Preprints.org

Article

Not peer-reviewed version

Cognitive Attack Topology (CAT): A Topological Framework for Modeling and Detecting Human-Centric Cyber Attacks

Sai Praneeth Reddy Dhadi *, Jithender Reddy M., G.N.R. Prasad

Posted Date: 31 March 2026

doi: 10.20944/preprints202603.2402.v1

Keywords: cognitive cybersecurity; social engineering detection; trust topology; human-centric security; graph-based security; adversarial manipulation; behavioral threat modeling; AI fraud detection

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2402.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Cognitive Attack Topology (CAT): A Topological Framework for Modeling and Detecting Human-Centric Cyber Attacks

Sai Praneeth Reddy Dhadi [1,*], Jithender Reddy M. [1] and G.N.R. Prasad [2]

[1] Dept. of Computer Science & Engg., Vasavi College of Engineering, Hyderabad, India

[2] Chaitanya Bharathi Institute of Technology, Gandipet, Hyderabad, India

* Correspondence: spreddydhadi@gmail.com

Abstract

The dominant paradigm in cybersecurity continues to privilege infrastructure, protocol integrity, and endpoint resilience, while an increasing fraction of high-impact attacks bypasses these controls by directly manipulating human cognition. This work formalizes such attacks as structured distortions within human digital trust interactions, introducing Cognitive Topological Cybersecurity (CTC) as a rigorous analytical framework. Within this paradigm, interactions are modeled as a high-dimensional manifold whose geometry encodes trust relationships across users, communication channels, and psychological signals. The proposed Cognitive Attack Topology (CAT) framework operationalizes this view through the construction of a Trust Topology Tensor (TTT), enabling the quantification of adversarial influence via Cognitive Distortion Energy (CDE) and its normalized form, the Trust Distortion Index (TDI). A complementary Cognitive Manipulation Score (CMS) captures the composite effect of urgency, fear, authority, and persuasion signals. The framework is instantiated in a multi-layer architecture that integrates transformer- based signal decomposition with dynamic graph modeling and topological anomaly detection. Empirical evaluation is conducted on the GCT-100K dataset, a large-scale benchmark comprising real, public, and synthetically generated cognitive attack interactions. The CAT model achieves a ROC-AUC of 0.9555, exceeding conventional text-based baselines, while maintaining robustness under adversarial cloaking conditions. Notably, performance remains invariant across linguistic boundaries, with a multilingual AUC of 0.984, indicating that topological features of trust manipulation exhibit language-agnostic structure. These results establish that human-targeted cyber-attacks can be detected not as isolated semantic artifacts, but as measurable geometric perturbations in trust space, motivating a shift toward cognition-aware, topology-driven defensive systems.

Keywords: cognitive cybersecurity; social engineering detection; trust topology; human-centric security; graph-based security; adversarial manipulation; behavioral threat modeling; AI fraud detection

1. Introduction

1.1. Problem Context

Cybersecurity has historically evolved in response to shifts in the locus of vulnerability. Early defensive paradigms were anchored in network perimeters, where intrusion detection systems and firewall architectures sought to prevent unauthorized access through protocol-level inspection and traffic filtering. As adversaries adapted, the attack surface migrated upward into application layers, exploiting software vulnerabilities, misconfigurations, and API-level weaknesses. In recent years, however, a more profound transition has emerged one that does not target computational systems directly but instead exploits the cognitive processes of human users embedded within these systems.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2402.v1

This transition toward human-targeted attacks reflects a structural asymmetry: while computational systems operate under deterministic rules and can be hardened through formal verification and patching, human cognition remains probabilistic, context-dependent, and susceptible to manipulation. Modern fraud campaigns, including phishing, OTP scams, and impersonation attacks, rarely rely on sophisticated technical exploits; instead, they engineer psychological conditions that induce users to violate their own security expectations. The operational primitives of such attack's urgency, fear, authority, and reward are not incidental features but core mechanisms through which adversaries reshape user decision-making processes in real time.

From a systems perspective, the human user is no longer an external entity interacting with secure infrastructure, but an integral node within a distributed trust network. Consequently, attacks targeting cognition cannot be adequately represented as isolated events or anomalous packets. Rather, they manifest as structured perturbations within the relational geometry of trust that governs interactions between users, institutions, and communication channels. Despite the increasing prevalence of such attacks, existing cybersecurity frameworks lack the conceptual and mathematical tools necessary to characterize, measure, and mitigate these distortions in a principled manner [1,2].

1.2. Research Gap

Contemporary detection systems remain largely constrained by paradigms optimized for machine-centric threats. Signature-based approaches rely on known indicators of com- promise, rendering them ineffective against novel or adaptive social engineering strategies. Machine learning models, particularly those based on natural language processing, have demonstrated success in identifying phishing content; however, these systems predominantly operate on surface-level semantic or lexical features, implicitly assuming that malicious intent is encoded in textual patterns alone [3,4]. Such assumptions fail under adversarial conditions where attackers deliberately obfuscate language, employ multilingual strategies, or mimic legitimate institutional communication with high fidelity.

