Preliminary Insights in the Language of Extortion in Ransomware Dialogues

Federico Cerutti[1,2,3,4,*], Massimiliano Giacomin[1], Andrea Loreggia[1] and Enrico Scala[1]

[1]University of Brescia, Italy
[2]Imperial College London, United Kingdom
[3]Cardiff University, United Kingdom
[4]University of Southampton, United Kingdom

Abstract

Ransomware negotiation constitutes a high-stakes form of digitally mediated interaction in which interlocutors pursue incompatible aims through a constrained textual channel. Existing scholarship has thoroughly documented the technical mechanics of ransomware, but the communicative processes through which attackers and victims construct pressure, credibility, and cooperation remain underexamined, despite their potential to have a massive impact on victims' economies. This study applies speech act theory and argumentation analysis to a corpus of authentic negotiation transcripts drawn from an open-source dataset. Using a large-language-model-assisted annotation with manual validation, each message is classified by illocutionary type and argumentative function. The analysis shows that attackers organise their discourse through directives, evaluative assertions, and conditional threats that project control and simulate professionalism, while victims rely on justificatory reasoning, requests, and appeals to fairness to reframe coercion as bargaining. By charting these pragmatic and argumentative routines across the unfolding of interaction, the study demonstrates that extortionary communication operates through stable patterns rather than ad-hoc improvisation, thereby extending pragmatic theory into a domain of digital conflict and providing an empirical foundation for examining the communicative dynamics that sustain cybercrime interaction.

1. Introduction

Ransomware negotiation represents one of the most complex forms of mediated interaction in contemporary cybersecurity. It combines high-stakes decision-making, asymmetric power relations, and intense emotional pressure within a constrained textual channel. In these exchanges, language is the only medium through which attackers and victims pursue incompatible aims: extortion versus recovery. Understanding how interlocutors use linguistic resources to threaten, justify, persuade, and cooperate is therefore essential for explaining how digital extortion unfolds in practice. Threat actors frame legitimacy, maintain pressure, and manage uncertainty through patterned discourse, whereas victims attempt to delay, reframe, or soften demands. Our empirical findings, drawn from the publicly available and anonymised **Ransomchats** dataset,[1] confirm that these exchanges follow a structured temporal logic: early turns are dominated by directives, disclosures, and metacommunicative work; mid-phase interactions centre on evaluative, justificatory, and conditional statements; and closing phases return to factual and procedural updates. This organisation suggests that negotiation language is neither ad-hoc nor chaotic but unfolds through recognisable pragmatic and argumentative routines.

Only a small number of research outputs have empirically analysed ransomware negotiations. Fujima et al. [1] use large language models to examine linguistic and strategic patterns in ransomware messages, but without a clear association to speech act theory, instead arguing for closer integration of computational linguistic analysis into cybersecurity workflows. Zhu et al. [2] propose a game-theoretic

Joint National Conference on Cybersecurity (ITASEC & SERICS 2026) February 09-13, 2026, Cagliari, IT
*Corresponding author.
federico.cerutti@unibs.it (F. Cerutti); massimiliano.giacomin@unibs.it (M. Giacomin); andrea.loreggia@unibs.it (A. Loreggia); enrico.scala@unibs.it (E. Scala)
0000-0003-0755-0358 (F. Cerutti); 0000-0003-4771-4265 (M. Giacomin); 0000-0002-9846-0157 (A. Loreggia); 0000-0003-2274-875X (E. Scala)

[1]https://github.com/Casualtek/Ransomchats (on 27 November 2025), MIT License.

framework supported by ChatGPT-generated insights to model decision-making behaviour in ransomware extortion scenarios. Faivre [3], from a psychological and social-engineering perspective, analyses asymmetric power dynamics, fear appeals, and coercive communication strategies that shape victims' behaviour. These studies demonstrate growing interest in the communicative dimension of ransomware, yet they do not provide a systematic account of the pragmatic and argumentative structures underlying negotiation moves over time.

The present study addresses this gap by integrating speech act theory [4, 5, 6] and argumentation analysis [7] into the examination of ransomware communication. Illocutionary categories capture how interlocutors assert, direct, commit, and evaluate, while argumentative functions identify the justificatory, coercive, evidential, or procedural roles of each message. The observed temporal patterns reveal how coercive framing appears early in interaction, how justificatory and fairness-oriented reasoning dominates the bargaining core, and how factual and procedural grounding resurfaces as dialogues approach closure. These observations show that extortionary communication, despite its adversarial setting, is structured by recurrent forms of pragmatic and argumentative organisation.

This paper's contributions are twofold. It first extends pragmatic theory into a domain characterised by high pressure, conflict, and limited contextual cues, showing that even under coercive conditions illocutionary and argumentative work remains systematic. Secondly, it offers operational insight for cybersecurity professionals, by supporting the identification of communicative indicators associated with escalation, stabilisation, or closure, as well as potentially group attribution.

2. Background

2.1. Ransomware Groups

To situate the negotiation transcripts within their operational contexts, this section introduces the principal ransomware groups represented in the dataset, with detailed profiles and OSINT references deferred to A. The ecosystem these groups constitute is fluid and internally heterogeneous, yet several common patterns emerge. Most operate through ransomware-as-a-service arrangements that pair data theft with encryption, rely on affiliate networks for intrusion activity, and maintain leak sites to exert bargaining pressure during negotiations. Many draw on shared tooling, inherited codebases, or infrastructural overlap, producing lineages in which one brand succeeds, absorbs, or imitates another. This recurrent reuse of technical and organisational components creates recognisable family structures across the ecosystem.

Despite these shared characteristics, the groups also differ in ways relevant to understanding their negotiation behaviour. Some, such as Conti, LockBit 3.0, and Hive, maintained extensive affiliate programmes and accumulated large victim portfolios before substantial law-enforcement intervention. Others, including DarkSide, BlackMatter, and BlackBasta, illustrate iterative rebranding and adaptation following operational setbacks or public exposure. A further category comprises derivative or successor groups—such as NoEscape, Hunters International, Cloak, and Pear—that emerged from leaked source code or from the dissolution of earlier brands. These groups often retain the technical signatures of their predecessors while experimenting with revised business models, including data-extortion-only variants.

There is also variation in scale and maturity. Well-established syndicates, including Conti and LockBit 3.0, exhibited high operational capacity and a quasi-corporate internal structure, whereas newer actors such as Fog, RunSomeWares, and Trinity remain comparatively small and less technically differentiated. Some groups, such as DragonForce, incorporate ideological messaging alongside financially motivated campaigns, illustrating how hybrid motivations can coexist within the broader landscape. Others, such as Mallox and Ranzy, persist as mid-tier operations exploiting widely accessible attack surfaces such as exposed SQL servers or weak VPN configurations.

Across these differences, a common reliance on multi-extortion strategies and rapid brand evolution underscores the unstable environment in which negotiations take place. The similarities allow for certain generalisations about extortion practices, while the divergences—particularly those relating to

scale, lineage, and operational maturity—provide necessary context for interpreting the rhetorical and strategic patterns observed in the negotiation transcripts.

