Closing Enterprise Network Visibility Gaps Using LLM-Assisted IP Traffic Attribution

Kevin Sam Tharayil[1], Evangelos Froudakis[1], Zane Ma[2], and Angelos D. Keromytis[1]

[1] Georgia Institute of Technology, Atlanta, USA
{kevinsam,efroudakis3,angelos}@gatech.edu
[2] Oregon State University, Corvallis, USA
zane.ma@oregonstate.edu

Abstract. A comprehensive understanding of network traffic is essential for enterprise security teams to effectively manage networks, enforce security controls, investigate incidents, and adhere to increasingly stringent regulatory and compliance requirements. Despite the wide range of monitoring systems and threat/IP intelligence feeds used to automatically assess and label IP traffic, a substantial "long tail" of IP addresses remains unexplained because it lacks domain associations, reputation signals, or sufficient structured metadata. These unattributed IPs create operational blind spots: they may conceal security relevant behavior, complicate incident response, or disrupt policy enforcement. Accurately assessing them requires reasoning over sparse behavioral patterns and heterogeneous contextual signals, a task typically performed manually by analysts through a slow and unscalable process. We propose CRISP (**C**ontextual **R**easoning and **I**P **S**emantic **P**rocessing), a hybrid system that combines traditional IP labeling methods with LLM-based reasoning, invoking the LLM only for ambiguous cases that require analyst attention. CRISP enriches each IP with its observed traffic features and external IP intelligence data, automatically labels high confidence cases using deterministic rules, and uses an LLM workflow to generate auditable assessments and labels for the remaining analyst attributable IPs. Our results show that off-the-shelf general-purpose LLMs, when augmented with domain specific adaptations, can effectively assess and label previously unattributed IP addresses. Compared to manual analyst investigation, CRISP reduces assessment time from minutes to seconds and lowers analysis cost by more than 99% while maintaining high agreement with experienced human analysts. These findings demonstrate that LLM-assisted reasoning offers an operationally feasible approach for reducing long-standing visibility gaps in enterprise traffic analysis and improving enterprise security posture.

Keywords: Enterprise network operations · Threat intelligence · LLM

1 Introduction

Comprehensive visibility into network traffic is essential for enterprise security teams to manage networks, investigate incidents, enforce policies, and satisfy

regulatory and compliance requirements. For instance, security paradigms such as zero trust and default-deny models, while powerful in principle, require precise knowledge of network behavior [8,26]. Without such visibility, applying these models risks breaking critical services or, conversely, lets malicious activity slip through undetected. To support traffic attribution, security teams rely on an array of IP blocklists, reputation services, VirusTotal, AbuseIPDB, commercial threat intelligence feeds, and other metadata driven classifiers. These mechanisms are effective for identifying well known threats and services, but they leave behind a persistent long tail of unattributed IP addresses that lack clear reputation signals, domain associations, or sufficient structured metadata. In this paper we refer to these as ***analyst attributable IPs***: IPs that cannot be labeled confidently by automated systems and therefore require manual analyst reasoning. These IPs may correspond to benign but business-critical services, suspicious infrastructure that does not yet meet threat intelligence thresholds, policy violating applications such as peer-to-peer or anonymization services, or traffic whose purpose remains genuinely ambiguous. In all cases, the lack of attribution creates operational network visibility blind spots complicating incident response, policy enforcement, and security governance.

The core challenge is that these IPs often cannot be classified through simple pattern matching. Their interpretation requires reasoning over sparse behavioral signals, low-level flow features, partial IP intelligence, port usage, ownership information, abuse reports, and weak or conflicting contextual evidence. Today, this task is typically handled manually by experienced security analysts through a slow attribution process that does not scale to the volume of analyst attributable IPs observed in large enterprise networks.

This is where Large Language Models (LLMs) offer a compelling new opportunity. LLMs, trained on diverse textual and semi-structured data, exhibit a remarkable ability to synthesize information, infer context, and reason over heterogeneous, informal, and fuzzy data that often surrounds these hard-to-classify analyst attributable IP addresses. However, a naive and broad application of LLMs directly for all network traffic attribution is unsuitable for operational use as LLMs are known to produce inconsistent outputs, lack transparency, or hallucinate unsupported conclusions [42]. We therefore present *CRISP (**C**ontextual **R**easoning and **IP** **S**emantic **P**rocessing)*, a hybrid system for explainable IP traffic attribution. Taking raw IP traffic in the form of NetFlows as input, CRISP first constructs an enriched IP profile for each IP by combining its observed traffic features with external IP intelligence. It then applies traditional attribution tools and deterministic rules to automatically label high-confidence IPs. For the remaining analyst attributable IPs, CRISP uses a two-stage LLM workflow with domain specific adaptations to generate auditable assessments and labels. This design limits LLM use to cases where contextual reasoning is required while preserving transparency and auditability. This paper investigates two research questions:

– ***RQ1:*** *Can off-the-shelf LLMs assess and label hard-to-attribute IP addresses in a way that is comparable to experienced security analysts?*

– ***RQ2:*** *Can LLM-assisted attribution scale to enterprise operational settings by converting large sets of unresolved IPs into explainable categories that support security triage, compliance review, and follow-up investigation?*

We answer RQ1 through an evaluation using analyst labeled IPs from prior work and a user study with six experienced security analysts. We answer RQ2 through a case study on previously unresolved orphan IPs from prior work. In a random sample of 1,000 external IPs from a large U.S. university network, CRISP automatically labels 76% using deterministic rules, leaving 24% as analyst attributable. On a ground-truth set reviewed by analysts, CRISP achieves high agreement with human analysts while reducing assessment time from minutes to seconds and lowering analysis cost by more than 99%. We further show that CRISP substantially outperforms a baseline that applies LLMs directly to raw NetFlow records, demonstrating the importance of domain specific adaptations and structured prompting.