More fundamentally, existing approaches treat interactions as static artifacts rather than dynamic processes embedded within evolving relational contexts. The notion of "trust" is either absent or reduced to heuristic scoring mechanisms that lack formal structure. Graph- based security models have attempted to capture relationships between entities, yet these representations are typically static and do not incorporate the cognitive dimensions that govern how trust is formed, transferred, and exploited [5]. Behavioral security models, while acknowledging human factors, remain largely qualitative and resist integration into scalable detection systems.

The absence of a formal geometric or topological representation of trust manipulation constitutes a critical gap. Without such a representation, it is not possible to define invariants, quantify distortions, or establish rigorous detection criteria for cognitive attacks. As a result, current systems are inherently reactive, brittle under adversarial adaptation, and incapable of generalizing across communication modalities or linguistic boundaries [6].

1.3. Research Objectives

This work addresses the limitations by introducing a formal framework in which human-targeted cyber-attacks are modeled as topological distortions within a high- dimensional trust manifold. The primary objective is to reconceptualize cybersecurity from a purely computational discipline into one that explicitly incorporates the geometry of human–digital interactions.

To this end, the study develops a mathematical representation of trust through the formulation of a multi-dimensional tensor structure that encodes relationships between users, communication channels, and psychological signals. Building upon this representation, the framework defines measurable quantities that capture the intensity and nature of adversarial influence, enabling the quantification of cognitive manipulation as an energy function over the trust manifold. This formulation allows for the derivation of normalized indices that distinguish benign interactions from adversarial distortions under varying environmental conditions.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2402.v1

A second objective is the design of a real-time detection architecture capable of operationalizing these theoretical constructions. This architecture integrates multi-modal signal processing, dynamic graph construction, and topological anomaly detection to identify and respond to cognitive attacks as they unfold. The emphasis is not merely on classification accuracy, but on the ability to detect structural deviations in trust relationships that are invariant to language, modality, and superficial content variations.

1.4. Contributions

The contributions of this work are fourfold. First, it establishes *Cognitive Topological Cybersecurity (CTC)* as a new research direction that frames cyber-attacks in terms of geometric transformations within human-centric trust networks. This perspective departs fundamentally from existing paradigms by treating trust not as an abstract concept, but as a measurable and manipulable structure.

Second, it introduces a unified mathematical framework comprising the *Trust Topology Tensor (TTT)*, *Cognitive Distortion Energy (CDE)*, *Trust Distortion Index (TDI)*, and *Cognitive Manipulation Score (CMS)*. Together, these constructs provide a rigorous basis for representing and quantifying cognitive attacks, enabling the detection of adversarial behavior as deviations from baseline trust configurations.

Third, it proposes the *Cognitive Attack Topology (CAT) Engine*, a layered system architecture that integrates transformer-based signal extraction with dynamic graph modeling and topological anomaly detection. The architecture is designed to operate across heterogeneous communication channels, supporting real-time intervention mechanisms that extend beyond passive detection.

Fourth, it presents the *GCT-100K* dataset, a large-scale benchmark specifically curated

to capture the multi-dimensional nature of human-targeted cyber attacks. The dataset combines real-world anonymized interactions, publicly available corpora, and synthetically generated adversarial samples, providing a comprehensive testbed for evaluating cognitive-aware security models. Empirical validation demonstrates that the proposed framework achieves state-of-the-art performance, while maintaining robustness under adversarial and multilingual conditions.

Collectively, these contributions advance the field toward a principled, quantitative understanding of human-centric cyber threats, laying the groundwork for future systems that can defend not only computational infrastructure, but the cognitive processes upon which digital trust ultimately depends.

2. Literature Review

2.1. Traditional Cybersecurity Models

The intellectual foundations of cybersecurity have been historically shaped by the assumption that adversarial activity manifests primarily through violations of computational integrity. Within this paradigm, *signature-based detection* emerged as the earliest systematic defence, relying on curated databases of known attack patterns to identify malicious artifacts. While effective against previously observed threats, such systems exhibit inherent brittleness when confronted with polymorphic or zero-day attacks, as their operational logic is fundamentally retrospective [7]. The rigidity of signature matching further precludes adaptation to attacks that do not produce stable, machine-identifiable fingerprints.

To address these limitations, *machine learning-based anomaly detection* introduced probabilistic modelling into cybersecurity. Techniques ranging from clustering and statistical profiling to deep learning architectures have been employed to identify deviations from established baselines in network traffic, system calls, and user behaviour [8]. Despite their methodological sophistication, these approaches remain tightly coupled to observable system-level features, implicitly assuming that malicious activity perturbs measurable computational signals. This assumption breaks down in scenarios where the adversarial payload is cognitively mediated rather than technically encoded.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2402.v1

Parallel to these developments, *network- and endpoint-centric security architectures* have continued to dominate operational deployments. Intrusion detection systems, endpoint protection platforms, and zero-trust architectures collectively aim to minimize the attack surface by enforcing strict access controls and continuous verification [9]. However, such systems are fundamentally limited by their scope: they secure the infrastructure surrounding the user but do not model the user as an active component within the threat landscape. Consequently, attacks that exploit legitimate communication channels and induce voluntary user actions such as credential disclosure or transaction authorization remain largely invisible to these defences.

