Data in Brief 68 (2026) 113171

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

A multidimensional cybercrime annotation dataset for a closed-access Russian–English underground hacking forum

Mariella Mischinger[a,b,*], Sergio Pastrana[b], Guillermo Suarez-Tangil[a,*]

[a] IMDEA Networks Institute, Av. Mar Mediterráneo, 22, 28918, Leganés, Madrid, Spain
[b] Universidad Carlos III de Madrid, Av. de la Universidad, 30, 28911, Leganés, Spain

ARTICLE INFO

Article history:
Received 11 July 2026
Revised 7 August 2026
Accepted 10 August 2026
Available online 24 August 2026

Dataset link: XIN-Annotations: A Multidimensional Cybercrime Annotation Dataset for a Closed-Access Russian–English Underground Hacking Forum (Original data)

Keywords:
Cyber threat intelligence
Text classification
Crime type labeling
Supply-chain categorization
Author intent
Post function
Multilingual corpus

ABSTRACT

This dataset provides a multidimensional cybercrime annotation layer for XIN, a closed-access Russian–English underground hacking forum comprising approximately 1.3 million post annotations across roughly 200,000 threads and 52 subforums, posted between February 2005 and August 2025. Every post is annotated along four dimensions drawn from established underground-forum coding schemes: post type (the communicative function of the post), author intent (the stance or social purpose), crime type (the category of cybercrime discussed), and supply-chain category (the role of the traded product or service). Labels were assigned through a codebook-based prompting procedure that applied a predefined scheme of category definitions, decision rules, and anonymized examples. We release the codebooks and the per-dimension label distributions over the full corpus. The annotations enable supervised and zero- or few-shot classification, benchmarking of automated labeling against a documented scheme, cyber threat intelligence extraction, and cross-lingual analysis of a community that is underrepresented in public corpora.

* Corresponding authors at: IMDEA Networks Institute, Av. Mar Mediterráneo, 22, 28918, Leganés, Madrid, Spain.
E-mail addresses: mariella.mischinger@imdea.org (M. Mischinger), guillermo.suarez-tangil@imdea.org (G. Suarez-Tangil).

https://doi.org/10.1016/j.dib.2026.113171
2352-3409/© 2026 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/)

© 2026 The Authors. Published by Elsevier Inc.
This is an open access article under the CC BY-NC license
(http://creativecommons.org/licenses/by-nc/4.0/)

Specifications Table

Subject	Computer Sciences
Specific subject area	Cybercrime; Underground Hacking Forum Analysis.
Type of data	JSON files referencing forum posts and all inferred labels.
Data collection	The dataset was generated by systematically parsing all posts in the XIN online forum and assigning cybercrime-related labels to each one. We employed a codebook-guided, systematic annotation procedure in which the model assigned labels according to a predefined coding scheme consisting of category definitions, decision rules, and illustrative anonymized examples. The secondary dataset was obtained with custom crawling and scraping software that emulated human browsing to collect raw post text into a relational database.
Data source location	Dataset: produced by applying a codebook through systematic labeling and distributed as part of this work. Secondary dataset: crawled and scraped from a closed-access Russian-English underground hacking forum (referred to as XIN) [1]. Coverage: 1,304,542 posts across 200,889 threads and 52 sub-forums, spanning February 2005 to August 2025 (crawled July 2020–August 2025).
Data accessibility	Repository name: XIN-Annotations Data identification number: 10.5281/zenodo.21237862 Direct URL to data: The released annotation artefacts (codebooks, full-corpus label distributions) are available at: https://zenodo.org/records/21237863 Instructions for accessing the data: The XIN-Annotations dataset is openly available at the link above under DOI 10.5281/zenodo.21237862. All post identifiers and annotations are released together with the codebook under a **Creative Commons Attribution 4.0 International** (CC BY 4.0) license. As mandated by our ethics protocol, the secondary dataset, the full corpus of XIN post data, is restricted to academic researchers who sign a sharing agreement. This data-sharing agreement is governed by the Cambridge Cybercrime Centre (www.cambridgecybercrime.uk/data.html). A curated sample can be swiftly accessed at DOI 10.5281/zenodo.17130287 (https://zenodo.org/records/17130287), from which forum names and personal data have been removed, and which contains no harmful forum content.
Related research article	M. Mischinger, S. Pastrana, G. Suarez-Tangil, Crime VIP: A Closed-Access Underground Hacking Forum, in: Proceedings of the International AAAI Conference on Web and Social Media (ICWSM), 2026 [1].

1. Value of the Data

- The dataset adds a four-dimensional cybercrime annotation — post type, author intent, crime type, and supply-chain category — to approximately 1.3 million posts of a closed-access, predominantly Russian-language underground hacking forum. Closed-access and non-English forums are largely absent from public corpora, which are dominated by English, open-access communities [2]. Researchers in cyber threat intelligence, security, criminology, and natural language processing can therefore study a community and a language that are otherwise hard to reach through the lens of valuable annotations.
- Because every label is tied to a documented codebook and the raw model output is preserved, the annotations can serve as a benchmark for automated underground-forum classification. The released scheme, examples, and per-post outputs let others reproduce, audit, and improve cybercrime analysis pipelines on a large, real-world corpus.
- The crime-type and supply-chain dimensions follow the taxonomies used to detect cybercrime categories [3] and to map cybercrime supply chains [4], while post type and author intent follow the function-and-intent scheme of Caines et al. [5]. The data are therefore directly reusable to train or evaluate classifiers, to extract indicators of compromise [6], and

to study how technical hacking has shifted toward a commoditized market of combinable services [7].
- Reliable machine translation now enables analysts who are not native speakers to work on non-English forum text [8]. Combined with the labels, the dataset supports cross-lingual classification, dark-jargon discovery, social-network analysis of key actors [9], and social studies of underground communities, lowering the barrier to multilingual cybercrime research.
- The codebooks, the full-corpus label distributions, and an identifier of each post are released so that the annotation step can be inspected and replicated independently of the access-controlled corpus. A sentinel *INVALID* label flags labels whose model output fell outside the scheme, letting the community filter or re-annotate them transparently.

2. Background

Underground forums are a central infrastructure of the cybercriminal ecosystem, where participants exchange technical knowledge and trade crime-related tools and services [2,10]. Their analysis supports early detection of emerging threats, because discussions often precede formal threat-intelligence reports. Studying these forums at scale, however, depends on annotated, real-world data. Public corpora such as CrimeBB are dominated by English-language, open-access communities [2], leaving a gap around non-English and closed-access forums, where entry is gated by invitation or payment and where more committed actors gather [11]. The XIN forum — a closed-access, predominantly Russian-language community of approximately 1.3 million posts spanning two decades — was released to help close this gap [1]. The raw corpus, however, carries no post-level labels, which limits its immediate use for supervised learning, cyber threat intelligence extraction, and comparative analysis. To fill this gap, the dataset that we release in this paper adds that missing layer: a four-dimensional cybercrime annotation of every post, built on established coding schemes [3–5]. It turns a large, hard-to-obtain corpus into a resource that can be used directly for classification, benchmarking, and the study of how cybercrime has become a commoditized industry [7].