This paper makes the following contributions:

– We design CRISP, a hybrid IP attribution system that combines deterministic IP intelligence rules with LLM assisted reasoning for ambiguous cases.
– We evaluate CRISP through a user study with experienced security analysts, showing high analyst agreement and substantial reductions in time and cost.
– We demonstrate that CRISP's domain specific adaptations substantially improve accuracy over naive LLM-based labeling of raw NetFlow records.
– As a case study, we apply CRISP to unlabeled IPs from prior work and show how it converts network blind spots into explainable categories useful for security triage, compliance review, and operational decision making.

2 Background and Related Work

Historically, IP traffic labeling has focused on detecting specific behaviors rather than supporting broader network explainability [33,10]. Early approaches relied on pattern recognition algorithms [19], while more recent systems use machine learning and deep learning (ML/DL) models for traffic classification [29,17]. These methods can achieve high accuracy when target classes are well defined, such as known malicious activity, popular services, hosting providers, or cloud platforms, and they underpin many commercial tools currently used for traffic classification. However, they provide limited support for IPs that lack strong intelligence signals or fall outside predefined labels, including the analyst attributable IPs we focus on. These IPs are not always missed instances of known classes; instead, they are traffic destinations for which the available evidence is sparse, ambiguous, and often outside the coverage of existing signatures, labels, or training data. As a result, attributing them requires analyst reasoning and explanation rather than standard supervised classification alone.

The growing utility of general-purpose LLMs, which require less task specific engineering and can operate over heterogeneous inputs, has prompted several

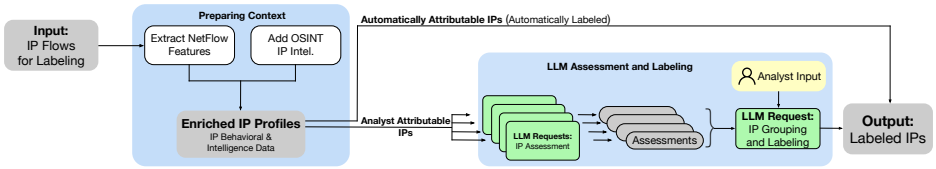


Fig. 1. System design of CRISP. The system uses existing tools and threat-intelligence techniques to automatically label IPs that can be attributed with high confidence, and relies on an LLM-based assessment and labeling pipeline for the remaining analyst attributable IPs.

works to explore their use for traffic classification [43,27,41,38]. LLMs are attractive for this setting because they can synthesize structured, semi-structured, and natural language data. However, they may hallucinate, produce inconsistent outputs, or generate conclusions that are difficult to verify [42], which is especially problematic in security contexts where explanation and accountability matter. They are also expensive and resource intensive to deploy at scale.

Recent studies attempt to mitigate these limitations using fine-tuning pipelines [4], custom tokenization schemes, and other engineering-heavy modifications [16]. Such approaches can improve performance, but they also reintroduce some of the engineering and generalization challenges that affect traditional ML/DL systems. Prior work has also shown that general-purpose LLMs often fail on complex tasks requiring extensive context and structured reasoning [31,6]. Moreover, Zhou et al. found that a fine-tuned LLM, despite achieving high accuracy, performed comparably to or slightly worse than state-of-the-art ML/DL systems on well-known network datasets [43]. These results suggest that LLMs offer limited advantage for straightforward pattern recognition tasks. Their strength instead lies in contextual reasoning over heterogeneous evidence. CRISP therefore adopts a hybrid design: conventional attribution tools and deterministic rules label high confidence IPs, while a constrained LLM workflow is reserved for analyst attributable IPs that require contextual reasoning. CRISP constrains the LLM through well established design principles, structured prompts, and a two-stage assessment and labeling workflow to mitigate hallucinations and maintain auditability and transparency.

3 System Design

As shown in Figure 1, CRISP takes as input a set of external IP addresses observed in outgoing enterprise NetFlow records and produces application level labels using a hybrid design: high confidence IPs are labeled using deterministic rules and existing intelligence sources, while analyst attributable IPs are forwarded to a constrained LLM workflow.

IP Profile Construction For each external IP, CRISP constructs an enriched IP profile from two classes of evidence: traffic features derived from Net-

Table 1. List of features that make up an IP profile that is given as input to CRISP

Feature	Source
Date Detected Protocol Distribution Source IP Distribution Destination (server) port distribution	Flow features obtained from NetFlow data [3]
ASN / IP Ownership Geolocation IP Type(Fixed/Mobile ISP, commercial, education etc.)	IPRegistry [12]
Associated Files [36] Associated Domain Names [37]	VirusTotal [35]
IP Abuse reports from community	AbuseIPDB [2]
Registered/known use of destination port(s)	IANA, Wikipedia, Speedguide.net

Table 2. Rules used to automatically label IPs based on industry practices and subject matter expert knowledge