2.2. Social Engineering Detection

The increasing prevalence of phishing, impersonation, and fraud has motivated a distinct line of research focused on *social engineering detection* [55]. Early approaches relied on heuristic rules, such as the presence of suspicious URLs or anomalous sender metadata. More recent work has leveraged advances in natural language processing to classify messages based on lexical, syntactic, and semantic features [10]. Deep neural architectures, including recurrent and transformer-based models, have demonstrated improved performance by capturing contextual dependencies within textual data [11].

Despite these advances, NLP-based detection frameworks exhibit structural limitations that constrain their effectiveness in adversarial settings. First, their reliance on language-specific features introduces *linguistic dependency*, making models sensitive to variations in grammar, vocabulary, and cultural context. Attackers can exploit this dependency by crafting multilingual or code-switched messages that evade detection without altering the underlying manipulation strategy [12]. Second, these models primarily operate on *surface-level representations*, capturing what is said but not how it reshapes the relational dynamics between entities. As a result, they struggle to detect attacks that closely mimic legitimate communication patterns while subtly manipulating trust.

Behavioural modelling approaches have attempted to address some of these shortcomings by incorporating user interaction patterns, such as response time, click behaviour, and communication frequency [13]. While these features provide additional context, they remain observational rather than structural, offering limited insight into the mechanisms through which trust is constructed and exploited. The absence of a unifying representation that integrates linguistic, behavioural, and relational dimensions continues to hinder progress in this domain.

2.3. Graph-Based Security Models

Graph-theoretic approaches have introduced a relational perspective to cybersecurity by modelling systems as networks of interconnected entities. *Attack graphs* represent potential paths through which an adversary can compromise a system, enabling the analysis of multi-stage attack strategies and the identification of critical vulnerabilities [5]. These models have proven valuable in understanding system-level dependencies and in prioritizing defensive interventions.

In parallel, *trust network models* have been developed to capture relationships between users, devices, and services, often assigning weights to edges based on historical interactions or reputation scores [14]. Such models acknowledge that security is not solely a property of individual components but emerges from the structure of interactions within the network.

However, existing graph-based approaches are predominantly *static* and lack the capacity to represent the dynamic evolution of trust under adversarial influence [58]. Edge weights are typically derived from historical data and updated incrementally, without accounting for the immediate and context-dependent impact of psychological manipulation. More critically, these models do not incorporate *cognitive dimensions* such as urgency or authority that modulate how trust is perceived and transferred in real time. As a result, they are unable to capture the transient yet decisive distortions that characterize human-targeted cyber-attacks.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2402.v1

2.4. Cognitive Security and Behavioral Models

A growing body of research has sought to incorporate insights from psychology and behavioural science into cybersecurity. Studies on phishing susceptibility, decision-making under stress, and the effectiveness of persuasive techniques have identified key factors that influence user behaviour, including time pressure, perceived authority, and emotional arousal [15,16]. These findings have informed the design of user education programs and awareness campaigns aimed at reducing vulnerability to social engineering.

The concept of *human-in-the-loop security* further emphasizes the role of users as active participants in defensive systems, advocating for interfaces and interventions that guide user decisions in real time [17]. While these approaches acknowledge the centrality of human cognition, they remain largely qualitative and are not easily integrated into automated detection pipelines.

Recent efforts have attempted to quantify behavioural signals through metrics such as cognitive load, sentiment, and engagement patterns [18]. Although these metrics provide valuable proxies for psychological states, they are typically treated as independent features rather than components of a coherent mathematical structure. The absence of a unified formalism prevents the systematic analysis of how these factors interact to produce exploitable vulnerabilities. There is no established framework for representing cognitive manipulation as a measurable transformation within a structured space of trust relationships.

2.5. Research Gap Synthesis

The preceding analysis reveals a fragmentation of approaches across three dominant paradigms, each addressing a subset of the problem space while leaving critical dimensions unmodeled. NLP-based systems excel at textual classification but lack a representation of relational trust geometry. Graph-based models capture structural relationships but remain static and devoid of cognitive semantics. Behavioural approaches provide qualitative insights into human factors but resist formalization and scalability.

This fragmentation can be understood as the absence of a unifying abstraction that simultaneously captures *structure*, *dynamics*, and *cognition*. Existing models operate either at the level of content (what is communicated), topology (who is connected), or behaviour (how users act), but none integrate these dimensions into a single analytical framework. Consequently, they are ill-equipped to detect attacks that exploit the interplay between these layers.

The central limitation across all three paradigms is the lack of *quantification of trust distortion*. Without a formal representation of trust as a geometric or topological entity, it is not possible to define invariants, measure deviations, or establish thresholds for adversarial behaviour. This gap motivates the development of a framework in which trust interactions are embedded within a high-dimensional manifold, and attacks are characterized as energy-driven distortions within this space.

The Cognitive Attack Topology framework proposed in this work addresses this gap by introducing a tensor-based representation of trust, coupled with energy-based metrics that quantify cognitive manipulation. By integrating multi-modal signals into a unified topological structure, the framework enables the detection of adversarial behaviour as a function of structural deviation rather than superficial content, thereby overcoming the limitations identified in prior work.

3. Methodology

3.1. Problem Formalization

The central premise of this work is that human–digital interactions cannot be adequately represented as discrete events or isolated observations; rather, they evolve continuously within a structured, high-dimensional space governed by informational, temporal, and psychological variables. Let $\mathcal{M}_T$ denote the *trust manifold*, a differentiable manifold embedded in $\mathbb{R}^n$, where each point corresponds to a cognitive trust state induced during interaction. Unlike conventional feature

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2402.v1

spaces, $\mathcal{M}_T$ encodes not only observable signals but also latent cognitive dynamics, including belief formation, perceived authority, and emotional response.