3. Data Description

The dataset associates each post of the XIN corpus with labels along four cybercrime-related dimensions. Two of the dimensions describe the communicative form of a post, and two describe its cybercrime content:

- **post_type** — the communicative function of the post (e.g., offering or requesting a product/service, proposing a trade, asking for information, or commenting).
- **author_intent** — the stance or social purpose expressed by the author (e.g., neutral, positive, negative, vouching, or redirecting to a private channel).
- **crime_type** — the category of cybercrime discussed, following an established crime-type taxonomy for underground forums [3].
- **supply_chain_category** — the role of the traded product or service within the cybercrime supply chain [4].

The repository is organized as follows. Four plain-text codebooks — *post_type.json, author_intent.json, crime_type.json,* and *supply_chain_category.json* — define every label (label name as key, definition as value, with an anonymized example where applicable). Figs. 1–4 show the label distributions across the released annotations for *post_type, author_intent, crime_type,* and *supply_chain_category*, respectively. A JSON file, *Dataset_labels.json* (1,294,070 records), contains labeled posts. According to our ethics protocol, the secondary XIN corpus is provided separately as a PostgreSQL database through the Cambridge Cybercrime Centre.

Each record is a JSON object with the post identifier (*post_id*), and the four parsed label fields (*post_type, author_intent, crime_type, supply_chain_category*). Three dimensions —

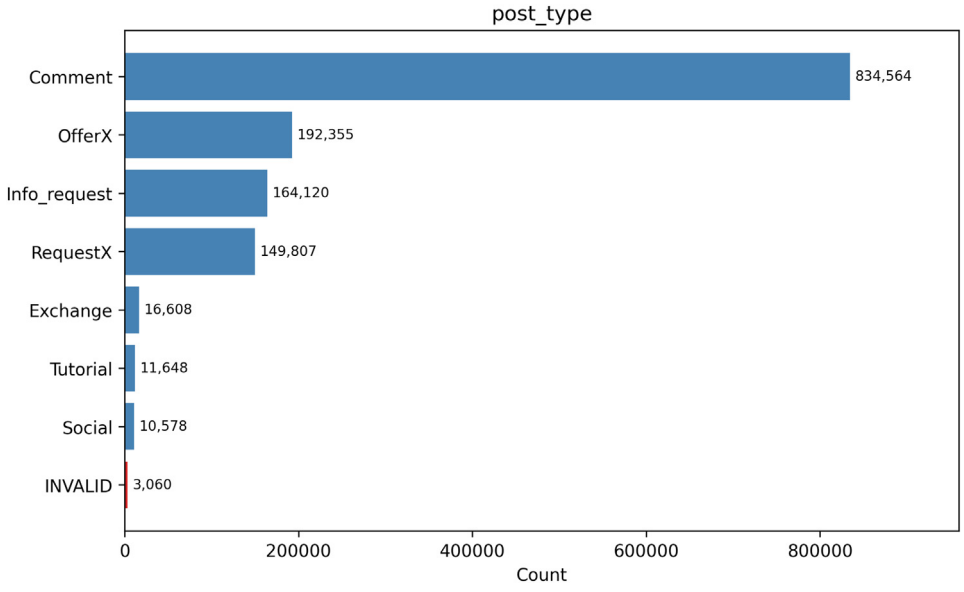


Fig. 1. Distribution of post_type labels over the released annotations. Comment is the most frequent label (834,564), followed by OfferX (192,355), Info_request (164,120) and RequestX (149,807); Exchange (16,608), Tutorial (11,648) and Social (10,578) are less frequent, with 3060 labels INVALID.

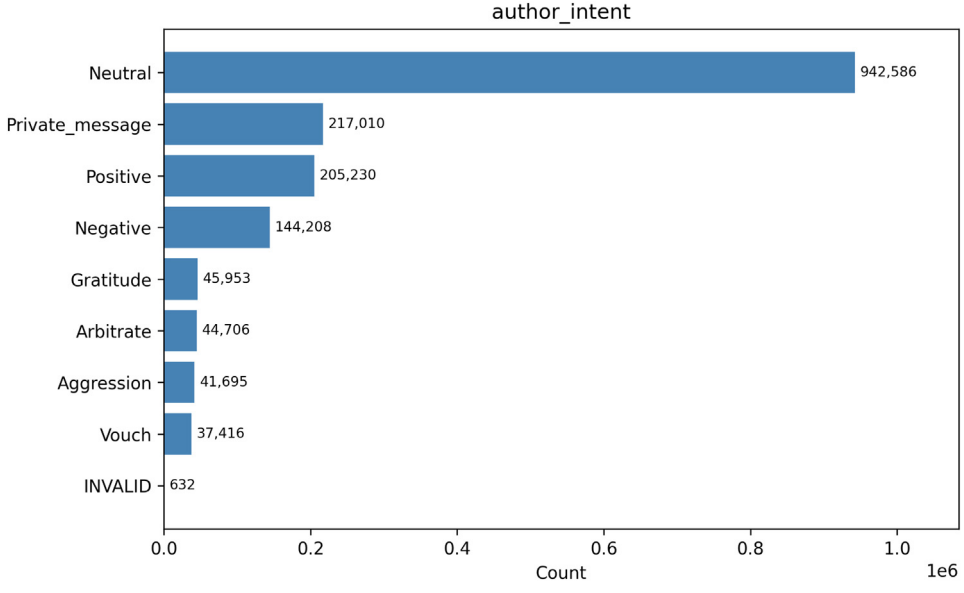


Fig. 2. Distribution of author_intent labels over the released annotations. Neutral dominates (942,586), followed by Private_message (217,010), Positive (205,230) and Negative (144,208); Gratitude (45,953), Arbitrate (44,706), Aggression (41,695) and Vouch (37,416) follow, with 632 INVALID labels.

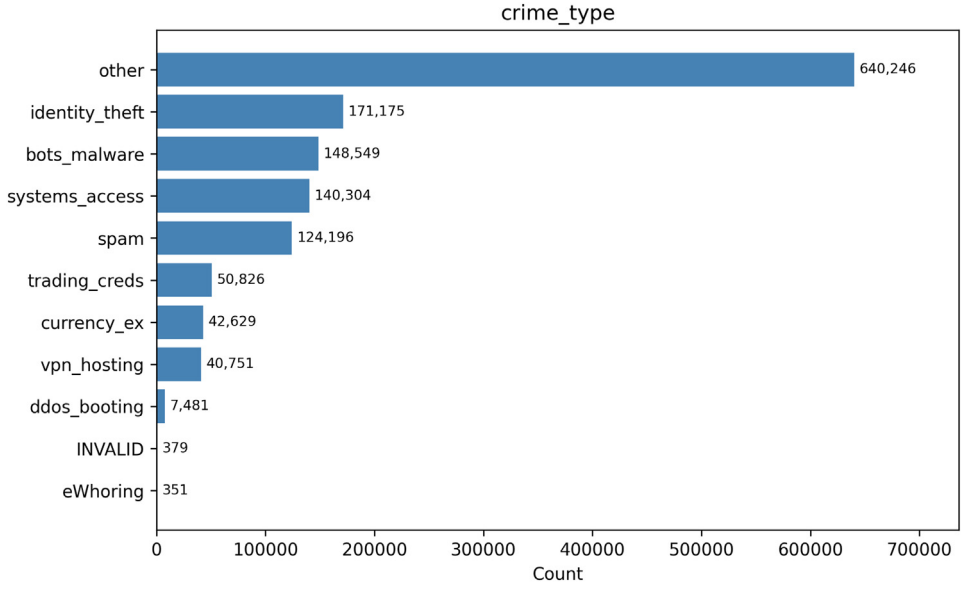


Fig. 3. Distribution of crime_type labels over the released annotations. The *catch-all* other is most frequent (640,246), followed by identity_theft (171,175), bots_malware (148,549), systems_access (140,304) and spam (124,196); trading_creds (50,826), currency_ex (42,629), vpn_hosting (40,751), ddos_booting (7481), eWhoring (351) and 379 INVALID complete the dimension.