Label	Rule
Malicious	>=3 reports on VT or IP in BLAG [25]
Potentially Malicious [32]	At least one malicious file or domain association [7]
Known Service	At least one domain in TRANCO top 200K [34]
IP Abuse	Abuse confidence score >=75 on AbuseIPDB [1]
Proxy, TOR, VPN, Bogon, Cloud Provider	Reported on IPRegistry [11]
Parking IP	Zirngibl et. al. [45]

Flow records and external IP intelligence (Table 1). Using subject-matter expertise, we selected features that are commonly used by analysts during manual IP investigation and are available through operationally deployed tools. Traffic features include the date observed, protocol distribution, source IP distribution, and destination port distribution. External intelligence includes ASN and ownership information, geolocation, IP type, associated files and domains from VirusTotal, AbuseIPDB reports, and known uses of observed destination ports. CRISP intentionally excludes source port distributions because, for outgoing flows initiated inside the enterprise network, source ports are typically ephemeral and did not provide useful signal during preliminary testing.

Automatic Attribution CRISP first applies deterministic rules to identify IPs that can be labeled with high confidence (Table 2). These rules use existing security intelligence sources and operational heuristics to identify categories such as malicious IPs, potentially malicious IPs, known services, abuse-reported IPs, proxies, Tor/VPN infrastructure, cloud providers, bogons, and parked IPs. IPs labeled at this stage are treated as automatically attributable and do not require LLM processing. This stage serves two purposes. First, it preserves transparency by using verifiable sources for high confidence cases. Second, it reduces cost and risk by limiting LLM use to IPs that remain ambiguous after conventional attribution.

IPs that remain unlabeled after automatic attribution are treated as analyst attributable and passed to the LLM workflow. Recent work has identified several effective practices for reducing hallucinations and improving reliability when using LLMs for analytical tasks [31,9,22,28]. Following these principles, we designed a two-step LLM workflow that separates assessment from labeling, which improves consistency, transparency, and prompt controllability.

Assessment Stage In the first stage, the LLM receives an enriched IP profile in JSON format and produces a concise 2–3 sentence assessment of the IP's most likely behavior or function, referencing only the provided data. The prompt adopts the persona of a network analyst [15] and constrains the model to rely solely on the supplied intelligence sources and flow features. This reduces opportunities for hallucination and produces an auditable rationale for each assessment [40]. In this stage, CRISP processes one IP at a time because large networks may contain thousands of analyst attributable IPs, far more than can fit in a single LLM context window.

LLM Labeling Stage In the second stage, CRISP sends all the generated assessments to the LLM and asks it to assign consistent application-level labels to each IP based solely on its assessment. Separating assessment from labeling allows the assessment step to serve as task specific context compression [31], stripping away extraneous profile details [21,18] while retaining the key behavioral signals. It also reduces inconsistent or overly granular labels across individual IPs and enables domain experts to curate the labeling prompt, such as by specifying preferred label sets, without modifying the assessment process. The labeling prompt permits ranked labels or an "ambiguous/unclear" label when evidence is insufficient, giving the model an explicit "way out", which has been shown to reduce hallucinations and improve overall accuracy [31]. CRISP uses a zero-shot approach because analyst-attributable IPs, by definition, do not map cleanly to known signatures or predefined categories; if such categories were already known, simpler rule-based or supervised techniques would often suffice. In our implementation, we ask the LLM to assign the application based label that best summarizes the likely purpose of traffic to each IP [14]. The complete prompts used for both stages are provided in Appendix A.

4 System Evaluation

We evaluate CRISP along four dimensions. First, we perform a preliminary validation using analyst labeled orphan IPs from prior work. Second, we measure how many IPs in a random sample of enterprise traffic can be automatically attributed and how many require analyst level reasoning. Third, we conduct a user study with experienced security analysts to evaluate the correctness and quality of CRISP's assessments and labels. Finally, we compare CRISP against a baseline that applies LLMs directly to raw NetFlow records to quantify the impact of CRISP's domain specific design.

Preliminary Validation on Analyst Labeled Orphan IPs Before evaluating CRISP on newly collected enterprise traffic, we first validate its workflow using analyst labeled *orphan IPs* from prior work [32]. Orphan flows are outgoing flows to external IP addresses that lack an associated domain name. The prior study identified 26,385 orphan IPs in a large enterprise network over a seven-month period: 11.6% were automatically labeled, 51.67% were labeled through manual analyst analysis, and the remaining 36.73% were left unresolved.

Table 3. Performance of various LLM models. Top ranked models from various vendors on OpenRouter during the time of writing were selected for evaluation.

Vendor	Model	Release Date	% Accuracy: 30 Orphan IPs	% Accuracy: 30 Ground Truth IPs	Avg. Cost for 30 IPs ($)	Avg. Variation
OpenAI	GPT-4o	May '24	100.0	93.33	0.32	6.67%
	GPT-5	Aug '25	100.0	70.00	1.02	11.67%
	GPT OSS 120b	Aug '25	100.0	86.67	0.01	10.0%
Google	Gemini 2.5 Pro	Jun '25	100.0	80.0	0.99	3.33%
	Gemini 2.5 Flash	Jun '25	100.0	93.33	0.56	3.33%
Meta	Llama 4 Maverick	Apr '25	100.0	73.33	0.03	5.0%
	Llama 4 Scout	Apr '25	100.0	86.67	0.01	5.0%
	Llama 3.3 70b In.	Dec '24	100.0	70.0	0.01	5.0%
Anthropic	Claude Sonnet 4.5	Sep '25	100.0	80.0	0.57	0.0%
xAI	Grok Code Fast 1	Aug '25	100.0	83.33	0.09	13.33%
DeepSeek	V3.1 Terminus	Sep '25	100.0	90.0	0.03	3.33%
	V3 0324	Mar '25	100.0	93.33	0.29	3.33%
Ensemble: Max. Accuracy (GPT-4o, Gemini 2.5 Flash, DeepSeek V3 0324)			100.0	100.0	0.67	3.33%
Ensemble: Best Open source Models (GPT OSS 120b, Llama 4 Scout, DeepSeek V3 0324)			100.0	93.33	0.31	6.67%