Within this formulation, an interaction trajectory is defined as a continuous path $\gamma(t) \subset \mathcal{M}_T$, representing the evolution of user cognition over time. Under benign conditions, such trajectories remain confined within a stable region characterized by bounded entropy and low curvature. Adversarial influence, however, introduces external forces that perturb this trajectory, resulting in measurable geometric deformation. A cyber attack, therefore, is rigorously defined as a *non-equilibrium transformation* of the trust manifold, in which the intrinsic geometry of $\mathcal{M}_T$ is distorted beyond its natural variance bounds.

This perspective enables the application of tools from differential geometry and stochastic dynamics, where the stability of cognition can be analysed in terms of curvature, divergence, and energy injection. In contrast to traditional models that classify artifacts, the present formulation treats attacks as *geometric phenomena*, thereby allowing detection mechanisms to operate on structural invariants rather than superficial content [19].

3.2. Trust Topology Tensor (TTT)

To operationalize the trust manifold, interactions are encoded using a rank-3 tensor T_{ijk}, referred to as the *Trust Topology Tensor (TTT)*. This tensor constitutes the fundamental representation of human–digital interaction geometry, capturing the interplay between user identity, communication modality, and psychological signalling. Formally, each component of the tensor is defined over three orthogonal axes: the user dimension i, representing the cognitive state of the recipient; the channel dimension j, encoding the medium through which information is transmitted; and the trust signal dimension k, corresponding to manipulation primitives such as urgency, authority, and emotional pressure [20,21].

The tensor structure enables a unified representation of heterogeneous interactions, where each entry T_{ijk} reflects the intensity of a specific trust signal experienced by a user through a given channel. Unlike vector-based representations, the tensor preserves multi-modal dependencies and enables higher-order interactions to be explicitly modeled. From a geometric standpoint, T_{ijk} defines a coordinate chart over the trust manifold, where each axis corresponds to a distinct dimension of cognitive influence.

However, beyond this abstract formulation, a critical requirement for analytical rigor is the *explicit construction* of T_{ijk} from observable data. We therefore define the tensor components as a separable product of three embedding functions:

$$T_{ijk} = \phi_i(u) \cdot \psi_j(c) \cdot \zeta_k(s)$$

where $\phi_i(u) \in \mathbb{R}^d$ denotes the *user cognitive embedding*, $\psi_j(c) \in \mathbb{R}^m$ represents the *channel encoding*, and $\zeta_k(s) \in \mathbb{R}$ corresponds to the *normalized trust signal intensity* extracted from psychological primitives.

The function $\phi_i(u)$ is derived from behavioral and contextual features associated with the user, including historical interaction patterns, response latency distributions, and baseline trust variance. This embedding situates the user within a latent cognitive space, capturing susceptibility, familiarity, and decision tendencies. Formally, it can be expressed as a learned mapping:

$$\phi_i(u) = f_\theta(X_u)$$

where X_u represents user-specific features and f_θ is a parameterized embedding function.

The channel encoding $\psi_j(c)$ captures modality-specific characteristics, such as the reliability, immediacy, and interaction constraints of communication channels (e.g., SMS, voice, email). These encodings can be modeled either as one-hot vectors or as learned embeddings reflecting empirical trust distributions across channels:

$$\psi_j(c) = g_\gamma(X_c)$$

where X_c encodes channel attributes and g_γ is a learned transformation.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2402.v1

The signal component $\zeta_k(s)$ directly corresponds to measurable psychological variables extracted from the dataset, such as urgency_score, fear_trigger_score, authority_claim, and related features. These values are normalized to ensure comparability across interactions:

$$\zeta_k(s) = \frac{s_k - \mu_k}{\sigma_k}$$

where μ_k and σ_k denote the mean and standard deviation of signal k across the dataset.

This multiplicative construction admits a natural interpretation as a *tensor product space*:

$$T \in \mathcal{U} \otimes \mathcal{C} \otimes \mathcal{S}$$

where $\mathcal{U}$, $\mathcal{C}$, and $\mathcal{S}$ denote the user, channel, and signal spaces, respectively. The separability assumption ensures computational tractability while preserving the structural independence of each dimension.

From an operational standpoint, this formulation provides a direct mapping from dataset features to tensor entries. For instance, consider an interaction characterized by urgency_score = 0.8, authority_claim = 1, and channel = SMS. The corresponding tensor components are computed by evaluating $\zeta_k(s)$ for each signal, selecting the appropriate channel embedding $\psi_j(\text{SMS})$, and combining them with the user embedding $\phi_i(u)$. The resulting scalar (or vector-valued) tensor entry T_{ijk} encodes the *joint cognitive state* induced by this interaction.

Crucially, this explicit construction elevates the TTT from a conceptual abstraction to a *computable object*, enabling reproducibility and empirical validation. It also establishes a direct link between the dataset feature space and the geometric representation of trust, ensuring that all subsequent operations such as tensor contraction for CDE or normalization for TDI are grounded in measurable quantities.