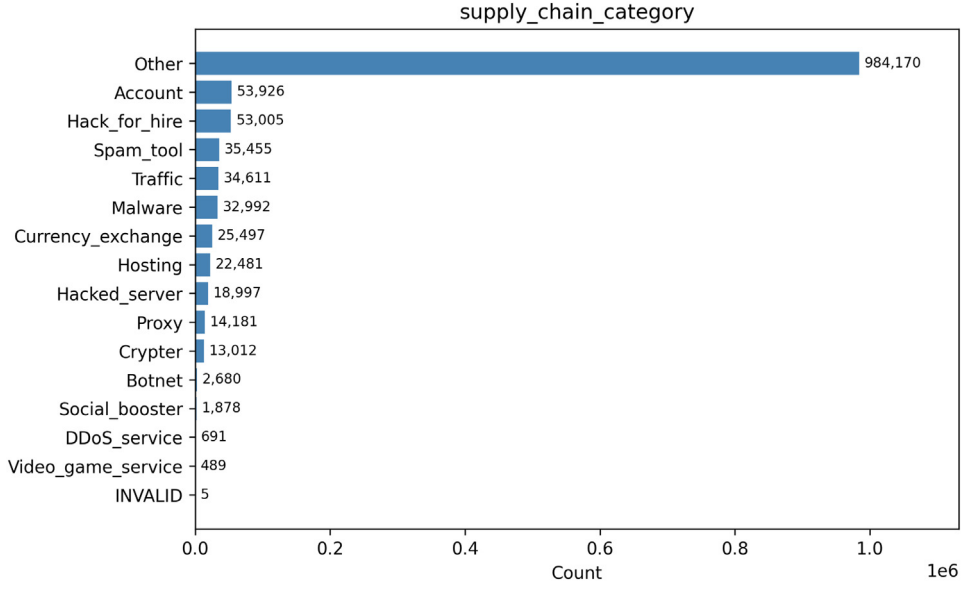


Fig. 4. Distribution of supply_chain_category labels over the released annotations. Other is most frequent (984,170), followed by Account (53,926), Hack_for_hire (53,005), Spam_tool (35,455), Traffic (34,611), and Malware (32,992); the remaining categories range from Currency_exchange (25,497) down to Video_game_service (489), with 5 INVALID values.

Table 1
The post_type dimension: communicative function of a post (7 labels).

Label	Description
OfferX	The user offers a product or service. *e.g., I am looking to sell this domain DOMAIN_NAME*
RequestX	The user requests a product or service. *e.g., Looking for a PRODUCT on here*
Exchange	The user proposes a trade. *e.g., I will buy it for CASH_VALUE*
Tutorial	The user points to a tutorial. *e.g., …here's the steps URL*
Info_request	The user requests information. *e.g., What's the estimated release of the API*
Comment	The user comments on a discussion. *e.g., PRODUCT need to be portforwarded*
Social	The user attempts to connect with another user for social reasons. *e.g., If anyone SOCIAL_ACTIVITY then post here*

Table 2
The author_intent dimension: stance or social purpose of a post (8 labels).

Label	Description
Positive	Posts which are helpful, give praise or encouragement, express amusement or a desire for a product/service. *e.g., nice design BTW*
Neutral	Posts which have no discernible sentiment, positive or negative. *e.g., There's currently NUMBER members registered*
Negative	Posts that state a negative opinion. *e.g., I hate PRODUCT not gonna use it*
Arbitrate	Posts containing admin-esque attempts to control discussion. *e.g., If it's a PERSON, do not charge*
Vouch	Posts that give a positive review of someone's product/service. *e.g., Vouch for this PERSON*
Gratitude	Posts that thank the authors of previous posts for their contributions. *e.g., ty ty hope u enjoy*
Private_message	Posts where the author suggests communication continues through another channel such as Telegram or Discord. *e.g., Going to MESSAGING-SERVICE you now*
Aggression	Posts that are abusive to the previous discussion participant(s) in some way. *e.g., (profanity-laden insult directed at another participant)*

post_type, *author_intent*, and *crime_type* — are multi-label and are stored as lists, whereas *supply_chain_category* is single-label and is stored as a string. Post labels that could not be mapped to a valid codebook label are marked as *INVALID*.

3.1. Label schemes

Table 1 lists the seven *post_type* labels; Table 2 lists the eight *author_intent* labels; Table 3 lists the ten *crime_type* labels; and Table 4 lists the fifteen *supply_chain_category* labels. Definitions and anonymized examples are reproduced verbatim from the released codebooks.

3.2. Label distributions over the corpus

Figs. 1–4 report how often each label is assigned across the released annotations. The four panels are descriptive summaries of the released annotations.

Table 3
The crime_type dimension: category of cybercrime discussed (10 labels), adapted from Atondo Siu et al. [3].

Label	Description
systems_access	Access to systems (excluding use of malware) and SQL-injection attacks. *e.g., How to access a phone's text messages and calls without physical access to it.*
bots_malware	Bots or malware and related services. *e.g., How to make my server file (of RAT) FUD?*
eWhoring	eWhoring (simulation of fraudulent cybersexual encounters for financial gain). *e.g., I am new to eWhoring; can someone give some tips/advice? PM me.*
currency_ex	Exchanging digital currencies. *e.g., Looking for Amazon.de giftcards*
ddos_booting	DDoS attacks, booting, stressing, and stress testing. *e.g., Would you be interested in investing in an SST service? Money would be made back plus more.*
identity_theft	Online identity theft, internet fraud, online scams or credit-card fraud. *e.g., I want to buy an eBay USA SMS account-verification service*
spam	Sending spam, sharing email addresses or containing marketing services. *e.g., Earn passive money with clickbank*
trading_creds	Trading accounts including gaming, social-network and streaming accounts. *e.g., Selling sickest kik*
vpn_hosting	VPN, hosting and proxy services. *e.g., Looking for someone to host DOMAIN in return for a share in the profits.*
other	None of the above crime types detected.

Table 4
The supply_chain_category dimension: role of the traded product or service (15 labels), following Bhalerao et al. [4].