We used this analyst labeled portion as an initial validation set. These labels span six application categories: Windows Update Delivery Optimization, Microsoft Teams, BitTorrent, NTP time servers, Apple FaceTime/iMessage, and Blizzard Games. We randomly sampled five IPs per category, yielding 30 IPs, and constructed enriched IP profiles using the features described in Table 1. As shown in Table 3, all evaluated LLMs correctly assessed and labeled these 30 IPs. This result suggests that, when sufficient behavioral and intelligence signals are available, CRISP can accurately recover analyst attributed behavior from enriched IP profiles. We return to the unresolved 36.73% of orphan IPs in Section 5.

4.1 Enterprise Traffic Dataset and User Study

To evaluate CRISP on operational traffic, we collected all outgoing NetFlow records from a large U.S. university network on 1 June 2025. From the set of unique external IP addresses observed in these flows, we randomly sampled 1,000 IPs and constructed enriched IP profiles for each. Applying the deterministic rules from Table 2, CRISP automatically labeled 76% of the sampled IPs. The remaining 24% lacked sufficient high confidence evidence and were treated as analyst attributable IPs. We processed these 240 analyst attributable IPs using CRISP's assessment and labeling workflow with OpenAI's GPT-4o. Manual inspection of the generated outputs showed that, despite the large number of analyst attributable IPs, many exhibited recurring behavioral patterns that could be grouped into a smaller set of high-level application or usage categories.

To evaluate the correctness of the LLM generated outputs and to create a ground truth dataset for further analysis, we conducted a user study among security analysts with varying levels of experience. Since analyst attributable IPs are inherently difficult to label, particularly in cases when IP intelligence data is sparse, we prioritized obtaining multiple analyst responses per IP rather than collecting a larger set with only one analyst response per IP. This approach

reduces ambiguity in the ground truth and helps resolve potential disagreements an analyst might have with the LLM's response. Consequently, based on the number of security analysts who agreed to participate in the study and the number of expert hours they were willing to contribute, we limited the user study, and subsequently the ground truth dataset, to 30 randomly selected IPs from the 240 analyst attributable IPs.

Participants completed the study through a Flask web application. The study had two parts. In the first part, analysts were shown the enriched IP profile together with CRISP's generated assessment and label. They rated the assessment and label using four Likert-scale questions: assessment comprehensibility, assessment correctness, label alignment with the assessment, and label correctness. The full questionnaire and response options are given in Appendix B. If an analyst assigned a score of 3 or below to any question, they were asked to provide their own assessment, label, and explanation. In the second part, analysts independently assessed and labeled two additional IP profiles without seeing CRISP's output, allowing us to measure the time required for manual attribution. Six analysts participated in total and were divided into two groups of three. The 30 IPs were randomly divided into two sets and the three users in each group evaluated the same set of IPs. The study was reviewed and approved by our Institutional Review Board.

Table 4 summarizes the results. Across all analysts, CRISP received mean scores of 4.6 for assessment comprehensibility, 4.3 for assessment correctness, 4.7 for label alignment, and 4.3 for label correctness. These results indicate strong agreement between CRISP and experienced human analysts. Disagreements were limited: among the four IPs that received a score of 2 for any question, the remaining analysts in the same group agreed with CRISP's output. The disagreements primarily reflected a preference for an "unknown" label or more conservative phrasing in the assessment when available evidence was weak.

Scalability The second part of the user study illustrates the scalability challenge of the manual attribution task. Analysts required an average of under 5 minutes to assess and label one IP. In contrast, GPT-4o processed all 30 IPs in 2 minutes and 20 seconds, averaging 4.7 seconds per IP, at a total cost of $0.32. Compared with manual analyst investigation, CRISP reduces per IP assessment time from minutes to seconds and lowers the estimated direct cost of generating assessments and labels by more than 99% [3], while producing assessments and labels that largely agree with experienced analysts.

4.2 Performance Across LLMs

We next evaluate how CRISP performs across different LLMs. Using the 30 IP ground truth set from the analyst study, we tested 12 models from multiple vendors through the OpenRouter API [23]. An output was considered correct only if both the assessment and the label were correct. We also estimated cost

[3] Based on the average salary of security analysts in the US reported by recruiting platforms at the time of writing [44].

Table 4. User study setup. Participants were divided into two groups, with all members of a group receiving the same data. In Part 1, they evaluated the LLM's assessments and labels; in Part 2, they produced their own assessments and labels based on the IP profiles while their completion time was measured.