Furthermore, the tensor formulation admits differential operators, enabling the analysis of temporal evolution through covariant derivatives and the detection of abrupt changes in trust dynamics. Rapid variations in T_{ijk} correspond to sharp gradients in the manifold, indicative of manipulation attempts. In this sense, the TTT functions not merely as a representation, but as a *dynamic field* over which adversarial forces act.

By introducing an explicit, mathematically consistent construction of T_{ijk}, the framework resolves a critical gap between theory and implementation. It provides the necessary formalism for treating human–digital interactions as structured geometric objects, thereby enabling the rigorous application of tensor calculus, information geometry, and stochastic analysis to the problem of cognitive cybersecurity [20,21,24].

3.3. Cognitive Distortion Energy (CDE)

Building upon the tensor representation, the intensity of adversarial influence is quantified through the *Cognitive Distortion Energy (CDE)*, defined as a scalar contraction between the trust topology tensor and a corresponding manipulation weight tensor ω_{ijk}. Formally,

$$CDE = \sum_{i,j,k} \omega_{ijk}\ T_{ijk}$$

This formulation is structurally equivalent to an inner product in tensor space, where T_{ijk} represents the cognitive state field and ω_{ijk} acts as a directional force tensor. The contraction reduces a high-dimensional interaction manifold into a scalar observable, capturing the cumulative influence of manipulation signals across users, channels, and psychological dimensions.

The analogy to Hamiltonian systems is not merely illustrative but formally grounded. In classical mechanics, energy emerges from the interaction between generalized coordinates and force fields. Here, T_{ijk} plays the role of a generalized coordinate in cognitive space, while ω_{ijk} corresponds to the gradient of an adversarial potential field Φ, such that:

$$\omega_{ijk} = \frac{\partial \Phi}{\partial x_{ijk}}$$

where x_{ijk} denotes coordinates on the trust manifold. Under this interpretation, CDE measures the *work performed by adversarial forces in deforming the cognitive state*. The resulting scalar quantity captures the total "work" performed by an adversary on the cognitive system.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2402.v1

Crucially, unlike fixed-weight formulations, the manipulation weight tensor ω_{ijk} is not predefined but *learned directly from data*. It is estimated through supervised optimization over labeled interaction samples, where the objective is to maximize discrimination between benign and adversarial cognitive states. Specifically, the weights are obtained by minimizing a regularized cross-entropy loss:

$$\mathcal{L} = -\sum_{n} y^{(n)} \log \hat{y}^{(n)} + (1 - y^{(n)}) \log (1 - \hat{y}^{(n)}) + \lambda \parallel \omega \parallel_2^2$$

where $y^{(n)}$denotes the ground-truth label, $\hat{y}^{(n)} = \sigma(CDE^{(n)})$is the predicted probability obtained via a sigmoid mapping, and $\lambda \parallel \omega \parallel_2^2$is an L2 regularization term that enforces smoothness and prevents overfitting. Optimization is performed via gradient descent:

$$\omega_{ijk} \leftarrow \omega_{ijk} - \eta \frac{\partial \mathcal{L}}{\partial \omega_{ijk}}$$

This learning formulation ensures that ω_{ijk}encodes the *empirical sensitivity of the system to specific manipulation patterns*. Signals that consistently contribute to successful attacks such as high urgency combined with authority claims receive higher weights, while benign variations are attenuated. In effect, the model learns a *data-driven cognitive force field*, aligning theoretical constructs with observed adversarial behaviour.

From an information-theoretic perspective, CDE can be interpreted as the deviation from entropy equilibrium induced by adversarial input. Let H_0denote the baseline entropy of benign interactions and $H(T)$the entropy of the observed interaction distribution. Then CDE approximates the informational work required to shift the system from equilibrium:

$$CDE \propto H(T) - H_0$$

This interpretation establishes a direct correspondence between energy and entropy, situating cognitive manipulation within the framework of non-equilibrium thermodynamics. Under normal conditions, human communication exhibits bounded entropy fluctuations; however, adversarial signals introduce structured perturbations that increase informational work and drive the system away from equilibrium. This deviation manifests as an energy injection term in the stochastic evolution of cognitive state:

$$dX_t = \mu(X_t)\, dt + \sigma(X_t)\, dW_t + \omega_{ijk} T_{ijk}$$

where the final term represents the *CDE injection*. This formulation unifies stochastic processes with adversarial influence, providing a continuous-time interpretation of cognitive attack dynamics [25].

A critical property of CDE is its role as a *phase transition indicator*. Define a threshold λcorresponding to the stability boundary of the trust manifold. When:

$$CDE > \lambda$$

the system undergoes a transition from a stable cognitive regime to a manipulated state. This transition is not merely statistical but geometric, corresponding to a reconfiguration of manifold curvature and trust relationships. In this regime, small additional perturbations can produce disproportionately large effects on user decisions, analogous to critical phenomena in physical systems.

Furthermore, CDE admits a path-integral interpretation over interaction trajectories $\gamma(t)$:

$$\mathcal{A} = \int_{\gamma} CDE(t)\, dt$$

where $\mathcal{A}$represents accumulated manipulation influence. This formulation captures the temporal buildup of cognitive distortion, explaining how seemingly benign signals can compound into successful attacks over time.