2.2. Speech Acts and Communicative Action

Speech act theory originates in the philosophy of language, seeking to explain how utterances do not merely convey information but perform actions. In his seminal Harvard lectures, Austin [4] proposed that saying something is often equivalent to *doing* something: for instance, when a speaker makes a promise, issues a warning, or declares a marriage, the utterance itself constitutes an act. He distinguished between the *locutionary act* (the production of a meaningful utterance), the *illocutionary act* (the action performed in saying something), and the *perlocutionary act* (the effect produced on the listener).

Building on Austin, Searle [5, 6] systematized illocutionary acts into five broad classes: *assertives* (statements that commit the speaker to the truth of a proposition), *directives* (attempts to get the hearer to do something), *commissives* (commitments to a future action such as promising or threatening), *expressives* (expressions of psychological states), and *declarations* (utterances that bring about changes in the social world). These categories have provided a durable framework for analysing both everyday and institutional communication [8, 9].

Later pragmatic theories refined this model by focusing on inference, context, and social interaction. Leech [10] emphasised the role of politeness and indirectness, while [11] and [12] explored the relationship between illocutionary force and social roles. Speech act analysis has since evolved into a central tool of discourse studies, allowing researchers to link linguistic form with interpersonal goals, power dynamics, and ideological positioning [13, 7]. In negotiation contexts, they provide the microfoundations for interactional strategies such as offering, counter-offering, justifying, and conceding [14].

2.3. Application to Ransomware Negotiation

Ransomware negotiations constitute an extreme form of goal-oriented dialogue where the traditional norms of institutional communication intersects with criminal coercion. The attacker must simultaneously demonstrate technical control (through *assertives*), exert pressure (*directives*), and project credibility (*commissives*), while the victim engages in information-seeking, bargaining, and moral appeals to mitigate damage and reframe the situation as a solvable transaction. Each utterance thus performs multiple pragmatic functions: coercion, persuasion, and face-work.

To study these dynamics empirically, we draw on the **Ransomchats** corpus [15], a publicly available and anonymised dataset of real ransomware negotiation transcripts collected from dark-web leak sites and incident response disclosures. The corpus is distributed under the MIT license and accessible via GitHub at https://github.com/Casualtek/Ransomchats. These dialogues are particularly valuable because they are unmediated records of asynchronous interaction, preserving both the linguistic texture and the pragmatic sequencing of cyber-extortion discourse. The dataset's open license allows reproducibility and transparent methodological reporting, facilitating comparative research across threat actor groups and negotiation styles.

3. Methodology

Our approach integrates linguistic pragmatics with argumentation theory to provide a theoretically grounded account of the communicative dynamics of ransomware negotiations. Classical speech act theory offers the foundational justification for treating attacker–victim exchanges as sequences of communicative actions. Austin's distinction between locutionary, illocutionary, and perlocutionary dimensions [4] and Searle's taxonomy of illocutionary forces [5, 6] establish the conceptual apparatus required to analyse how threats, demands, assurances, or concessions are performed linguistically in these dialogues.

However, conventional taxonomies alone cannot account for the strategic and coercive nature of digital extortion. Ransomware communication unfolds under conditions of anonymity, pronounced power asymmetry, and time pressure, in which coercion and cooperation co-exist. Pragmatic work emphasising the role of contextual inference in interpreting force [8, 9] justifies the need to adapt classical categories to these interactional constraints. This adaptation allows us to capture how interlocutors modulate directness, imply consequences, or frame threats while maintaining plausible deniability.

Argumentation theory provides a further theoretical justification by modelling how interactants use reasons, evaluations, or principles to influence counterparts. The pragma-dialectical framework [7] would then argue that even adversarial exchanges rely on structured justificatory moves. Work on negotiation and conflict discourse [16, 13] demonstrates that actors invoke fairness, legitimacy, responsibility, and consequences to manage disagreement and pursue settlement. This supports the integration of argumentative functions into our analytical scheme.

To operationalise our theoretical framework, we employ a two-layer annotation scheme. This dual-layer structure is necessary because ransomware negotiation combines performative and justificatory dimensions. The primary speech act captures what the utterance *does*, while the argumentative function captures how it *advances* the negotiation. Together, these layers allow us to represent both the pragmatic action and the strategic rationale embedded in each communicative move. The first layer assigns a single *primary speech act* to each message, capturing the illocutionary force of the utterance. The category Informative–Declarative denotes statements that provide information or establish a position, as in "Your files were encrypted at 02:00 UTC." The category Directive captures attempts to prompt an action, such as "Upload the test file." The category Commissive identifies commitments to future behaviour, exemplified by "We will pay once we verify the decryptor." The category Negotiative–Evaluative applies to assessments or appraisals of proposals, as in "Your offer is unacceptable." The category Expressive–Metacommunicative includes stance-taking and interaction-management moves, such as "We are trying our best to cooperate."

The second layer assigns an *argumentative function* to each message, capturing the justificatory or strategic role the utterance plays in the negotiation. The category Grounds/Facts identifies factual claims used to support positions, for example "The attached screenshot shows the encrypted directory." The category Action Pressure denotes attempts to increase urgency or perceived cost, as in "If you delay, the price will double." The category Conditional Reasoning captures the use of if–then constructions to frame consequences, such as "If you refuse, we will leak the data." The category Value/Fairness Appeal applies to invocations of proportionality or legitimacy, for instance "We offer discounts for small companies." The category Face/Ethos Work includes moves to project credibility or manage social stance, as in "We are honest actors; we always provide decryptors." The category Process Maintenance identifies utterances aimed at sustaining the interactional flow, such as "Please respond within two hours."

An LLM-based classifier (gpt-4o, temperature zero) generated labels for each message. This approach follows recent studies [17] showing that, when coming to annotating text with argumentative guidelines, gpt-4o (and Claude) perform similarly to humans. The model was constrained by a structured prompt requiring a JSON dictionary containing the following fields: verbatim message text, party (victim or other), primary act, secondary subtypes, argumentative function, negotiation phase, and a short rationale explicitly citing linguistic cues. The schema enforced one primary act and one argumentative function per message, with optional secondary acts. Output was validated for JSON conformity before human review.

4. Results

4.1. Pragmatics Across Stages of Negotiations

We normalised each message's position within its conversation by dividing its turn index by the total number of turns, yielding a continuous progress value in $[0, 1]$. For comparative analysis across dialogues of different lengths, this value was discretised into twenty equal-width bins.

Table 1
Speech act categories with illustrative examples and argumentative functions

Speech Act Type	Illustrative Example (Akira Corpus)	Argumentative Function / Pragmatic Role
Informative–Declarative	"We have downloaded 400 GB of your sensitive data." "Payment received. The post deleted."	Provides situational grounding and factual updates; states conditions or changes of status that structure the negotiation environment.
Directive	"Send us 2–3 encrypted files so we can prove that we can restore them." "Please confirm payment within 24 hours."	Enables coercion and operational coordination; requests, instructs, or demands actions that drive the extortion process.
Commissive	"If you don't pay, we will publish your data." "After payment we will provide the decryptor and evidence of deletion."	Commits the speaker to future action; encodes threats, promises, and offers that structure conditional reasoning.
Negotiative–Evaluative	"We can reduce the price to 75,000 USD if you pay today." Victim: "We are a non-profit helping women; please consider our situation."	Modifies terms, introduces justifications, or evaluates proportionality; supports bargaining dynamics.
Expressive–Metacommunicative	"We appreciate your patience." "Please confirm you received our message."	Maintains the interactional channel and relational framing; expresses stance or coordinates process flow.