Group	Analyst ID	Years of Experience	Mean Scores for LLM Assessment		Mean Scores for LLM Label		Avg. time to assess & label 1 IP
			Comprehensibility	Correctness	Alignment	Correctness	
Group 1	analyst1	9	4.1	3.7	4.5	4.0	475s
	analyst2	4	4.4	4.1	4.5	4.0	126s
	analyst3	2.5	4.6	4.3	4.2	4.4	397s
Group 2	analyst4	9	4.8	4.2	4.9	4.2	565s
	analyst5	7	5.0	4.9	4.9	4.4	90s
	analyst6	4	4.8	4.6	5.0	4.6	87s
Mean			4.6	4.3	4.7	4.3	290s

using token counts reported by the API and each model's published price per-million-tokens.

As shown in Table 3, model performance varied substantially. GPT-4o, Gemini 2.5 Flash, and DeepSeek V3 achieved 93.33% accuracy, while some newer or larger models performed worse. Errors were concentrated in IPs with sparse evidence, particularly those involving UDP flows to registered ports without clear application mappings or additional intelligence. Lower performing models often hallucinated in these cases, for example, by associating applications like BitTorrent with unknown ports, or by speculating that the traffic might be peer-to-peer, gaming related, or even malware C2. Although such interpretations are possible, the provided IP profiles do not contain evidence to support them. As such, preferred labels would be "non-standard or custom service" or "ambiguous/unclear", which was also confirmed by analysts in the user study. Even though these fallback options were explicitly offered in the prompt, several models appeared reluctant to use these "way out" labels consistently. Additional misclassifications arose for IPs with traffic to multiple ports with known uses, as well as for IPs for which more than one label was reasonably plausible. In these cases, some models either "tried too hard" or overlooked clear evidence in the IP profile in an effort to force a single label, even though the prompt explicitly allowed returning multiple ranked labels when warranted.

Variation Determinism is important for an IP labeling system. Similar or identical inputs should consistently produce the same output. Since LLMs are known to have issues with consistency, we also measured output variation by running each model twice on both the orphan-IP validation set and the analyst-study ground-truth set. We define variation as the fraction of IPs for which the same model produced different labels across the two runs. Models with higher variation were often the same models that speculated more or hallucinated details not present in the input IP profile. In addition, newer architectures such as GPT-5, which does not support a `temperature` parameter, may also contribute to this increased variability [30].

Ensemble One of the reported best practices for building effective LLM agents for complex tasks is an orchestrator-worker architecture [28]. Motivated

by this, we created two ensemble configurations where multiple LLMs independently assess and label each IP, and a synthesizer assigns the final label via majority vote. Table 3 shows the results for two such ensembles. The best performing ensemble achieved 100% accuracy on both datasets with low variation, while remaining cheaper than using a single large model such as GPT-5.

4.3 Impact of Domain Specific Adaptations

To evaluate the contribution of CRISP's domain specific adaptations, we compared it with a baseline that applies an LLM directly to raw NetFlow records. The baseline uses a single prompt to assess and label each IP from its raw flows, without enriched IP profiles or the two stage assessment and labeling workflow. Table 5 shows that CRISP substantially outperforms this baseline across all evaluated models. For example, GPT-4o improves from 13.33% accuracy with the baseline to 93.33% with CRISP. Similar gains appear across DeepSeek and Llama models. Baseline outputs often relied on shallow port based inference, such as labeling port 443 traffic as HTTPS, and frequently hallucinated unsupported applications or inconsistent labels for IPs with similar profiles. These results demonstrate the benefit of CRISP's combined domain specific adaptations, including enriched IP profiles, constrained prompts, and separate assessment and labeling stages, over directly applying an LLM to raw network traffic. We also compare model pairs from the same vendors released approximately one year apart. CRISP outperforms the baseline in each case, suggesting that domain specific task structure remains important even as underlying LLMs evolve over time.

Vendor	Model	Release Date	% Accuracy Baseline	% Accuracy CRISP
OpenAI	GPT-4o	May '24	13.33	93.33
	GPT-3 Turbo	May '23	20.0	63.33
DeepSeek	V3.1 Terminus	Sep '25	36.67	90.0
	DeepSeek R1	Jan '25	10.0	86.67
Meta	Llama 4 Scout	Apr '25	10.0	86.67
	Llama 3 70b In.	Apr '24	13.33	76.67

Table 5. CRISP outperforms a baseline system that naively uses an LLM to assess and label IPs, even as the underlying LLMs improve.

5 Case Study: Explaining Previously Unresolved Orphan IPs

To demonstrate how CRISP can reduce operational blind spots, we apply it to the unresolved orphan IPs from prior work [32]. As described in Section 4, this orphan IP dataset contains outgoing traffic to external IPs with no associated domain name. Although 63.27% of the 26,385 orphan IPs were labeled through automated tools or manual analyst effort, the remaining 9,651 IPs were left unresolved because they required case-by-case reasoning that did not scale. We treated these unresolved IPs as analyst attributable and constructed enriched IP profiles using the same traffic features and intelligence sources described in

Table 1. CRISP labeled a majority of these IPs (67.4%) as traffic associated with benign services such as web services, media streaming, CDNs, cloud hosting, and gaming. Approximately 5% of IPs were labeled as security relevant, including scanning activity, brute force attempts, potential DDoS activity, trojans, or botnet related traffic. After manually reviewing the profiles of these IPs, we found that although the available intelligence signals did not meet the threshold required for immediate action, the LLM was able to connect observed flow behavior with suspicious patterns reported by the security community or with historically abused ports. These cases highlight how CRISP can surface IPs that warrant further investigation, even when traditional indicators are inconclusive.