Label	Description
Account	Selling or requesting an account, multiple accounts, or access codes; also includes account-creation automation software.
Botnet	Selling or renting access to computers infected with malicious software.
Crypter	A piece of software which obfuscates malware.
Currency_exchange	Exchanging one form of currency for another.
DDoS_service	Selling or requesting a DDoS attack.
Hacked_server	Selling or requesting a single hacked server.
Hack_for_hire	Offering targeted hacking, malware coding, or requesting a specific service.
Hosting	Hosting a website, game server, or otherwise maintaining it; includes DDoS mitigation.
Malware	A piece of malicious software executed on a victim's machine (e.g., cryptocurrency miners, ransomware).
Proxy	Selling or requesting a proxy/VPN.
Social_booster	Supports gaining social-media attention (e.g., buying likes/views, selling followers).
Spam_tool	Selling or requesting an email/chat spam tool or spamming service.
Traffic	Selling real or fake visitors to a site; does not include social-media traffic.
Video_game_service	Selling or requesting a video-game service (mods, points, power-leveling); excludes video-game accounts.
Other	Anything that does not fall into the previous categories.

4. Experimental Design, Materials and Methods

4.1. Secondary dataset and data acquisition

The annotation layer is built on the XIN corpus, whose collection is described in the related research article [1]. In brief, the forum is closed-access and was reachable only through a single account provided by an industry partner, so the crawl could not be parallelized. To avoid detection and account banning, custom crawling and scraping software emulated human navigation in terms of connection patterns and fetch rates — an approach previously described by CARONTE [12] and discussed for adversarial collection environments [13] — and downloaded only raw post text. The crawl ran from 20 July 2020 to 27 August 2025 and required several iterations because new posts appeared while earlier iterations were still running. The resulting corpus is stored

as a relational database with three tables (Forum/sub-forums, Thread and Post) and comprises 1,304,542 posts across 200,889 threads and 52 sub-forums, posted between February 2005 and August 2025. No special hardware was required.

4.2. Unit of annotation, inclusion and exclusion

The unit of annotation is the individual post. Every post available in the corpus was submitted for labeling; the annotation does not sub-sample, so the label distributions in Figs. 1–4 are computed over the entire corpus. Content that the crawling account could not retrieve, as well as permission-granted posts hidden behind post-count thresholds (approximately 0.5% of the forum corpus) are also absent from the annotations. No language filter was applied: posts were labeled regardless of language, the corpus being predominantly Russian (72.4%), followed by English (14.4%) and other languages [1].

4.3. Label schemes and their provenance

The four dimensions reuse coding schemes that are well established for underground-forum research. The *post_type* and *author_intent* dimensions follow the function-and-intent scheme of Caines et al., which classifies posts by their function, the author's intent, and the addressee [5]. The *crime_type* dimension adopts the crime-type taxonomy of Atondo Siu et al. [3], with the original "Not criminal" category rendered here as *other* so that posts discussing criminal activity outside the listed categories can still be captured. The *supply_chain_category* dimension follows the supply-chain mapping of Bhalerao et al. [4], which organizes traded products and services by their role in the cybercrime supply chain. Each scheme is materialized as a codebook (label, definition, and, where the source provides one, an anonymized example), released alongside the data.

4.4. Annotation

Labels were assigned using a codebook-guided prompting procedure, building on the summary provided in the specifications table. In particular, we adopt an in-context learning approach with few-shot prompting. The model performs classification without any parameter updates, relying solely on task instructions, label definitions, and a small number of labeled demonstrations, i.e., example posts paired with their correct labels. We adopt this approach because prior work has demonstrated its reliability for labeling crime-related content [8].

For each post, the model received the relevant codebook as a dictionary whose keys are label names and whose values are definitions and examples. The prompt also encoded decision rules for ambiguous cases — for example, posts that discuss a crime without committing or transacting it are assigned the catch-all category, and 'botting' or 'hosting' activity is judged in context because it is not always criminal — together with the post's context: the sub-forum title, the thread headline and the post content.

Multiple labels were permitted for post_type, author_intent and crime_type, whereas supply_chain_category received a single label. The model returned a structured JSON object; the raw response is parsed into the four label fields. Within a response if a label could not be mapped to a valid codebook label, it was replaced by "INVALID". This procedure mirrors the labeling setup validated for crime-type classification on multilingual forum data in related work, where the same prompt structure and a temperature of zero were used to keep outputs deterministic [8].

Software and configuration: custom crawling/scraping software (relational database back-end); PostgreSQL for the corpus; an LLM accessed through its API for codebook-guided labeling; and standard plotting tools for Figs. 1–4. The model and version used for annotation is "gpt-5.4" (snapshot gpt-5.4-2026-03-05), via the Chat Completions API (v1/chat/completions), decoding

Table 5
Invalid-label counts and rates by annotation field. L_f = label instances; E_f = INVALID instances; $r_f = \frac{E_f}{L_f}$; $|A_f|$ = posts with ≥1 INVALID label in field f.

Field	Instances L_f	Invalid E_f	Rate r_f	Posts affected $\|A_f\|$
post_type	1,382,740	3060	0.22%	3059
author_intent	1,679,436	632	0.04%	632
crime_type	1,366,887	379	0.03%	374
supply_chain_category	1,294,070	5	0.0004%	5
Total	**5,723,133**	**4076**	**0.07%**	**4056**

temperature is 0, and using JSON mode to enforce structured output (response_format = {"type": "json_object"}). All annotations were made with the prompt found in Appendix I of this work. The annotation processes used over 2,500,000,000 tokens and cost between $3000 and $4000,[1] which was covered through a gift awarded as part of the OpenAI Cybersecurity Grant.

4.5. Preprocessing and content handling

The corpus is released as raw text, and the annotation operated on that text without additional cleaning to offer wider context, so labeled posts retain the crawler's in-text markers. Non-textual or referential content is represented by delimiters placed around keywords: ***IMG*** (image URL), ***IFRAME*** (external embed, usually video), ***LINK*** (other external URL), ***CODE*** (explicit source code), ***CITING*** (identifier of a cited post) and ***CITING MISSED POST*** (plain text of a cited post whose identifier could not be retrieved).

4.6. Quality control and validation

Our first approach to control quality is deriving a standard and auditable annotation scheme.

- For this, a fixed codebook with explicit definitions, decision rules and anonymized examples was applied uniformly to every post, reducing inconsistency across the corpus.
- The post-processed model output is also stored together with the post identifier in *Dataset_labels.json*, so each parsed label can be traced back to the post it was derived from.

We further assess two aspects of our dataset to ensure label quality and traceability: (1) out-of-scheme handling, and (2) reliability of the labeling process; as discussed next.

Out-of-scheme handling. Labels that fall outside the codebook are flagged with *INVALID* rather than silently dropped. Examples are empty lists or strings, or labels taken from another category (a supply_chain_category label appears in the crime_type list). Remaining valid labels are retained. This allows the community to decide whether affected label categories can still be used, should be filtered out, or re-annotated.

For each annotation field f, let L_f be the number of label instances, E_f the number of flagged INVALID labels, and A_f the set of posts containing at least one INVALID label in f. We report the per-field invalid-label rate and the overall affected-post fraction as

$$r_f = \frac{E_f}{L_f}, \quad \rho = \frac{\left|\cup_f A_f\right|}{N}, \quad N = 1,304,542. \tag{1}$$

Table 5 gives L_f, E_f, r_f and $|A_f|$ per field. This affects 3060 label instances (E_f) across 3059 posts (A_f) for post_type (one post has two INVALID labels), 632 label instances (E_f) across 632 posts (A_f) for author_intent, 379 label instances (E_f) across 374 posts (A_f) for crime_type and 5 label instances (E_f) across 5 posts (A_f) for supply_chain_category. Posts may have invalid labels

[1] The exact cost cannot be determined precisely due to the limited granularity of the billing overview.