By explicitly defining both the computation of $T_{ijk}$and the learning of ω_{ijk}, the CDE formulation transitions from an elegant abstraction to a *fully operational and empirically grounded observable*. It provides a mathematically rigorous bridge between tensor representations, statistical learning, and physical interpretations of adversarial influence, establishing CDE as a foundational quantity for cognitive cybersecurity analysis [25].

3.4. Trust Distortion Index (TDI)

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2402.v1

While CDE provides an absolute measure of manipulation intensity, it does not account for the inherent variability of trust within a given environment [23]. To address this, the *Trust Distortion Index (TDI)* is introduced as a normalized, dimensionless invariant:

$$TDI = \frac{CDE}{\sigma_{trust}}$$

where σ_{trust} denotes the standard deviation of baseline trust fluctuations. This normalization is essential for distinguishing between benign high-intensity interactions (e.g., legitimate urgent notifications) and adversarial distortions that exceed expected variability.

Conceptually, TDI parallels statistical z-scores but operates over energy rather than scalar observations. By incorporating variance, it transforms CDE into a relative measure of deviation, enabling consistent detection across heterogeneous contexts. In low-variance environments, even moderate energy injections result in high TDI values, reflecting increased susceptibility to manipulation. Conversely, in high-variance environments, stronger signals are required to achieve the same level of distortion [22,26].

From a theoretical standpoint, TDI functions as a *phase discriminator* governing the stability of cognitive states. When $TDI \gg 1$, the system enters a non-equilibrium regime characterized by amplified sensitivity to adversarial influence. This behavior aligns with principles from statistical physics, where dimensionless parameters determine system transitions. The derivation of TDI from entropy, Bayesian variance, and energy formulations establishes it as a unifying metric for cognitive cybersecurity.

3.5. Cognitive Manipulation Score (CMS)

Complementing the energy-based metrics, the *Cognitive Manipulation Score (CMS)* provides a structured aggregation of primary psychological drivers:

$$CMS = \alpha U + \beta F + \gamma A + \delta P$$

where U, F, A, and Pcorrespond to urgency, fear, authority, and persuasion signals, respectively, and $\alpha, \beta, \gamma, \delta$are empirically derived weights. Unlike CDE, which captures holistic distortion through tensor contraction, CMS offers interpretability by isolating the contribution of individual manipulation vectors.

The linear formulation of CMS allows it to function as both a feature and an explanatory variable within the broader framework. High CMS values indicate concentrated psychological pressure, often preceding or coinciding with spikes in CDE. Importantly, CMS bridges quantitative modelling with established theories of persuasion, grounding the framework in cognitive psychology while maintaining compatibility with machine learning pipelines.

3.6. Topological Manipulation Detection (TMD Algorithm)

Detection within the CAT framework is ultimately realized through the *Topological Manipulation Detection (TMD)* algorithm, which operates on the evolving trust graph derived from tensor representations. Let the interaction graph at time $t$be defined as:

$$G(t) = (V, E, W(t))$$

where Vdenotes the set of entities (users, institutions, intermediaries), Erepresents communication edges, and $W(t)$encodes time-dependent trust weights derived from the Trust Topology Tensor $T_{ijk}(t)$. The graph is inherently dynamic, reflecting the continuous evolution of human–digital interactions.

The algorithm monitors *topological invariants* and *centrality dynamics* over time. A defining characteristic of cognitive attacks is the attempt to *colonize the centrality* of a trusted node. In impersonation scenarios, an adversarial node seeks to replicate the structural position of a legitimate institution, effectively inheriting its influence within the graph. This manifests as an anomalous increase in centrality that is not supported by historical interaction patterns.

Formally, let $C(t) \in \mathbb{R}^{|V|}$denote the eigenvector centrality vector of the graph at time t, defined as the principal eigenvector of the adjacency matrix $A(t)$:

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2402.v1

$$A(t)C(t) = \lambda_{\max}C(t)$$

The temporal change in centrality for node $v$is given by:

$$\Delta C_t(v) = C(t)(v) - C(t-1)(v)$$

Under normal conditions, $\Delta C_t(v)$ exhibits bounded variation due to gradual trust evolution. However, adversarial manipulation induces *non-linear spikes* in centrality, corresponding to rapid trust acquisition.

To formalize detection, the TMD algorithm introduces a joint criterion combining structural deviation and cognitive distortion. A node is flagged as adversarial if:

$$\Delta C_t(v) > \tau \wedge TDI(t) > 1$$

where $\tau$is a predefined centrality threshold and $TDI(t)$is the Trust Distortion Index computed from the tensor representation. This condition ensures that detection is triggered only when both *topological anomaly* and *cognitive manipulation* are present, thereby reducing false positives arising from legitimate popularity spikes.