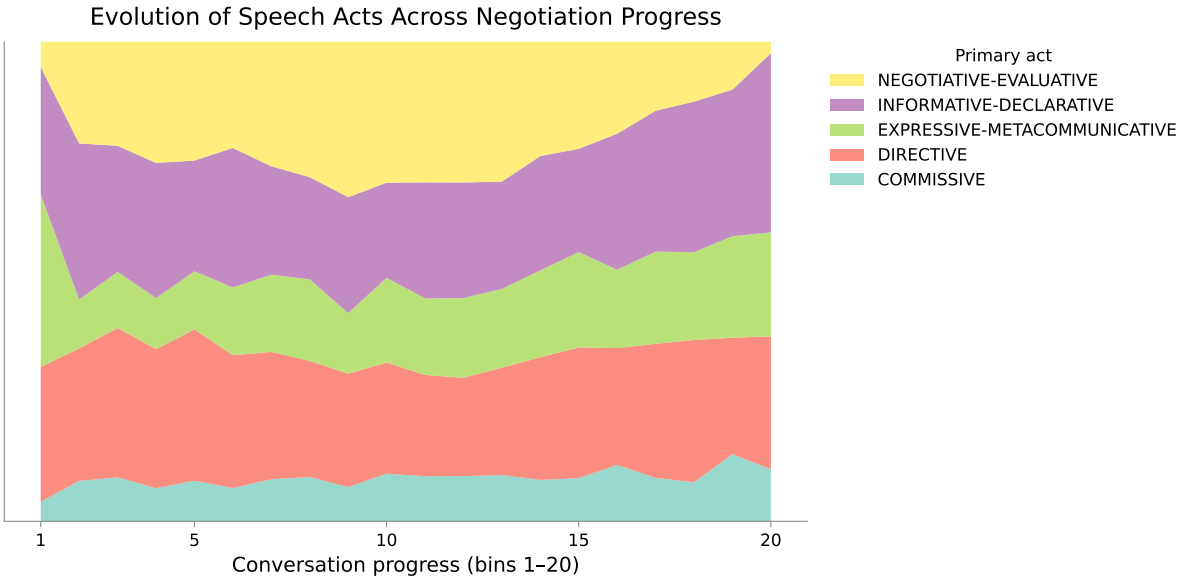

Figure 1: Evolution of primary speech act categories across normalised negotiation progress. Values represent relative frequencies within each of the twenty progress bins.

Evolution of speech acts across the negotiation process. Figure 1 presents these trends as a stacked distribution over the twenty progress bins. At the outset (bins 1–3), messages are dominated by Directive and Informative–Declarative acts. These early exchanges combine procedural instructions with status-setting disclosures. Expressive–Metacommunicative acts are also frequent in the initial bins, reflecting routine channel coordination and the establishment of interactional footing.

From the fourth bin onward, Negotiative–Evaluative acts increase steadily and become the most common category through the mid-conversation intervals (bins 7–13). This pattern indicates that once the basic parameters of contact and capability are established, the dialogue shifts toward bargaining,

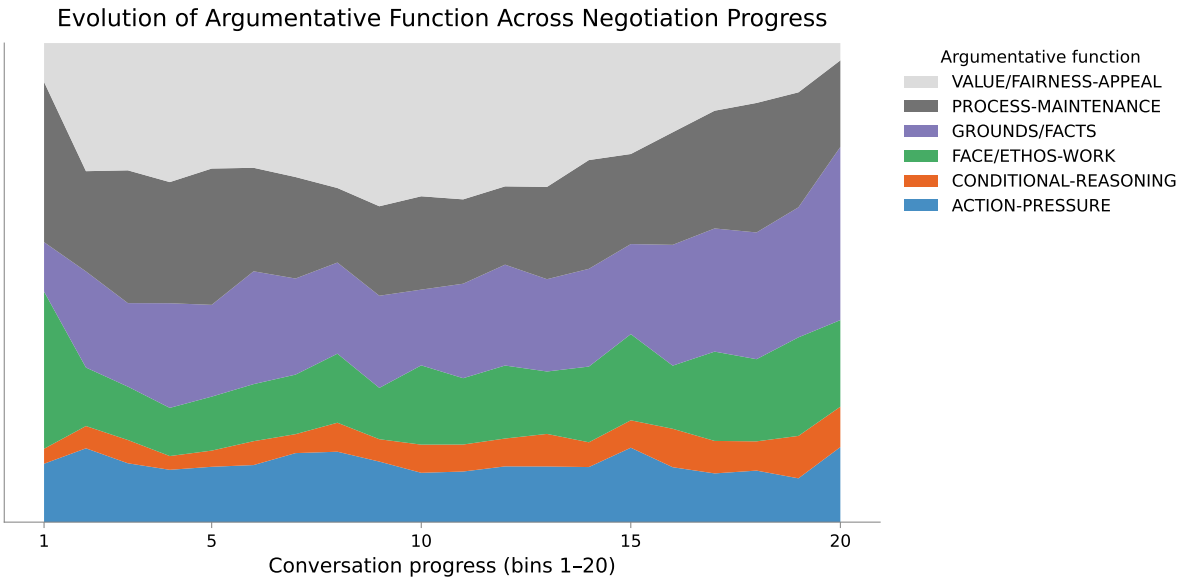

Figure 2: Evolution of argumentative functions across normalised negotiation progress. Values represent relative frequencies within the twenty progress phases.

justification, and adjustment of terms. The proportion of Commissive acts rises moderately across the same span, mirroring the introduction of conditional promises, concessions, or hardening of positions.

In the later bins (15–20), Informative–Declarative acts increase again. This trend corresponds to status updates, confirmations, or closing information as negotiations approach resolution. Directive acts remain consistently represented throughout, though their relative share decreases when bargaining peaks and increases again toward the end.

Evolution of argumentative functions across the negotiation process. Moving to the argumentative functions, Figure 2 visualises their distribution across the conversation segments, which shows a clear temporal structure. In the earliest phases (bins 1–3), Face/Ethos Work and Process Maintenance are prominent. This pattern indicates that the opening exchanges are devoted to establishing a workable channel, managing politeness or stance, and securing minimal interactional alignment. At the same time, Action Pressure is already present at moderate levels, signalling that coercive expectations are introduced early even while relational grounding is still under construction.

From phases 4 to 12, the proportion of Value/Fairness Appeal increases and becomes the most frequent category. This shift corresponds to the central bargaining stages, where both attackers and victims advance justificatory or proportionality-based arguments to reshape demands. Over the same interval, Grounds/Facts grows steadily, reflecting the exchange of evidence, situational clarifications, or status updates that support negotiation claims. Conditional Reasoning also becomes more common, marking the introduction of conditional promises or implicit sanctions as part of the bargaining dynamic.

In the later phases (13–20), Grounds/Facts becomes dominant. This trend reflects the movement toward resolution, where messages increasingly provide confirmations, technical information, or closing statements. Process Maintenance remains consistently represented, supporting the procedural coordination necessary to finalise agreements or end stalled conversations. Value/Fairness Appeal declines sharply toward the end, indicating that justificatory negotiation is largely suspended once the interaction approaches closure.