Table 6
Annotator A1 and A2 agreement scores. Agreement of annotators A1 and A2 regarding each codebook.

label_type	P_o	$P\emptyset$	$P_\emptyset$	Cohen κ_{exact}	Krippendorff α_{exact}	Cohen κ_{label}	Krippendorff α_{label}
post_type	70.0% (105)	8.7% (13)	21.3% (32)	0.56	0.56	0.70	0.70
author_intent	54.0% (81)	16.0% (24)	30.0% (45)	0.34	0.34	0.59	0.58
crime_type	62.0% (93)	6.0% (9)	32.0% (48)	0.37	0.37	0.61	0.61
supply_chain	57.3% (86)	0%	42.7% (64)	0.34	0.34	0.55	0.55
Average	**60.8% (365)**	**7.7% (46)**	**31.5% (189)**	**0.40**	**0.40**	**0.61**	**0.61**

in more than one codebook field (10 posts have INVALID labels in two fields, 2 posts in three fields), hence there are in total 4076 INVALID label instances (E_f) across 4,056 posts (A_f) making $\rho = 0.31\%$ in the dataset.

Reliability evaluation. Furthermore, to assure reliability of the labeling, we compare the generated labels with a set of posts manually annotated by domain experts. We first report the annotation process, then assess annotator agreement, and finally compare the ground truth with the automatically generated labels.

Annotation Process. In a first labeling round, two human domain experts A^1 and A^2, first and second authors of the paper, manually annotate a random sample of 150 to-English translated posts (related work found annotations on translated posts are as effective as in their original language [8]). The process is blinded so each annotator knows neither the LLM's nor the other human's choices. In our annotation scheme, an annotation field for post_type, author_intent, and crime_type is not restricted to a single value; a field may be assigned several labels at once. After the first annotation round, each post *i* is therefore associated with a family of annotator-specific label sets $A^1{}_i$ and $A^2{}_i$, where $A^{1/2}{}_i$ denotes the set of labels assigned to post i by annotator 1 or 2.

Inter-annotator agreement. Because fields can hold multiple labels rather than a single one, comparing the two annotators' outputs requires distinguishing three situations:

- Full agreement: $A^1{}_i = A^2{}_i$ – both annotators assign exactly the same labels. We report P_o as the proportion of posts where $A^1{}_i = A^2{}_i$.
- Partial agreement: otherwise – the annotators agree on some but not all labels. We report $P\emptyset$ as the proportion of posts where $A^1{}_i \cap A^2{}_i \neq \emptyset$ and $A^1{}_i \neq A^2{}_i$.
- No agreement: $A^1{}_i \cap A^2{}_i = \emptyset$ – the two label sets are disjoint, sharing no labels. We report $P_\emptyset$ as the proportion of posts where $A^1{}_i \cap A^2{}_i = \emptyset$.

To correct for the agreement expected by chance, we compute Cohen's and Krippendorff's scores κ and α, two standard measures of inter-annotator agreement.

Standard Cohen's κ and Krippendorff's α scores assume that each item is assigned exactly one category from a mutually exclusive set, an assumption that does not accommodate cases in which a single field may take on multiple labels. We handle this by computing the following two approaches:

- Exact set match, either "Full agreement", else "No agreement". That is, we adopt a strict exact-match criterion, in which a pair of annotations is scored as full agreement (1) if and only if the two label sets coincide exactly, and as no agreement (0) otherwise, thereby treating any partial overlap as complete disagreement. We use κ_{exact} and α_{exact} to denote Cohen's and Krippendorff's exact set matches.
- Partial set match as per-label averaging. We compute agreement for each label, as a binary present/absent decision [14], and report the pooled score across labels, reflecting the complete set of binary decisions. We use κ_{label} and α_{label} to denote Cohen's and Krippendorff's partial set matches.

Table 6 shows the inter-annotator agreement results. We see that observed agreement (P_o) ranges between 54% and 70%, averaging 60.8% for all codebook-decisions, with the dual, full disagreements ($P_\emptyset$), ranging from 21% to 42.7%, and averaging 31.5%. Except for the sup-

ply_chain_category, all other label types observe some partial agreement (P∅), averaging 7.7% of the codebook decisions. Turning to the chance-corrected measures, the exact-match scores average $\kappa_{\text{exact}} = \alpha_{\text{exact}} = 0.404$, placing overall agreement in the *moderate* band. When agreement is instead credited per label rather than requiring an identical set match, both metrics rise to $\kappa_{\text{label}} = \alpha_{\text{label}} = 0.61$, corresponding to *substantial* agreement.

The difference between the exact-match and per-label coefficients arises because exact-set matching treats partial overlaps as complete disagreements, whereas per-label scoring credits the labels on which the annotators concur. Under both scoring schemes, however, the coefficients exceed the levels attributable to chance agreement given the observed label prevalence, indicating that the agreement reflected in $P_o$is not solely an artifact of the skewed label distribution. In this context, the prevalence of P∅ and the difference between the exact-set and per-label for both Cohen's and Krippendorff's scores reflect a difficult or fine-grained scheme rather than an unreliable annotation.

To further characterize partial agreements, we perform a second iteration of annotations, recruiting a third annotator, the last co-author. In particular, the annotators met to review disagreements, discussing posts' content. As a result, the annotators found various reasons for disagreements that overall made it difficult to disambiguate discrepancies. Given this difficulty, rather than resolving the disagreements into a single reference set, we assess how the generated labels align with either annotator, A^1 or A^2. Appendix II provides further detail on this rationale and, for completeness, reports the results obtained over the consensus set derived from the second iteration.

The inherent difficulty in annotating this type of datasets, even through human experts, is further supported by previous findings that point to there being more than one "correct" answer due to differences in context and interpretation [8].

LLM performance. We test the LLM's performance considering both label sets of annotators A1 and A2 as valid options. Fig. 5 shows the agreement with at least one human annotator at