Approximately 11% of IPs were labeled as services such as database operations, network configuration and management, education/research, government or military-related infrastructure, security services, file servers, and video conferencing platforms. For such IPs, naively applying restrictive policies such as default deny could disrupt critical organizational functions. CRISP also identified 23 IPs associated with peer-to-peer traffic, BitTorrent, or Tor. As regulatory and compliance requirements increasingly mandate strict control over such traffic [5,24,20], the presence of these services within network blind spots presents a compliance risk, potentially exposing organizations to violations without security teams being aware of it. Finally, 15.8% of IPs remained labeled as "non-standard or custom service" or "ambiguous/unclear". These labels do not provide complete attribution, but they still reduce uncertainty by separating genuinely ambiguous traffic from benign services, suspicious activity, and policy relevant applications. Overall, this case study shows that CRISP can reduce enterprise network visibility blind spots at operational scale by transforming unresolved IPs into explainable categories that support security triage, compliance review, and follow-up investigation.

6 Discussion, Limitations and Future Work

Our evaluation shows that CRISP can produce assessments and labels that largely agree with experienced security analysts, while the case study demonstrates that the system can scale to unresolved IPs observed in an operational enterprise network. While CRISP substantially reduces the size of network blind spots, some IPs remain difficult to interpret and are appropriately labeled as custom service or ambiguous. Although such labels may offer limited immediate insight, they play an important role by explicitly identifying cases that require additional data collection, policy review, or targeted investigation. Due to the limited number of experienced network analysts who agreed to participate in the study, and the expert hours they were able to contribute, the ground truth dataset was limited to 30 IPs. Furthermore, analysts evaluated CRISP's outputs before providing alternative labels for lower-scored IPs, which may introduce anchoring bias into the resulting ground truth. However, each IP was independently reviewed by multiple analysts to reduce reliance on individual judgments and improve label reliability, prioritizing annotation quality over dataset size. Addi-

tionally, the rule set used for automatic labeling in CRISP is limited. Rather than attempting exhaustive coverage, we focused on industry best practices derived from subject-matter expertise and prioritized high precision over high recall. Future work could expand CRISP through the integration of dynamic data sources via MCP servers [39], retrieval augmented generation (RAG) [13], and agentic architectures that query live intelligence sources on demand. Future work also includes evaluating safeguards for practical deployment of CRISP, including input sanitization to mitigate prompt injection attempts and the use of locally hosted models where confidentiality requirements prohibit external processing.

7 Conclusion

Enterprise networks contain a persistent long tail of IP traffic that cannot be confidently attributed using conventional tools alone. These analyst attributable IPs create visibility gaps that complicate security incident response, policy enforcement, compliance review, and monitoring. We presented CRISP, a hybrid system that combines deterministic IP-intelligence rules with constrained LLM-assisted reasoning to generate explainable assessments and application-level labels for such IPs. CRISP achieves high agreement with experienced analysts while reducing assessment time from minutes to seconds and lowering analysis cost by more than 99%. Compared with applying LLMs directly to raw NetFlow records, CRISP substantially improves accuracy and reduces unsupported conclusions. Collectively, these findings demonstrate that LLM-assisted analysis can make large scale attribution of enterprise network traffic operationally feasible, enabling security teams to substantially reduce long standing network visibility gaps.

Acknowledgments. We thank the anonymous reviewers for their insightful comments and suggestions. This work is supported by National Science Foundation SaTC CORE grant no. 2247562. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Appendix A: LLM Prompts Used for CRISP and Baseline System

```
crisp_assessment_system_prompt = """You are a network telemetry analyst. I will give you
JSON metadata about a single IP. Your job is to infer the most likely PURPOSE of the traffic
and briefly explain why.

Requirements:
- Only use the fields provided in the JSON.
- Prefer precise usage labels over generic ones.
- If the data is ambiguous, pick the most likely usage and list 1-2 alternatives.
- Be concise and deterministic.
```

```
Consider (when present): destination_port_info (IANA/ISC notes),
destination_port_distribution, source/destination ports, reverse DNS/hostnames,
ASN/organization, VT domain/file notes, IP registry "type" (isp/cloud), and common service
mappings.

Output strictly as JSON with this schema:
{
    "ip": "<IP address>",
    "Analysis": "<2-4 sentences citing the strongest fields/signals from the JSON>",
    "alternatives": [{"usage_label":"<alt>", "why":"<one sentence>"}]
}"""
```

```
crisp_labeling_system_prompt = """You are a network telemetry analyst. You will be given a
list of IP addresses and their expert assessments. Each assessment describes the observed
behavior of the IP, based on traffic patterns, port usage, reverse DNS, ASN, WHOIS data, and
other metadata. Your task is to assign a short, application-based label to each IP address
that best summarizes the likely PURPOSE of its traffic. The goal of labeling is to help
analysts group IP addresses by their usage or role on the network.

Requirements:
- Label each IP address only based on the assessment of each IP address.
- After labeling an analyst should be able to group the IPs based on the labels.
- Use consistent phrasing across similar cases (be deterministic).
- If the assessment suggests that the traffic could be from a custom, specialized,
proprietary service or non-standard service, label it as "non-standard or custom service".
- Based on the assessment, when necessary, you can assign a maximum of two ranked labels
separated by  "/" if the assessment suggest more than one possible usage for the IP address.
The first label should be the most likely usage and the second label should be the less
likely usage.
- When absolutely necessary, if the assessment is too ambiguous, have too many possible
interpretations, or lacks enough information to confidently assign a specific label the IP
as "ambiguous/unclear".

Output strictly as JSON with this schema:
{
    "<IP address>":"<assigned label>",
    "<IP address>":"<assigned label>",
    ...
}"""
```