4.2. Group Attribution

Primary speech acts as attributional signals. To examine whether primary speech acts can assist group attribution, we consider their distribution across groups as reported in Figure 3. The

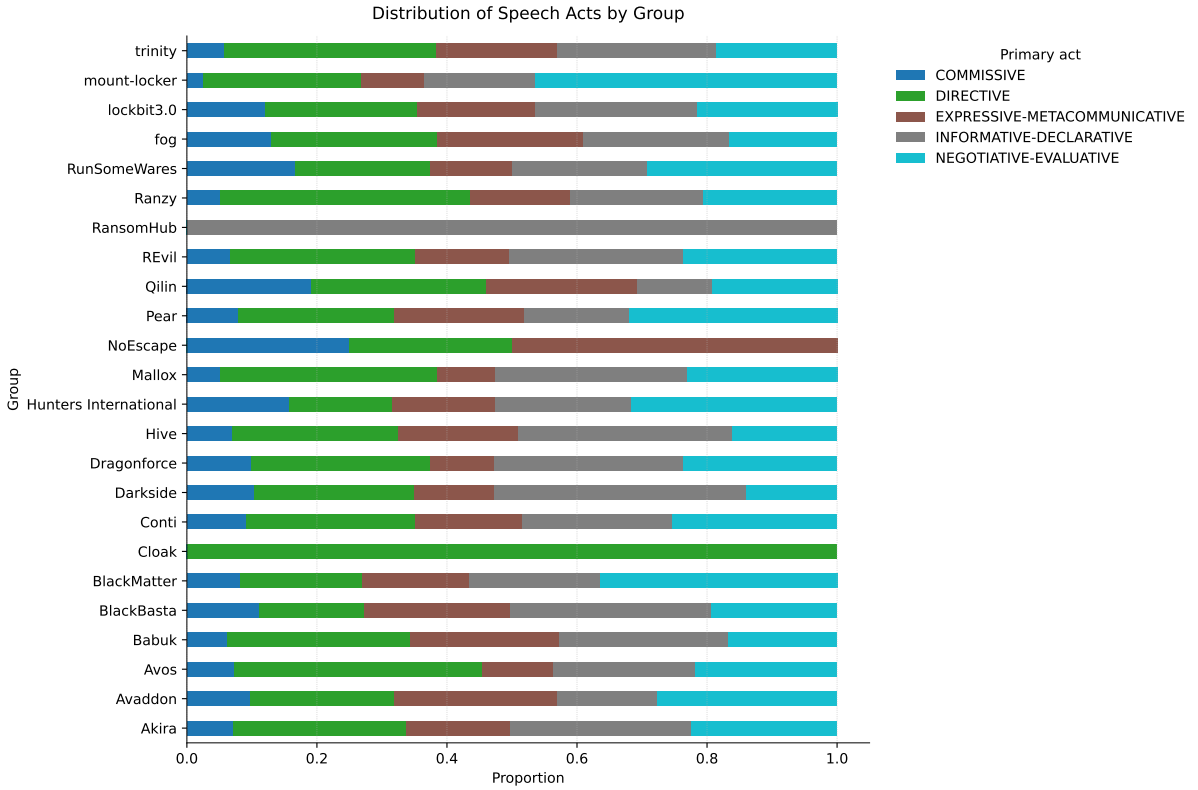

Figure 3: Distribution of primary speech acts across ransomware groups. The figure reports, for each group, the proportion of messages annotated with the five primary speech act categories.

descriptive patterns indicate that groups differ in how they deploy the five speech-act categories, yet these differences emerge with varying clarity. A balanced assessment therefore requires recognising both the indications of stylistic divergence and the empirical limits attached to them.

A first glance reveals that a few groups exhibit strikingly skewed profiles. *Cloak* uses only Directive acts, *RansomHub* relies exclusively on Informative–Declarative acts, and *NoEscape* shows an unusually high concentration of Expressive–Metacommunicative moves – albeit they are severely unrepresented in the dataset, hence such conclusions should not be taken as definitive without further data collection: it remains unclear whether such extreme distributions would persist with larger samples. For this reason, they offer useful but fragile attributional cues.

More established groups present a different picture. *Conti*, *REvil*, *Hive*, and *BlackBasta* share broadly similar proportions of Informative–Declarative, Directive, and Negotiative–Evaluative acts, while reserving Commissive and Expressive–Metacommunicative acts for more specific purposes. The overlap among these distributions indicates that speech-act proportions, taken alone, are not sufficiently discriminative to separate major groups with confidence. Even so, smaller but consistent differences appear. *BlackMatter* and *mount-locker* make heavier use of Negotiative–Evaluative acts, while *Avos* and *Mallox* rely less on Expressive–Metacommunicative acts than comparable groups. These nuances do not decisively determine attribution, but they contribute incremental evidence that can narrow down likely candidates in a broader analytical context.

Some groups also display internally coherent but distinctive equilibria. *Hunters International*, for example, distributes its speech acts in a near-uniform manner, a pattern that sets it apart from directive-heavy or evaluative-heavy groups. Such equilibrium does not offer a strong marker on its own, yet it becomes informative once combined with other traits, as it reduces the range of plausible attributions without overstating its discriminative value.

Taken together, these observations indicate that primary speech acts offer a limited but meaningful layer of evidence for attribution. Their value is strongest where distributions are extreme and weakest