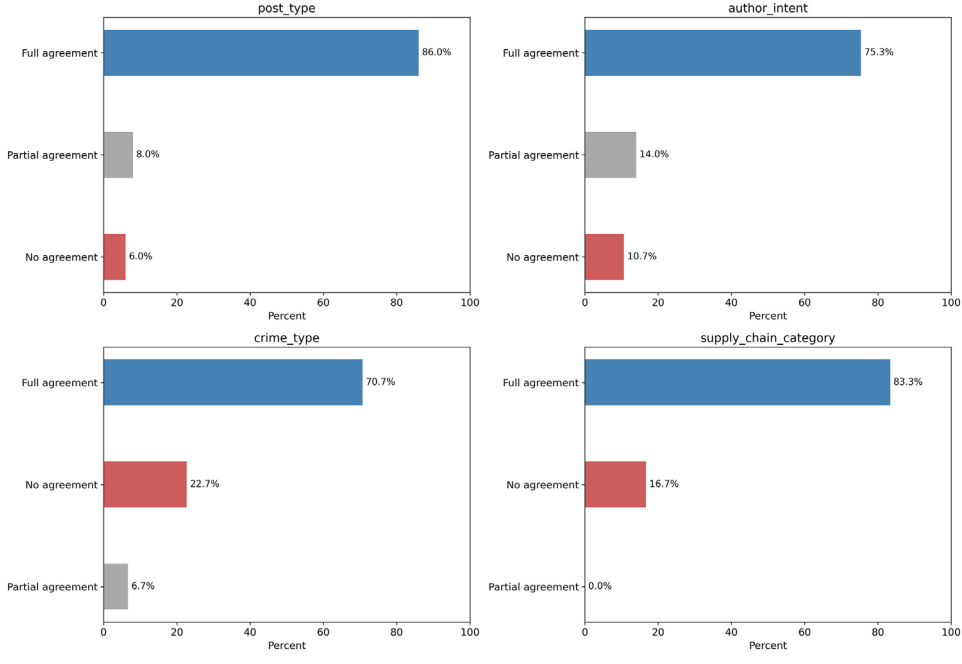


Fig. 5. Agreement of LLM with at least one annotator regarding each codebook. The LLM shows a high agreement with 86% for post_type, 83.3% of supply_chain_category, 75.3% for author_intent, and 70.7% for crime_type. Furthermore, the LLM partially agreed on post labels for 14% of author_intent, 8% of post_type, and 6.7% of crime_type.

Table 7
Agreement of the LLM with either of the annotators A1 and A2 at the label-level. We observe a good performance with an F1-score of 90–75%.

Codebook	gpt_ label_count	human_ label_count	TP	FP	FN	Precision	Recall	F1-score
post_type	160	157	143	17	14	0.89	0.91	0.90
author_intent	193	170	154	39	16	0.8	0.91	0.85
supply_chain_category	150	150	125	25	25	0.83	0.83	0.83
crime_type	161	153	117	44	36	0.73	0.76	0.75
Total	**664**	**630**	**539**	**125**	**91**	**0.81**	**0.86**	**0.83**

the post-level. In most cases the LLM fully agrees with the labels given by an annotator (86–70.7%), performing best for post_type (86%), supply_chain_category (83.3%), and a few partial agreements (6.7–14%). For a minority of cases the LLM fully failed the annotation (full disagreement) for crime_type (22.7%), followed by supply_chain_category (16.7%), author_intent (10.7%) and post_type (6%).

A more nuanced measurement of the LLM performance is presented in Table 7, where we calculate precision and recall of the LLM labels with A1 and A2 separately for each post, retain the higher scoring set, and summarize across all posts. For all categories we find the LLM works reliably – it matches the annotators most for post_type with an F1-score of 0.90, followed by author_intent (0.85), supply_chain_category (0.83), and crime_type (0.75), making an overall F1-score of 0.83.

Takeaway and Qualitative analysis. Overall the LLM performs reliably both on posts where annotators agreed, and, where they disagreed, in matching one of them. This supports the reliability of this work.

A qualitative analysis on human and LLM disagreements reveals insights aligning with our previous work [8], where we find that for all codebooks the human annotators, as well as the LLM disagree on labels due to ambiguous post content or different interpretations. For example, many disagreements in the codebook "author_intent" stem from the subjectivity of "Positive", "Negative" and "Neutral". For "supply_chain_category" annotators disagreed in the type of category, while the LLM appears to have taken the instructions more strictly and sorted posts more often into "Other".

4.7. File organization

The released annotation artifacts are: the four codebooks (**.json*); the four full-corpus distribution figures (Figs. 1–4); and the labels (*Dataset_labels.json*). To prevent misuse, the secondary dataset XIN (PostgreSQL) is distributed under controlled access through the Cambridge Cybercrime Centre.

4.8. Worked example

To illustrate how the four dimensions combine, consider a representative request post from a sub-forum dedicated to servers, proxies and hosting (anonymized; identifying details removed). The thread headline reads "Need VPS that ignores abuse," and the post body is:

"Looking for VPS providers that can ignore abuse, notably malware-hosting abuse."

Given the codebook and this context, the procedure returns the following labels (the model's structured output, stored in the labels field):

```
{
"post_type": ["RequestX"],
"author_intent": ["Neutral"],
```

Table 8
Anonymized worked examples of four-dimensional annotation drawn from the labeled sample.

Post (anonymized)	post_type	author_intent	crime_type	supply_chain_ category
"Looking for VPS providers that can ignore abuse, notably malware-hosting abuse."	RequestX	Neutral	bots_malware, vpn_hosting	Hosting
"CASH_VALUE — verification on your victim data; full escrow."	OfferX	Neutral	identity_theft	Hack_for_hire
"RDP, local admin, DE; industry: call/business centres; AV: Defender; escrow; start CASH_VALUE, step CASH_VALUE."	OfferX	Neutral	systems_access	Hacked_server

"crime_type": ["bots_malware", "vpn_hosting"],
"supply_chain_category": "Hosting"
}

The post is a request (*RequestX*) phrased without discernible sentiment (*Neutral*); it touches two crime types simultaneously — malware-related activity (*bots_malware*) and hosting/VPN provision (*vpn_hosting*) — which illustrates the multi-label nature of the *crime_type* dimension; and the traded role is *Hosting*. The labeled record additionally stores the post identifiers so that every annotation can be traced back to its input. Table 8 shows three anonymized examples spanning different post types and supply-chain roles (monetary amounts replaced by CASH_VALUE).

Reproducibility: the codebooks, the structured per-post outputs and the released outputs make it possible to re-derive, audit and extend the annotation independently of the access-controlled corpus.

5. Limitations

The dataset reflects a single, closed-access forum collected through one account, and it inherits the constraints of that source. The restricted-access sub-forum could not be retrieved, and hence is not included in the labeling. Furthermore, permission-gated posts (approximately 0.5% of the corpus) are excluded from labeling, as their actual content is hidden, so a small number of threads are partial. The corpus is linguistically and geographically skewed: it is predominantly Russian (72.4%), followed by English (14.4%) and other languages, so the labels may not generalize to other communities or languages. It is also temporally broad (2005–2025); because data freshness affects conclusions, older posts may carry historical rather than operational value.

With full-scale human annotation of 1.3M posts completely infeasible, the annotations themselves carry uncertainty: labels were assigned by a language model following a codebook. We note that LLMs may introduce errors and biases, both inherited from training data and arising from ambiguous cases. While we introduced the INVALID flag to capture cases the model itself deemed unclassifiable, this mechanism only accounts for uncertainty the model explicitly recognizes. However, our evaluation puts this in perspective: expert annotators themselves disagree, and the LLM's performance with individual annotators ($F1 = 0.83$) approaches this human level of consistency. While the resulting labels should be treated as model-mediated estimates rather than ground truth, we argue that such automated annotation remains a valuable tool for navigating collections of this scale, especially when downstream conclusions rest on aggregate patterns or individual LLM inferences are used alongside a human-in-the-loop approach as machine labeling cannot match human evaluation quality [8].

Finally, the full corpus is access-controlled under a data-sharing agreement that permits academic use only.