Appendix B: User Study Questionnaire and Scoring Scale

LLM Assessment Evaluation Question	Scoring Scale
How comprehensible is the system's assessment?	1: Confusing or incomprehensible 2: Vague and hard to follow 3: Somewhat clear but needs improvement 4: Mostly clear and easy to understand 5: Very clear, concise, and fully understandable
How correct is the system's assessment?	1: Completely inaccurate 2: Mostly inaccurate 3: Moderately accurate / mixed accuracy 4: Mostly accurate 5: Completely accurate and reliable

LLM Labeling Evaluation Questions	Scoring Scale
How well does the system's label align with the assessment?	1: Not aligned at all 2: Poorly aligned 3: Partially aligned 4: Mostly aligned 5: Fully aligned and strongly supported by the assessment
How correct is the system's label for this IP?	1: Completely incorrect 2: Mostly incorrect 3: Partially correct 4: Mostly correct 5: Completely correct and appropriate

References

1. AbuseIPDB: Abuseipdb & opencti integration (2025), https://www.abuseipdb.com/opencti, accessed: 2025-12-12
2. AbuseIPDB: Abuseipdb api documentation — introduction (2025), https://docs.abuseipdb.com/#introduction, accessed: 2025-12-12
3. Claise, B.: Cisco systems netflow services export version 9. Tech. rep. (2004)
4. Cui, T., Lin, X., Li, S., Chen, M., Yin, Q., Li, Q., Xu, K.: Trafficllm: Enhancing large language models for network traffic analysis with generic traffic representation. arXiv preprint arXiv:2504.04222 (2025)
5. Cybersecurity and Infrastructure Security Agency (CISA) and Federal Bureau of Investigation (FBI): Aa20-183a: Defending against malicious cyber activity originating from tor (Jul 2020), https://www.cisa.gov/news-events/cybersecurity-advisories/aa20-183a, accessed: 2025-12-12
6. Deng, G., Liu, Y., Mayoral-Vilches, V., Liu, P., Li, Y., Xu, Y., Zhang, T., Liu, Y., Pinzger, M., Rass, S.: {PentestGPT}: Evaluating and harnessing large language models for automated penetration testing. In: 33rd USENIX Security Symposium (USENIX Security 24). pp. 847–864 (2024)
7. Faulkenberry, A., Avgetidis, A., Ma, Z., Alrawi, O., Lever, C., Kintis, P., Monrose, F., Keromytis, A.D., Antonakakis, M.: View from above: Exploring the malware ecosystem from the upper dns hierarchy. In: Proceedings of the 38th Annual Computer Security Applications Conference. pp. 240–250 (2022)
8. FireMon: Why zero trust fails in the real world and what you can do about it (Jun 2025), https://www.firemon.com/blog/why-zero-trust-fails-in-the-real-world-and-what-you-can-do-about-it/, accessed: 2025-11-28
9. Google Project Zero: Project naptime: Evaluating offensive security capabilities of large language models (Jun 2024), https://googleprojectzero.blogspot.com/2024/06/project-naptime.html, accessed: 2025-12-12
10. Guerra, J.L., Catania, C., Veas, E.: Datasets are not enough: Challenges in labeling network traffic. Computers & Security **120**, 102810 (2022)
11. Ipregistry: Ip address response fields — ipregistry api documentation (2025), https://ipregistry.co/docs/responses/ip-address-fields, accessed: 2025-12-12
12. Ipregistry: Ipregistry: Ip geolocation and threat intelligence api (2025), https://ipregistry.co/, accessed: 2025-12-12
13. Jiang, Z., Xu, F.F., Gao, L., Sun, Z., Liu, Q., Dwivedi-Yu, J., Yang, Y., Callan, J., Neubig, G.: Active retrieval augmented generation. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 7969–7992 (2023)

14. Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y.: Large language models are zero-shot reasoners. Advances in neural information processing systems **35**, 22199–22213 (2022)
15. Kong, A., Zhao, S., Chen, H., Li, Q., Qin, Y., Sun, R., Zhou, X., Wang, E., Dong, X.: Better zero-shot reasoning with role-play prompting. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 4099–4113 (2024)
16. Li, Q., Zhang, Y., Jia, Z., Hu, Y., Zhang, L., Zhang, J., Xu, Y., Cui, Y., Guo, Z., Zhang, X.: Dollm: How large language models understanding network flow data to detect carpet bombing ddos. arXiv preprint arXiv:2405.07638 (2024)
17. Lin, X., Xiong, G., Gou, G., Li, Z., Shi, J., Yu, J.: Et-bert: A contextualized datagram representation with pre-training transformers for encrypted traffic classification. In: Proceedings of the ACM Web Conference 2022. pp. 633–642 (2022)
18. Liu, P., Liu, J., Fu, L., Lu, K., Xia, Y., Zhang, X., Chen, W., Weng, H., Ji, S., Wang, W.: Exploring {ChatGPT's} capabilities on vulnerability management. In: 33rd USENIX Security Symposium (USENIX Security 24). pp. 811–828 (2024)
19. Madhukar, A., Williamson, C.: A longitudinal study of p2p traffic classification. In: 14th IEEE international symposium on modeling, analysis, and simulation. pp. 179–188. IEEE (2006)
20. Mintz, D.G.: Peer-to-peer (p2p) file sharing (Jul 2007), https://www.transportation.gov/testimony/peer-peer-p2p-file-sharing, testimony before the Committee on Oversight and Government Reform; Accessed: 2025-12-12
21. Mirzadeh, I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S., Farajtabar, M.: Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. arXiv preprint arXiv:2410.05229 (2024)
22. OpenAI: Optimizing llm accuracy (2025), https://platform.openai.com/docs/guides, accessed: 2025-12-12
23. OpenRouter: Quickstart (2025), https://openrouter.ai/docs/quickstart, accessed: 2025-12-12
24. Palo Alto Networks, Inc.: Controlling peer-to-peer applications. Tech. rep., Palo Alto Networks, Sunnyvale, CA (2010), https://media.paloaltonetworks.com/documents/Controlling_P2P_Apps.pdf, accessed: 2025-12-12
25. Ramanathan, S., Mirkovic, J., Yu, M.: Blag: Improving the accuracy of blacklists. In: NDSS (2020)
26. Regev, O.: Why you can't trust zero trust solutions. Cyber Protection Magazine (Sep 2025), https://cyberprotection-magazine.com/why-you-cant-trust-zero-trust-solutions, accessed: 2025-11-28
27. Sarabi, A., Yin, T., Liu, M.: An llm-based framework for fingerprinting internet-connected devices. In: Proceedings of the 2023 ACM on Internet Measurement Conference. pp. 478–484 (2023)
28. Schluntz, E., Zhang, B.: Building effective agents (Dec 2024), https://www.anthropic.com/engineering/building-effective-agents, accessed: 2025-12-12
29. Shi, Z., Luktarhan, N., Song, Y., Tian, G.: Bfcn: A novel classification method of encrypted traffic based on bert and cnn. Electronics **12**(3), 516 (2023)
30. Swept AI: Gpt-5 removed the one thing digital health & the enterprise ai needs (Oct 2025), https://www.swept.ai/post/gpt-5-removed-the-one-thing-digital-health-the-enterprise-ai-needs, accessed: 2025-12-12
31. Sypherd, C., Belle, V.: Practical considerations for agentic llm systems. arXiv preprint arXiv:2412.04093 (2024)

32. Tharayil, K.S., Avgetidis, A., Ma, Z., Antonakakis, M., Keromytis, A.D.: Understanding orphan flows. In: 2025 9th Network Traffic Measurement and Analysis Conference (TMA). pp. 1–11. IEEE (2025)
33. Torres, L.M., Magana, E., Izal, M., Morato, D.: Strategies for automatic labelling of web traffic traces. In: 37th Annual IEEE Conference on Local Computer Networks. pp. 196–199. IEEE (2012)
34. Tranco: Tranco: A research-oriented top sites ranking hardened against manipulation (2025), https://tranco-list.eu/, accessed: 2025-12-12
35. VirusTotal: How it works (2025), https://docs.virustotal.com/docs/how-it-works, accessed: 2025-12-12
36. VirusTotal: Ip object — communicating files (2025), https://docs.virustotal.com/reference/ip-object-communicating-files, accessed: 2025-12-12
37. VirusTotal: Ip object — resolutions (2025), https://docs.virustotal.com/reference/ip-object-resolutions, accessed: 2025-12-12
38. Wang, T., Xie, X., Zhang, L., Wang, C., Zhang, L., Cui, Y.: Shieldgpt: An llm-based framework for ddos mitigation. In: Proceedings of the 8th asia-pacific workshop on networking. pp. 108–114 (2024)
39. Wang, Z., Chang, Q., Patel, H., Biju, S., Wu, C.E., Liu, Q., Ding, A., Rezazadeh, A., Shah, A., Bao, Y., et al.: Mcp-bench: Benchmarking tool-using llm agents with complex real-world tasks via mcp servers. arXiv preprint arXiv:2508.20453 (2025)
40. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)
41. Zhang, H., Sediq, A.B., Afana, A., Erol-Kantarci, M.: Large language models in wireless application design: In-context learning-enhanced automatic network intrusion detection. In: GLOBECOM 2024-2024 IEEE Global Communications Conference. pp. 2479–2484. IEEE (2024)
42. Zhang, X., Meng, H., Li, Q., Tan, Y., Zhang, L.: Large language models powered malicious traffic detection: Architecture, opportunities and case study. arXiv preprint arXiv:2503.18487 (2025)
43. Zhou, H., Huang, X., Deng, L.: Enhancing network traffic classification with large language models. In: 2024 IEEE International Conference on Big Data (BigData). pp. 7282–7291. IEEE (2024)
44. ZipRecruiter: Network analyst salary. https://www.ziprecruiter.com/Salaries/Network-Analyst-Salary (2026), accessed: 2026-05-21
45. Zirngibl, J., Deusch, S., Sattler, P., Aulbach, J., Carle, G., Jonker, M.: Domain parking: Largely present, rarely considered! In: TMA (2022)