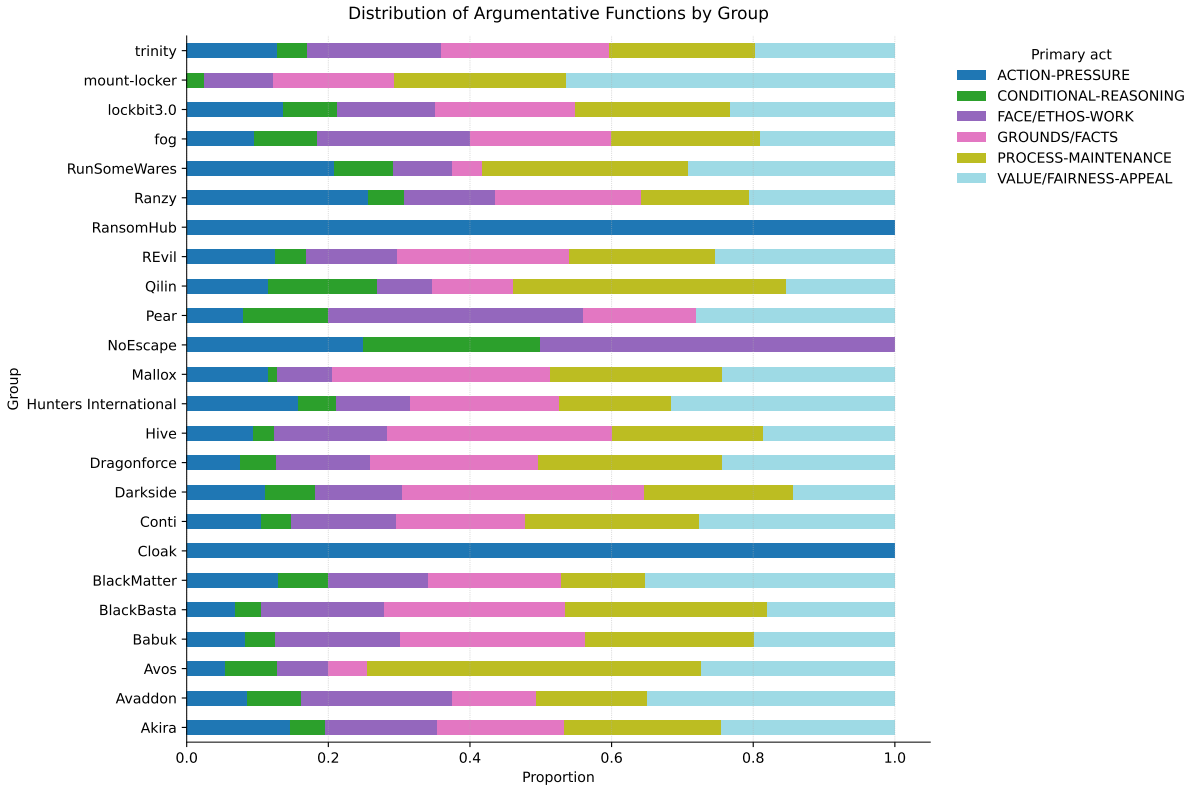

Figure 4: Distribution of argumentative functions across ransomware groups. Each bar represents, for a given group, the proportion of messages annotated with the six argumentative-function categories.

where groups converge on similar communicative styles. Their reliability improves when integrated with additional features—semantic cues, argumentative functions, temporal organisation, and contextual metadata—reflecting the multi-layered reasoning typical of cyber-threat-intelligence work. Within such composite frameworks, speech acts contribute a pragmatic dimension that supports, rather than replaces, richer forms of linguistic and behavioural evidence.

Argumentative functions as attributional signals. The distribution of argumentative functions across groups, summarised in Figure 4, reveals that groups differ not only in what they say but in how they structure justification, pressure, and interpersonal positioning. These patterns do not form sharp boundaries in every case, yet they introduce stylistic tendencies that, when read with due caution, contribute to an attributional perspective.

Some groups present highly distinctive profiles. *Cloak* and *RansomHub* consist almost entirely of Action–Pressure moves, suggesting a communicative style that relies on direct coercive framing rather than justification or procedural guidance. *NoEscape* contrasts sharply with this pattern through its heavy reliance on Face/Ethos–Work and Conditional–Reasoning, combined with the absence of factual or procedural content. These patterns stand out clearly, yet their interpretive weight is constrained by the limited number of chats underlying them. Their divergence is therefore notable but not, on its own, a stable foundation for attribution.

Larger and more representative groups exhibit a subtler landscape. *Conti*, *REvil*, *BlackBasta*, and *Hive* consistently combine Grounds/Facts and Process–Maintenance with a moderate amount of Face/Ethos–Work. Their relative balance gives them communicative profiles that are broadly similar but not identical. What differentiates them is often a matter of emphasis: *BlackBasta* places slightly more weight on procedural work, while *Hive* foregrounds factual content. These differences are not decisive, yet they show that argumentative functions encode recurrent preferences in how groups manage explanation, pressure, and rapport.

A further set of groups introduces more pronounced thematic tendencies. *Avaddon* and *BlackMatter* rely extensively on Value/Fairness–Appeal, reflecting a style where moral or evaluative arguments play a central role. *Mallox* and *Darkside* lean heavily on factual justification, grounding their interactions in claims about evidence, impact, or technical detail. *Avos* is unusual in placing a large share of its interactional effort in Process–Maintenance, suggesting a strong concern with keeping the exchange operationally coherent. These patterns, while not unique markers, provide a richer sense of how each group frames negotiation tasks and sustains its preferred narrative.

A few groups display internally balanced, almost patterned distributions. *Hunters International* and *Qilin* maintain a steady distribution across multiple categories, without leaning strongly on any single argumentative resource. This equilibrium does not generate strong separability, but it provides a recognisable communicative footprint that becomes informative when combined with other linguistic layers.

Overall, the argumentative-function layer introduces a dimension of variation that does not produce clear-cut attribution boundaries but offers consistent stylistic cues. Its contribution lies less in sharp distinctions than in recurrent tendencies: some groups argue by pressure, others by values, others by facts or procedural guidance. These tendencies matter analytically when combined with complementary indicators—speech acts, semantic content, temporal patterns, and contextual metadata. Within such integrated frameworks, argumentative functions help capture the strategic organisation of extortion dialogues, adding nuance to attribution without claiming determinative power.

5. Conclusion

This study shows that ransomware negotiation unfolds through stable pragmatic and argumentative routines, and the empirical patterns across the corpus substantiate this claim. Early interactional phases are marked by directives, disclosures, and metacommunicative moves that establish procedural alignment and define the parameters of engagement. As conversations progress, interlocutors shift towards evaluative, justificatory, and conditional acts that animate the bargaining core. In the closing stages, factual and procedural updates reassert themselves as negotiations move toward resolution. The corresponding distribution of argumentative functions reinforces this developmental trajectory: initial phases centre on interactional grounding, middle phases rely on proportionality-based reasoning and evidential clarification, and final phases consolidate factual grounding. Taken together, these findings demonstrate that extortion exchanges, despite their adversarial nature, follow structured communicative patterns rather than ad-hoc or chaotic forms of discourse.

The analysis also illustrates that pragmatic and argumentative features contribute a modest but meaningful layer of evidence for group attribution. While extreme stylistic profiles appear mainly in groups supported by limited material, more established actors display consistent tendencies in how they exert pressure, justify positions, maintain procedure, or manage interpersonal stance. These preferences do not support attribution in isolation, yet they offer discriminative value when integrated with semantic, temporal, and contextual indicators. The combined results therefore demonstrate that open-source ransomware communication enables systematic linguistic analysis, that negotiation language is organised by recurrent pragmatic and argumentative structures, and that these structures hold analytical value for both academic research and operational cyber-threat intelligence.

Future research can extend this study in two directions. A first line of work involves developing a multi–LLM annotation framework in which several state-of-the-art models independently label each message with both its primary speech act and its argumentative function. This approach would enable computation of inter-annotator agreement across models – allowing also to test recent results [17] in a complex and articulated domain – identification of systematic divergences, and construction of an aggregated gold-standard layer via majority voting or alternative ensemble strategies. Such a pipeline would strengthen the reliability of large-scale pragmatic annotation while enabling the release of a fully labelled dataset.

A second line of work concerns a focused analysis of those negotiations in which explicit monetary

bargaining occurs. Concentrating on these dialogues would allow a systematic examination of the argumentative moves that lead to reductions in the ransom demanded. To support this, formal argumentation [18] offers an established theoretical framework for modelling consistency and reinstatement of arguments [19], meta-argumentation [20] and value-driven considerations [21].