6. Ethics Statement

The work uses textual data collected from an online underground forum. Informed consent from individual members was not sought: doing so would have required private messaging that could be construed as interference and could have altered behavior, and the data are analyzed only as a collective whole, without aiming to identify individual users — consistent with established ethical guidance for research on datasets of illicit origin [2,15]. Forum members generally have strong incentives to remain anonymous and rarely disclose personally identifiable information; no such information was encountered, the authors did not attempt to deanonymize any user, and the data-sharing agreement governing distribution prohibits any deanonymization attempt. The crawler was designed to emulate human browsing to avoid disrupting the platform's operation. To prevent harm, the name of the forum and all references to specific individuals or groups are withheld, and the corpus is distributed under controlled access through the Cambridge Cybercrime Centre, restricted to academic researchers who sign the agreement. Ethics approval for collecting and analyzing the forum data was obtained under application number LRS-19/20-17377, following a risk assessment submitted to the institutional review body of the former institution of the senior author and to the principal investigator of the project in which data collection was initiated. The present annotation layer derives solely from the already-collected text and introduces no new data collection from the platform or its members. The authors confirm that they have read and follow the ethical requirements for publication in Data in Brief.

CRediT Author Statement

Mariella Mischinger: Conceptualization, Methodology, Software, Data curation, Investigation, Visualization, Writing – original draft. **Sergio Pastrana:** Conceptualization, Methodology, Supervision, Writing – review & editing (including guidance on ethical and data-sharing procedures, drawing on prior development of the CrimeBB research infrastructure). **Guillermo Suarez-Tangil:** Conceptualization, Supervision, Funding acquisition, Project administration, Writing – review & editing.

Data Availability

XIN-Annotations: A Multidimensional Cybercrime Annotation Dataset for a Closed-Access Russian–English Underground Hacking Forum (Original data) (Zenodo).

Acknowledgments

This work was supported by the OpenAI Cybersecurity Grant Program and by project PID2022-143304OB-I00 funded by MICIU/AEI/10.13039/501100011033 and by the ERDF, EU. Guillermo Suarez-Tangil is a 2020 Ramón y Cajal fellow (RYC2020-029401-I) funded by MCIU/AEI/10.13039/501100011033 and the ESF Investing in your future; the same grant funded Mariella Mischinger's initial work. AI tools were used to support the LLM-assisted annotation, coding, as well as to summarize and contextualize the related research article, improve the writing style, clarity, grammar, and syntax.

Declaration of Competing Interest

The authors declare that their results are presented as derived from the methodology and processing pipeline and that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix I

Label generation prompt and system message adapted from the crime-type classification methodology in [8], originally proposed by [16], which we adapt for XIN through role assignment to suit the requirements of our labeling task. messages = [

{“role”: “system”, “content”: “You are a helpful assistant for labeling posts.”},

{“role”: “user", “content": chatGPT_label_prompt}] chatGPT_label_prompt = ’”You are a labeling bot specializing in labeling forum posts in RU and EN.

You have deep knowledge of RU and EN, including slang related to hacking, internet, military, finance, and vulgar or colloquial terms.

Do not explain the labels, just provide them.

There are four annotation categories: post_type, author_intent, crime_type, and supply_chain_category.

Each category and label instructions are explained in detail below. The labels and explanations will always be provided as dictionaries where the labels are the keys and the explanations are the values.

Multi-labels are allowed for post_type, author_intent, and crime_type. However supply_chain_category can only have one label. Only pick the exact label names provided in the dictionary.

Do not mix labels across categories. Before sending your answer confirm if the labels given are indeed existent in each category (e.g. Private_message is an author_intent label and can not appear in post_type).

You should return your answer as valid JSON. Do not explain your labels.

Return the JSON in this exact structure:

{{

"post_type": ["<label1>", “<label2>"],

"author_intent": ["<label1>”, “<label2>"],

"crime_type": ["<label1>", “<label2>"],

"supply_chain_category": “<label1>"

}}

Explanation for post_type:

Post type indicates the general function of the post. An offerX and a requestX involve products and services and are opposites of each other.

Similarly, we see exchange posts when a user proposes the trade of a product and/or service for a currency amount or another product and/or service in kind.

Informational post types include tutorials in which a link or guide showing how to perform a certain task is posted for others to follow.

An info request is a request for help, guidance or information, often being the first post in a thread conversation.

A comment is any response which is simply informative, whether it is the first post in a thread, or a response to the post(s) of others.

Finally, a social post is one in which a user proposes subsequent interaction, perhaps through another medium such as networked gaming.

This is the dict with the post_type labels: {post_type}

Explanation for author_intent:

Author intent relates to author affect and what they seem to have intended by the specific wording of their posting.

This is the dict with the author_intent labels: {author_intent}

Explanation for crime_type:

Some posts discuss crime types but not the actual commission of the crime being discussed. For example, one of the posts on the bulletin board called ‘Suggestions and Ideas’ asks for ‘Death Removal of eWhoring’. The discussion focuses on whether the topic of ‘eWhoring’ should or could be removed from the forum. These types of posts are classified as ‘other’.

Some posts discuss products or services, but do not indicate they are selling or using them. For example, some posts discuss software that can automatically increase forum participants' social network channels' subscribers by the hundreds. Where there is no indication the products were being used or sold, these posts are classified as 'other'.

Posts about 'Botting' and 'Hosting' can be particularly difficult to classify as they are not always related to criminal activity. For example, the use of bots for enhanced game playing is a common topic. While this may be against a game's terms of service, no crime is being committed. Therefore, analysing the post context is crucial for categorising the post correctly. As such, these types of posts are classified as 'other'.

This is the dict with the crime_type labels: {crime_type}

Explanation for supply_chain_category:

First decide whether the post is about a product or service that could be part of a cybercrime supply chain.

A post is product-related only if it clearly offers, requests, buys, sells, trades, reviews, discusses availability of, or asks about a concrete product or service.

Other posts such as mere social chat, insults, thanks, moderation, general discussion, vague comments, tutorials without a product/service, or unrelated technical discussion should be marked as "Other".

There can only be one label for supply_chain_category.

This is the dict with the supply_chain_category labels: {supply_chain_category}

Ensure accuracy and context-appropriate labeling as per these guidelines.

You have information about the post, the thread and the sub-forum the post appears in.

This is the sub-forum title: {subforum_title}

This is the general thread headline: {thread_title}

This is the post content: {text}

"""

Appendix II

To further characterize partial agreements, we perform a second iteration of annotations, recruiting a third annotator, the last co-author. In particular, the annotators met to review disagreements, deriving a root cause analysis, and assigning a final consensus label.

Root Cause Analysis: During the final label assignment, the annotators examined the sources of disagreement and found that these did not stem from ambiguity in the codebook itself, which is sourced from previous work, but rather from the inherent complexity of the material, which requires expert judgement to interpret. Two recurring patterns illustrate this.

First, determining whether a post constitutes criminal activity (a requirement for assigning a crime_type) or merely borderline behavior (e.g., violating terms of service) often depends on inferring intent from contextual cues such as the headline and the subforum in which the post appears. Annotators drew on their domain expertise to weigh these cues, and this occasionally led them to reach different, yet individually defensible, conclusions. For example, a headline hints that malware is being sold, while the post itself merely requests information about its functionality; whether the surrounding context outweighs the literal content of the post is a matter of expert assessment rather than mechanical rule application.