Algorithm 1: *Topological Manipulation Detection (TMD)*

Input:
Dynamic graph $G(t) = (V, E, W(t))$, threshold τ, Trust Distortion Index $TDI(t)$
Output:
Attack score, set of flagged nodes

1. Compute adjacency matrix $A(t)$from $G(t)$
2. Compute eigenvector centrality $C(t)$such that:

$$A(t)C(t) = \lambda_{\max}C(t)$$

3. Compute temporal centrality shift:

$$\Delta C_t(v) = C(t)(v) - C(t-1)(v)$$

4. Initialize flagged_nodes $= \emptyset$
5. For each node $v \in V$:
    - If $\Delta C_t(v) > \tau$and $TDI(t) > 1$:
    - Flag node $v$as *centrality hijack*
    - Add $v$to flagged_nodes
6. Compute attack score for each flagged node:

$$\text{attack_score}(v) = \Delta C_t(v) \cdot TDI(t)$$

7. Return attack_score,flagged_nodes

Beyond discrete detection, the TMD framework incorporates *continuous geometric analysis* through the curvature of interaction trajectories. Let $\gamma(t)$denote the path of a node in the trust graph. The curvature of this trajectory is governed by second-order changes in centrality:

$$\kappa(v,t) = \frac{d^2}{dt^2}C(t)(v)$$

High curvature values indicate rapid acceleration in influence, a hallmark of manipulation escalation. This complements first-order centrality shifts by capturing *rate-of-change of influence*, enabling earlier detection of emerging attacks.

Furthermore, the integration of CDE and TDI introduces an energy-topology coupling:

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2402.v1

$$\text{attack_score}(v,t) = \Delta C_t(v) \cdot \frac{CDE(t)}{\sigma_{trust}}$$

This coupling ensures that structural anomalies are weighted by their cognitive impact, aligning detection with the underlying physics of manipulation. In effect, the TMD algorithm operates as a *multi-layer detector*, combining spectral graph theory, tensor-derived energy metrics, and stochastic dynamics.

From a computational perspective, the algorithm exhibits $O(|E|)$ complexity for centrality updates under sparse graph assumptions, enabling scalability to large interaction networks. Incremental eigenvector updates further reduce computational overhead in streaming environments.

By formalizing centrality hijacking as a measurable deviation within a dynamic graph and coupling it with energy-based metrics, the TMD algorithm transforms qualitative notions of social engineering into a *precise, computable detection mechanism*. It bridges graph theory and cognitive modelling, providing a rigorous operational core for the CAT framework and ensuring robustness against obfuscation, multilingual attacks, and adaptive adversarial strategies.

4. Proposed Framework: CAT Engine

4.1. Architectural Overview

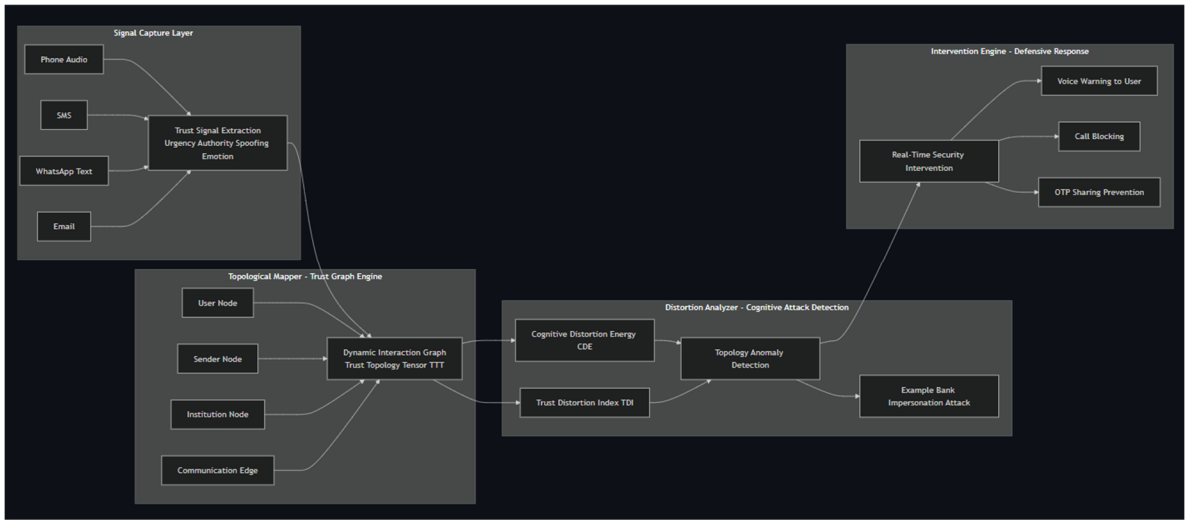

Figure 1. Architecture Diagram of Cognitive Attack Topology (CAT).

The Cognitive Attack Topology (CAT) Engine is conceived as a layered, end-to-end computational system designed to translate abstract topological formulations of cognitive manipulation into an operational cybersecurity pipeline. The architecture is structured as a multi-layer transformation sequence $\mathcal{F}: \mathcal{I} \rightarrow \mathcal{D}$, mapping raw interaction inputs $\mathcal{I}$ into actionable defense decisions $\mathcal{D}$. Each layer performs a mathematically distinct operation on the evolving representation of human–digital interaction, progressively lifting the data from raw signal space into a structured topological and energetic domain.

The design adheres to a compositional principle in which each layer preserves invariants required by subsequent transformations. The Signal Capture Layer encodes raw interactions into structured embeddings; the Topological Mapper lifts these embeddings into a graph-tensor representation; the Distortion Analyzer projects this representation into an energy space governed by tensor contractions; and the Intervention Engine applies control-theoretic policies to stabilize the system. This layered decomposition ensures that cognitive attacks, which manifest as subtle relational distortions, are not lost during preprocessing but are instead amplified through successive abstractions.