A further line of future work concerns the study of planning-based models [22] as a way to provide a formal account for the strategic negotiations. Automated planning provides a principled way to specify of how agents pursue goals through course of actions that could take uncertainty into account too [23]. Contemporary surveys in multi-agent systems demonstrate how planning frameworks capture interdependent decision processes, coordinated policy generation, and adaptive responses to dynamically evolving constraints [24]. In parallel, work on automated negotiation shows how strategic sequencing of communicative moves can be modelled as rational choice under incomplete information, with agents selecting offers, counteroffers, and threats to optimise expected utility while accommodating adversarial behaviour [25]. Extending ransomware-negotiation analysis through automated planning would provide a model-based approach to examine how threat actors structure coercive strategies over time, how victims attempt to preserve or expand their feasible options, and how specific pragmatic or argumentative moves operate as steps within a broader strategic plan. Bringing together these strands would provide a rigorous bridge between linguistic structure and reasoning, and would open a path toward cross-fertilisation between computational linguistics, formal argumentation, and automated planning.

Acknowledgments

The work was partially supported by the European Office of Aerospace Research & Development and the Air Force Office of Scientific Research under award numbers FA8655-22-1-7017 and FA8655-25-1-7067. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States government. This work was supported by project ACRE (AI-Based Causality and Reasoning for Deceptive Assets - 2022EP2L7H).

Declaration on Generative AI

Apart from using gpt-4o for text annotation, as described in the main text, GPT-5.1 was used for text refinement and language editing. All outputs were reviewed and edited by the authors, who retain full responsibility for the final content.

References

[1] H. Fujima, T. Kumamoto, Y. Yoshida, Using chatgpt to analyze ransomware messages and to predict ransomware threats, Research Square (2023). doi:10.21203/rs.3.rs-3645967/v1, preprint.

[2] T. Zhu, X. Li, W. Zhang, Applying chatgpt-powered game theory in ransomware negotiations, TechRxiv (2023). doi:10.36227/techrxiv.170244324.48846520, preprint.

[3] J. Faivre, Negotiations in tech: An analysis of asymmetric ransomware negotiations, SSRN Electronic Journal (2022). URL: https://ssrn.com/abstract=4530094. doi:10.2139/ssrn.4530094, preprint.

[4] J. L. Austin, How to Do Things with Words, Oxford University Press, Oxford, 1962.

[5] J. R. Searle, Speech Acts: An Essay in the Philosophy of Language, Cambridge University Press, Cambridge, 1969.

[6] J. R. Searle, A taxonomy of illocutionary acts, in: K. Gunderson (Ed.), Language, Mind, and Knowledge, University of Minnesota Press, 1975, pp. 344–369.

[7] F. H. van Eemeren, R. Grootendorst, A Systematic Theory of Argumentation: The Pragma-Dialectical Approach, Cambridge University Press, Cambridge, 2004.

[8] K. Bach, R. M. Harnish, Linguistic Communication and Speech Acts, MIT Press, Cambridge, MA, 1979.
[9] J. Thomas, Meaning in Interaction: An Introduction to Pragmatics, Routledge, London, 1995.
[10] G. Leech, Principles of Pragmatics, Longman, London, 1983.
[11] J. L. Mey, Pragmatics: An Introduction, 2 ed., Blackwell, Oxford, 2001.
[12] J. Culpeper, M. Haugh, (Im)politeness and Interaction: From Attitude to Practice, Cambridge University Press, Cambridge, 2014.
[13] N. Fairclough, Discourse and Social Change, Polity Press, Cambridge, 1992.
[14] L. L. Putnam, M. E. Roloff, Communication processes in negotiation, in: F. M. Jablin, L. L. Putnam (Eds.), Handbook of Organizational Communication, Sage, 2004, pp. 425–476.
[15] Casualtek, Ransomchats: Ransomware negotiation chat dataset, https://github.com/Casualtek/Ransomchats, 2023. MIT Licence.
[16] R. Fisher, W. Ury, B. Patton, Getting to Yes: Negotiating Agreement Without Giving In, Penguin Books, New York, 1991.
[17] A. Lindahl, Llms as annotators of argumentation, in: Proceedings of the 14th Joint Conference on Lexical and Computational Semantics (SEM 2025), 2025, pp. 242–252.
[18] P. Baroni, D. Gabbay, M. Giacomin, L. Van der Torre, Handbook of formal argumentation (2018).
[19] P. Baroni, F. Cerutti, M. Giacomin, On generalized notions of consistency and reinstatement and their preservation in formal argumentation, Artificial Intelligence 336 (2024) 104202.
[20] P. Baroni, F. Cerutti, M. Giacomin, G. Guida, Afra: Argumentation framework with recursive attacks, International Journal of Approximate Reasoning 52 (2011) 19–37.
[21] T. J. Bench-Capon, Persuasion in practical argument using value-based argumentation frameworks, Journal of Logic and Computation 13 (2003) 429–448.
[22] M. Ghallab, D. S. Nau, P. Traverso, Automated Planning and Acting, Cambridge University Press, 2016.
[23] D. Aineto, E. Scala, Cost-optimal fond planning as bi-objective best-first search, volume 35, 2025, p. 140 – 148. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-105017473122&doi=10.1609%2ficaps.v35i1.36110&partnerID=40&md5=ef810522b1ba6f155322790bed17b366. doi:10.1609/icaps.v35i1.36110, cited by: 0; All Open Access, Gold Open Access.
[24] A. Torreño, E. Onaindia, O. Sapena, A review of multi-agent planning, Artificial Intelligence 305 (2022) 103682. doi:10.1016/j.artint.2021.103682.
[25] T. Baarslag, K. V. Hindriks, E. H. Gerding, A survey of automated negotiation methods for multi-agent systems, Decision Support Systems 138 (2020) 113382. doi:10.1016/j.dss.2020.113382.
[26] CISA and FBI, #stopransomware: Avoslocker ransomware (update), 2023. URL: https://www.cisa.gov/news-events/cybersecurity-advisories/aa23-284a, cybersecurity advisory AA23-284A. Accessed 27 November 2025.
[27] CISA, Threat profile: Black basta, 2023. URL: https://www.cisa.gov/sites/default/files/srware/black-basta-threat-profile.pdf, tLP:CLEAR threat profile report 202303151200. Accessed 27 November 2025.
[28] CISA, FBI, and NSA, #stopransomware: Blackmatter ransomware, 2021. URL: https://www.cisa.gov/news-events/cybersecurity-advisories/aa21-291a, cybersecurity advisory AA21-291A. Accessed 27 November 2025.
[29] UK National Crime Agency, Ransomware criminals sanctioned in joint uk/us crackdown on international cyber crime, 2023. URL: https://www.nationalcrimeagency.gov.uk/news/ransomware-criminals-sanctioned-in-joint-uk-us-crackdown-on-international-cyber-crime, news release. Accessed 27 November 2025.
[30] U.S. Department of Justice, Department of justice seizes $2.3 million in cryptocurrency paid to the darkside ransomware group, 2021. URL: https://www.justice.gov/archives/opa/pr/department-justice-seizes-23-million-cryptocurrency-paid-ransomware-extortionists-darkside, press release, 7 June 2021. Accessed 27 November 2025.
[31] U.S. Department of Justice, U.s. department of justice disrupts hive ransomware variant, 2023. URL: https://www.justice.gov/archives/opa/pr/us-department-justice-disrupts-hive-ransomwar

e-variant, press release, 26 January 2023. Accessed 27 November 2025.
[32] Group-IB, Qilin ransomware: Tactics, attack methods & mitigation, 2023. URL: https://www.group-ib.com/blog/qilin-ransomware/, technical blog. Accessed 27 November 2025.
[33] U.S. Department of Health and Human Services, HC3, Qilin ransomware threat profile, 2024. URL: https://www.hhs.gov/sites/default/files/qilin-threat-profile-tlpclear.pdf, threat profile TLP:CLEAR. Accessed 27 November 2025.
[34] UK National Crime Agency, Lockbit leader unmasked and sanctioned, 2024. URL: https://www.nationalcrimeagency.gov.uk/news/lockbit-leader-unmasked-and-sanctioned, news release on Operation Cronos. Accessed 27 November 2025.
[35] U.S. Department of Health and Human Services, HC3, Trinity ransomware: Threat actor profile, 2024. URL: https://www.hhs.gov/sites/default/files/trinity-ransomware-threat-actor-profile.pdf, hC3 threat-actor profile report 202410041500. Accessed 27 November 2025.