Second, a similar dynamic arose for posts that discuss products or services without indicating that these are being sold or used. Although such posts are classified as "other" in the absence of evidence of sale or use as per the annotation guidelines, the contextual signals surrounding a post sometimes provided annotators with grounds to infer such evidence, and their expertise led them to weigh these signals differently.

Underlying both patterns is a common root cause: each annotator made a judgement call informed by their individual expertise, and such expert judgement, by its nature, cannot be exhaustively codified. Rather than reflecting a shortcoming of the annotation guidelines, these disagreements capture genuine interpretative variation that is intrinsic to labeling underground forum data, where meaning is often implicit and context dependent.

As a result, for many posts it is difficult to assign a clear set of labels. Although the annotators decided on a final set of labels for the consensus set *C*, in many cases the decision is rather forced than meaningful, prompting us to base the analysis in the main body of the paper on the premise that both $A^1$and A^2 could have been correct. However, for completeness, we next provide details concerning the consensus set *C*.

Results on Subset C. We derive the final consensus label G_i as follows. Where A1 and A2 agree, their shared label is retained. In cases of disagreement, a third annotator (A3) moderated a discussion to resolve the conflict; where no agreement could be reached, A3 breaks the tie by casting the deciding vote. This procedure yields the consensus set $C = \{(i, G_i) : i = 1, \ldots, 150\}$.

Using consensus subset C, we measure the general LLM performance, followed by a nuanced label-level analysis. Table 9 shows the results, where κ_{exact} and α_{exact}measure if a post is entirely correctly labeled, not considering partial agreements. We find that the full agreement overall shows reliable kappa scores, post_type being most reliable with 0.74, followed by author_intent (0.52), supply_chain_category (0.47), and crime_type (0.42).

Table 9
LLM performance compared to the consensus set C.

label_type	P_o	$P\emptyset$	$P_\emptyset$	Cohen κ_{exact}	Krippendorff α_{exact}
post_type	125 (83.3%)	12 (8%)	13 (8.7%)	0.74	0.74
author_intent	100 (66.7%)	19 (12.7%)	31 (20.7%)	0.52	0.52
crime_type	93 (62%)	9 (6%)	48 (32%)	0.42	0.41
supply_chain_category	111 (74%)	0 (0%)	39 (26%)	0.47	0.46
Average	429 (71.5%)	40 (6.7%)	131 (21.8%)	0.53	0.53

Next, Table 10 shows how accurately the LLM assigned labels for each post, summarized across all posts, observing an average F1-score of 0.76 across all codebooks (88–65%). This is only slightly below the F1-score of 0.83 obtained when comparing it with A1 and A2 individually (Table 7), indicating that the two results are broadly comparable. The LLM's reasoning thus largely agrees with the individual annotators, although it deviates somewhat more from the final consensus labels assigned in the second human labeling iteration.

Table 10
LLM performance at the label-level compared to the consensus set C. Overall we observe a positive performance with label-level F1-score of 88–65%.

Codebook	Gpt label count	Human label count	TP	FP	FN	Precision	Recall	F1
post_type	160	159	140	20	19	0.88	0.88	0.88
author_intent	193	178	143	50	35	0.74	0.80	0.77
supply_chain_category	150	150	111	39	39	0.74	0.74	0.74
crime_type	161	155	103	58	52	0.64	0.66	0.65
Total	664	642	497	167	145	0.75	0.77	0.76

References

[1] M. Mischinger, S. Pastrana, G. Suarez-Tangil, Crime VIP: a closed-access underground hacking forum, in: Proceedings of the International AAAI Conference on Web and Social Media (ICWSM), 2026.
[2] S. Pastrana, D.R. Thomas, A. Hutchings, R. Clayton, CrimeBB: enabling cybercrime research on underground forums at scale, in: Proceedings of the 2018 World Wide Web Conference (WWW), 2018, pp. 1845–1854.
[3] G. Atondo Siu, B. Collier, A. Hutchings, Follow the money: the relationship between currency exchange and illicit behaviour in an underground forum, in: 2021 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), 2021, pp. 191–201.
[4] R. Bhalerao, M. Aliapoulios, I. Shumailov, S. Afroz, D. McCoy, Mapping the underground: supervised discovery of cybercrime supply chains, in: 2019 APWG Symposium on Electronic Crime Research (eCrime), IEEE, 2019, pp. 1–16.
[5] A. Caines, S. Pastrana, A. Hutchings, P.J. Buttery, Automatically identifying the function and intent of posts in underground forums, Crime Sci. 7 (1) (2018) 1–14.
[6] M. Mischinger, S. Pastrana, G. Suarez-Tangil, IoC Stalker: early detection of indicators of compromise, 2024 Annual Computer Security Applications Conference (ACSAC), IEEE, 2024.

[7] K. Huang, C. Grier, T.J. Holt, C. Kruegel, D. McCoy, S. Savage, G. Vigna, Framing dependencies introduced by underground commoditization, Workshop on the Economics of Information Security (WEIS), 2015.
[8] M. Mischinger, J. Hughes, F. Vitiugin, S. Pastrana, A. Hutchings, G. Suarez-Tangil, Lost in translation: analyzing non-English cybercrime forums, 2025 APWG Symposium on Electronic Crime Research (eCrime), IEEE, 2025.
[9] S. Pastrana, A. Hutchings, A. Caines, P. Buttery, Characterizing Eve: analyzing cybercrime actors in a large underground forum, in: International Symposium on Research in Attacks, Intrusions, and Defenses (RAID), Springer, 2018, pp. 207–227.
[10] B. Stone-Gross, T. Holz, G. Stringhini, G. Vigna, The underground economy of spam: a botmaster's perspective of coordinating large-scale spam campaigns, 4th USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET), 2011.
[11] B. Dupont, A.-M. Côté, J.-I. Boutin, J. Fernandez, Darkode: recruitment patterns and transactional features of "the most dangerous cybercrime forum in the world, Am. Behav. Sci. 61 (11) (2017) 1219–1243.
[12] M. Campobasso, P. Burda, L. Allodi, CARONTE: Crawling adversarial resources over non-trusted, high-profile environments, in: 2019 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), 2019, pp. 433–442.
[13] K. Turk, S. Pastrana, B. Collier, A tight scrape: Methodological approaches to cybercrime research data collection in adversarial environments, in: 2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), 2020, pp. 428–437.
[14] A. Williams, T. Thrush, D. Kiela, ANLIzing the adversarial natural language inference dataset, in: Proceedings of the Society for Computation in Linguistics (SCiL), 2022, pp. 23–54.
[15] D.R. Thomas, S. Pastrana, A. Hutchings, R. Clayton, A.R. Beresford, Ethical issues in research using datasets of illicit origin, in: Proceedings of the 2017 Internet Measurement Conference (IMC), 2017, pp. 445–462.
[16] V. Valeros, A. Širokova, C. Catania, S. Garcia, Towards better understanding of cybercrime: the role of fine-tuned LLMs in translation, in: 2024 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), IEEE, 2024, pp. 91–99.