21, 1990.

[38] S. Min, X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, and L. Zettlemoyer, "Rethinking the role of demonstrations: What makes in-context learning work?" in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 11 048–11 064.

[39] H. Zhao, H. Chen, F. Yang, N. Liu, H. Deng, H. Cai, S. Wang, D. Yin, and M. Du, "Explainability for large language models: A survey," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 2, pp. 1–38, 2024.

[40] M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kawas, and P. Sen, "A survey of the state of explainable ai for natural language processing," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 2020, pp. 447–459.

APPENDIX A
RESULTS OF LANGUAGE DETECTION MODEL

TABLE V: Performance Metrics of Language Identification Models

Metric	**Accuracy**	**Precision**	**Recall**	**F1**
LangID_org	0.727	0.182	1.000	0.308
LangDetect_org	0.939	0.500	0.875	0.636
FastText_org	0.917	0.412	0.875	0.560
LangID_dec	0.917	0.421	1.000	0.593
LangDetect_dec	0.947	0.533	1.000	0.696
FastText_dec	0.909	0.375	0.750	0.500
Hybrid	0.992	1.000	0.875	0.933

APPENDIX B
PROMPT FORMAT

prompt= f"""You are an expert at analyzing and labeling criminal group marketplace group posts . Your goal is to assign the most appropriate predefined category label(s) to each post, based on **predefined categories, their definitions, and intentions**.
—
TASK OVERVIEW:
—
PREDEFINED CATEGORIES:
—
MANY-SHOT EXAMPLES: fewshot_example
—
OVERALL DISTINCTION GUIDE
—
INSTRUCTIONS:
—
FINAL MATCHING STEP (Before answering):
—
Task: Post: "{post}" A:Let's think step by step.
CATEGORY ALIGMENT CHECK:
"""