A. Ransomware Groups

Akira. Akira is a financially motivated ransomware-as-a-service (RaaS) operation first observed in early 2023, noted for double-extortion attacks against mid-sized and large organisations in Europe and North America. Recent advisories describe Akira affiliates exploiting the SonicWall SonicOS vulnerability CVE-2024-40766, as well as unpatched backup infrastructure, and extending their focus to Nutanix AHV virtual machines in addition to earlier VMware ESXi and Hyper-V targets.[2] Public reporting characterises Akira as a Russian-language RaaS outfit, but open sources stop short of attributing it to a state actor.

Avaddon. Avaddon operated as a global RaaS programme from 2020 until mid-2021, combining phishing campaigns and exploitation of exposed services in double-extortion attacks on enterprises and critical infrastructure. In June 2021, the group abruptly ceased operations and released more than 2,900 decryption keys, a move documented by both media and CTI vendors.[3] Subsequent analysis has linked Avaddon's operator or affiliate base to later brands such as NoEscape.

Avos (AvosLocker). The group labelled *Avos* in the dataset corresponds to AvosLocker, a double-extortion RaaS first seen in 2021 and targeting US critical infrastructure sectors, including finance, manufacturing, and government. A joint CISA–FBI advisory documents AvosLocker's evolution from Windows-only payloads to Linux and VMware ESXi variants and provides indicators of compromise and tactics used in recent campaigns.[26] Open sources classify AvosLocker as a profit-driven criminal group; there is no public evidence of formal state control.

Babuk. Babuk (Babuk Locker) emerged in early 2021 and quickly gained notoriety for a high-profile attack on the Washington, DC Metropolitan Police Department in which sensitive internal police files were exfiltrated and leaked after negotiations failed.[4] Shortly afterwards, Babuk announced a shift away from encryption while its leaked source code was repurposed by other actors, including in the ARCrypter family. Public reporting consistently locates Babuk within the Russian-speaking cybercrime milieu, while law-enforcement and CTI sources continue to treat it as a criminal rather than state-directed operation.

[2] https://www.bleepingcomputer.com/news/security/cisa-warns-of-akira-ransomware-linux-encryptor-targeting-nutanix-vms/ (on 27 November 2025).
[3] https://www.bleepingcomputer.com/news/security/avaddon-ransomware-shuts-down-and-releases-decryption-keys/ (on 27 November 2025).
[4] https://www.databreachtoday.com/babuk-ransomware-gang-posts-more-dc-metro-police-data-a-16575 (on 27 November 2025).

BlackBasta. BlackBasta appeared in early 2022 and is widely identified as a double-extortion RaaS operation with strong links to the earlier Conti/Ryuk ecosystem. Official and CTI threat profiles describe BlackBasta as a Russian-speaking group that rapidly accumulated victims across manufacturing, professional services, and health care using targeted intrusions rather than spray-and-pray campaigns.[27] The available evidence points to a financially motivated syndicate rooted in Russian-speaking cybercrime networks; public sources have not provided conclusive proof of direct state command.

BlackMatter. BlackMatter surfaced in mid-2021 shortly after the disruption of DarkSide and was quickly assessed by CISA and others as a possible rebrand or successor, based on code similarities and overlapping infrastructure.[28] It operated as a RaaS service targeting large enterprises with a double-extortion model, initially claiming to exclude certain critical sectors from its targeting. Open sources discuss the likelihood of BlackMatter operating from Russian-speaking jurisdictions, but treat it as a financially motivated criminal enterprise rather than an openly acknowledged state proxy.

Cloak. Cloak is a relatively recent ransomware operation first publicly observed in late 2022–2023, using an ARCrypter variant derived from the leaked Babuk source code. CTI reporting highlights Cloak's focus on small and medium-sized organisations, particularly in Europe, and its use of loaders that terminate security and backup services, delete shadow copies, and deploy encrypted payloads.[5] Current analysis treats Cloak as a financially motivated actor whose technical lineage traces back to Babuk, without robust evidence tying it to any specific state sponsor.

Conti. Conti was one of the most prolific RaaS ecosystems of the late 2010s and early 2020s, closely associated with TrickBot and Ryuk and responsible for large numbers of attacks on health-care providers and local government. UK–US sanctions documentation and investigative reporting portray Conti as a highly organised, Russia-based cybercrime syndicate with a quasi-corporate structure and deep integration with other malware operators.[29] Western government statements have suggested alignment between elements of the Conti/TrickBot network and Russian strategic interests, but they continue to frame Conti primarily as a transnational criminal enterprise.

DarkSide. DarkSide operated as a high-impact RaaS group in 2020–2021 and is best known for the Colonial Pipeline attack of May 2021, which resulted in fuel shortages along the US east coast. A US Department of Justice case study documents the subsequent seizure of part of the ransom payment (63.7 bitcoin), highlighting law-enforcement focus on cryptocurrency tracing in disrupting the group.[30] After the Colonial incident and associated pressure, DarkSide announced its "closure", and its code and infrastructure are widely believed to have fed into later brands such as BlackMatter.

DragonForce. DragonForce began as a hacktivist collective but, by 2023–2025, evolved into a RaaS operation combining ideological messaging with profit-driven multi-extortion campaigns. CTI analyses and vendor blogs describe DragonForce targeting high-street retailers and media organisations, especially in the UK and Europe, and leveraging code bases drawn from other major ransomware families.[6] Recent reporting also notes public rivalry with other RaaS brands such as RansomHub, illustrating the competitive dynamics in the current ransomware ecosystem.

Hive. Hive was a RaaS programme active from mid-2021 until its disruption in January 2023, when a joint FBI–DOJ operation infiltrated its infrastructure, obtained decryption keys, and quietly assisted more than 1,500 victims in over 80 countries.[31] Before takedown, Hive focused heavily on health-care and public-sector entities and generated at least USD 100 million in ransom payments. Official

[5] https://www.halcyon.ai/threat-group/cloak (on 27 November 2025).
[6] https://www.sentinelone.com/blog/dragonforce-ransomware-gang-from-hacktivists-to-high-street-extortionists/ (on 27 November 2025).

statements describe Hive as a transnational criminal group; no public attribution links it directly to a state.

Hunters International. Hunters International appeared in late 2023 and has been widely assessed as a likely Hive successor or rebrand, based on code reuse and overlapping infrastructure. Early technical analyses show that Hunters International adopted many of Hive's tools and techniques and initially pursued a similar RaaS model before experimenting with data-extortion-only operations.[7] Public CTI sources characterise it as a Russian-speaking, profit-driven actor; no robust evidence has been published of state direction.

Mallox. Mallox (also known as TargetCompany, Fargo, and Tohnichi) is a ransomware strain and group active since June 2021, notable for exploiting unsecured Microsoft SQL servers for initial access. Palo Alto Networks' Unit 42 describes Mallox as a Windows-focused ransomware that follows the double-extortion model and has shown sustained activity across multiple years.[8] Open sources consistently treat Mallox as part of the broader financially motivated cybercrime ecosystem, with no state nexus identified.

NoEscape. NoEscape is a multi-extortion RaaS first observed in mid-2023, widely believed to be a rebrand or successor of Avaddon based on strong overlaps in infrastructure and negotiation style. CTI reporting documents NoEscape's support for Windows, Linux, and VMware ESXi payloads and its aggressive targeting of enterprises across sectors in Europe and North America.[9] As with Avaddon, current public sources classify NoEscape as a financially motivated Russian-speaking operation without formal state attribution.

Pear. Pear (PEAR, "Pure Extraction And Ransom") is an extortion-focused operation first reported in mid-2025, distinguished by its emphasis on data theft and Tor-only infrastructure rather than encryption-led attacks. Threat-intelligence bulletins describe PEAR as a "data-broker" style actor that rapidly publishes victim data via multiple onion services, with early victims in legal, technology, and professional services.[10] Given its recent emergence, Pear is best understood as an evolving financially motivated crew rather than a fully institutionalised RaaS franchise.

Qilin. Qilin, also tracked under its earlier alias Agenda, is a RaaS operation active since around 2022 and targeting both Windows and Linux environments, with a strong presence in health-care and manufacturing. Group-IB and subsequent HHS threat profiles describe Qilin as a Russia-based operation offering customisable Rust and Go payloads and operating a double-extortion leak site.[32, 33] Public sources emphasise its aggressiveness and scale but do not attribute it to a specific state.

REvil. REvil (Sodinokibi) was a highly influential RaaS cartel active from 2019 until its disruption in 2021–2022, associated with major incidents such as the Kaseya supply-chain attack and the JBS meat-processing breach. Media and court reporting describe REvil as a Russia-linked group whose affiliates caused hundreds of millions of dollars in losses before coordinated US–Russian law-enforcement actions led to arrests in early 2022.[11] Public sources frame REvil as a criminal enterprise embedded in Russian cybercrime ecosystems, rather than as an openly acknowledged state actor.

[7] https://www.bleepingcomputer.com/news/security/new-hunters-international-ransomware-possible-rebrand-of-hive/ (on 27 November 2025).
[8] https://unit42.paloaltonetworks.com/mallox-ransomware/ (on 27 November 2025).
[9] https://www.bleepingcomputer.com/news/security/meet-noescape-avaddon-ransomware-gangs-likely-successor/ (on 27 November 2025).
[10] https://redpiranha.net/news/threat-intelligence-report-august-5-august-11-2025 (on 27 November 2025).
[11] https://cybernews.com/news/hacker-jailed-revil-ransomware-attacks/ (on 27 November 2025).

RansomHub. RansomHub is a rapidly growing RaaS programme first advertised in early 2024 and, by late 2024–2025, assessed as one of the most active ransomware groups globally. A detailed Group-IB profile highlights RansomHub's high affiliate profit share, dual-extortion tactics, and wide victim base across sectors.[12] While some reporting emphasises rivalry and "turf wars" with DragonForce, there is no reliable open-source evidence of direct state sponsorship.

Ranzy. Ranzy Locker is a RaaS variant first seen in 2020 as an evolution of earlier families such as Ako and ThunderX/MedusaLocker, with a particular focus on US organisations in sectors including construction, academia, government, IT, and transportation. An FBI flash alert, reported in secondary CTI sources, attributes at least dozens of successful intrusions to Ranzy and notes the group's use of a double-extortion leak site.[13] Ranzy is classified as a financially motivated criminal actor, with no substantiated state link.

RunSomeWares. RunSomeWares is an emerging ransomware group first tracked in early 2025, when its leak site went live and it claimed a small initial cluster of victims. CTI roundups describe RunSomeWares as a conventional double-extortion operation with limited bespoke tooling, and note that available information about its operators and geography remains sparse.[14] At the time of writing, there is insufficient open-source evidence to place RunSomeWares within a specific regional or state-backed ecosystem.

fog. Fog is a newer ransomware operation that has gained visibility for its role in exploiting the SonicWall CVE-2024-40766 SSL-VPN vulnerability, often in parallel with Akira, to compromise enterprise networks. Technical reporting documents Fog and Akira affiliates using stolen or brute-forced VPN credentials and the SonicOS flaw to gain initial access, followed by standard lateral movement and encryption phases.[15] Public sources describe Fog as a financially motivated actor; their coverage does not yet support precise conclusions about its organisational structure or state connections.

lockbit3.0. LockBit 3.0 (LockBit Black) is an evolution of the LockBit family that, by the early 2020s, had become one of the most active global RaaS operations, with thousands of victims in more than 120 countries. CTI analyses outline LockBit 3.0's sophisticated affiliate model and custom tooling, while an international law-enforcement operation led by the UK National Crime Agency disrupted LockBit infrastructure in early 2024 and publicly identified its administrator "LockBitSupp" as Dmitry Khoroshev.[34] Despite repeated disruptions, LockBit-derived variants continue to appear, underscoring the resilience of this Russia-based criminal ecosystem.

mount-locker. Mount Locker is a RaaS operation first observed in mid-2020, targeting enterprises worldwide with high ransom demands and double-extortion tactics. BlackBerry's technical analysis details Mount Locker's use of Windows Active Directory APIs for lateral movement and its subsequent evolution into related brands such as Astro Locker.[16] Open-source reporting treats Mount Locker as a conventional financially motivated syndicate; its precise geographic base remains uncertain.

trinity. Trinity is a relatively new ransomware operation first observed in 2024 and associated with attacks on critical infrastructure and health-care organisations, including incidents reported in Europe and the United States. An HHS threat-actor profile describes Trinity's use of double extortion, ChaCha20-based encryption, and extensive system reconnaissance and privilege escalation prior to data exfiltration

[12] https://www.group-ib.com/masked-actors/ransomhub/ (on 27 November 2025).
[13] https://www.acronis.com/en-gb/tru/posts/fbi-warns-of-ranzy-locker-ransomware-threat/ (on 27 November 2025).
[14] https://www.cyfirma.com/research/tracking-ransomware-february-2025/ (on 27 November 2025).
[15] https://www.bleepingcomputer.com/news/security/fog-ransomware-targets-sonicwall-vpns-to-breach-corporate-networks/ (on 27 November 2025).
[16] https://blogs.blackberry.com/en/2020/12/mountlocker-ransomware-as-a-service-offers-double-extortion-capabilities-to-affiliates (on 27 November 2025).

and encryption.[35] Current open-source analysis notes possible overlaps with earlier code families but provides no firm evidence of state sponsorship, framing Trinity as a high-impact but still maturing criminal group.