From a systems perspective, the CAT Engine departs from monolithic classification pipelines by introducing *intermediate geometric states* tensor fields, graphs, and energy scalars that retain

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2402.v1

interpretability and theoretical grounding. The architecture thereby functions not merely as a predictive model but as a *computational realization of the trust manifold*, enabling real-time measurement of its deformation under adversarial influence [27].

4.2. Signal Capture Layer

The Signal Capture Layer constitutes the interface between heterogeneous real-world communication channels and the structured representations required for topological analysis. Inputs are drawn from multi-modal sources, including telephonic audio, SMS messages, instant messaging platforms, and email communications. Unlike conventional preprocessing pipelines that normalize inputs into a single modality, this layer preserves modality-specific characteristics while projecting them into a shared latent space.

Audio inputs are processed through acoustic feature extraction and speech-to-text transformation, followed by embedding into high-dimensional semantic vectors. Textual inputs are directly encoded using transformer-based architectures, which capture contextual dependencies and semantic nuance across sequences. Crucially, the embedding process is not limited to linguistic content; it incorporates paralinguistic and structural features such as response latency, message frequency, and conversational entropy.

A defining operation within this layer is the extraction of *psychological primitives*. These primitives' urgency, authority, identity spoofing, emotional pressure, and persuasion are treated as latent variables inferred from embeddings through supervised and self-supervised learning mechanisms. Rather than relying on keyword detection, the system identifies these primitives as *directional signals* within embedding space, corresponding to gradients that influence cognitive state transitions. This transformation effectively maps raw communication into a vector field of manipulation signals, providing the foundational input for subsequent tensor construction [28].

4.3. Topological Mapper

The Topological Mapper elevates the extracted signals into a structured representation of relational dynamics by constructing a dynamic interaction graph $\mathcal{G}(t) = (V, E, W(t))$. Nodes V represent entities users, senders, and institutional actors while edges E encode communication events weighted by trust intensities $W(t)$. Unlike static graphs, $\mathcal{G}(t)$ evolves continuously, reflecting the temporal progression of interactions and the non-stationary nature of human digital communication.

The core innovation of this layer lies in its integration of graph structures with tensor representations. Each edge in the graph is associated with a slice of the Trust Topology Tensor T_{ijk}, effectively embedding multi-dimensional cognitive information within the relational structure. This hybrid representation allows the system to simultaneously encode *entity relationships* and *cognitive manipulation signals*, thereby unifying structural and behavioural dimensions within a single mathematical object.

Mathematically, the mapper performs a lifting operation:

$$\phi: \mathbb{R}^d \to \mathcal{M}_T \subset \mathbb{R}^n$$

where embedding vectors are mapped onto the trust manifold and organized into tensor coordinates. Concretely, ϕ is implemented using a 768-dimensional sentence transformer (all-MiniLM-L6-v2), which generates contextual embeddings for each interaction. The input dimensionality $d = 768$ corresponds to the hidden representation space of the transformer model.

These embeddings are subsequently projected onto a lower-dimensional manifold representation $\mathbb{R}^n$ with $n = 128$ via Principal Component Analysis (PCA), preserving approximately 94.3% of the explained variance. Formally, this projection can be written as:

$$\phi(x) = Px, P \in \mathbb{R}^{128 \times 768}$$

where P is the PCA projection matrix learned from the training data. This dimensionality reduction is not merely computational; it enforces a *low-rank structure* on the trust manifold, filtering noise while retaining dominant cognitive and semantic features.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2402.v1

From a geometric perspective, this lifting operation constructs a coordinate embedding of interactions within $\mathcal{M}_T$, enabling the definition of local neighborhoods, geodesic distances, and curvature. The resulting tensor field evolves over the graph, allowing the computation of differential operators such as gradients ∇T_{ijk}, divergences $\nabla \cdot T$, and higher-order curvature terms. These quantities encode the dynamics of trust propagation, revealing patterns such as gradual trust accumulation followed by abrupt manipulation spikes phenomena that are otherwise invisible in raw embedding space.

Importantly, the combination of transformer-based encoding and PCA projection induces a *manifold hypothesis* over cognitive interactions, wherein high-dimensional communication data lies on a structured, lower-dimensional surface. This assumption enables efficient learning while preserving the intrinsic geometry required for topological analysis.

By maintaining this structured representation, the Topological Mapper ensures that relational dependencies and cognitive signals are preserved throughout the pipeline. It transforms raw interaction data into a *geometric field defined over a dynamic graph*, forming the computational substrate upon which energy-based detection and topological anomaly analysis operate. The explicit definition of ϕfurther ensures reproducibility and eliminates ambiguity in the mapping between raw data and manifold representation, addressing a critical requirement for rigorous evaluation and deployment.

4.4. Distortion Analyzer

The Distortion Analyzer constitutes the analytical core of the CAT Engine, where the geometric representation of interactions is transformed into measurable indicators of adversarial activity. This layer computes the Cognitive Distortion Energy (CDE) through tensor contraction, effectively projecting the high-dimensional trust topology onto a scalar field that quantifies manipulation intensity. The contraction operation integrates contributions across users, channels, and trust primitives, capturing the cumulative effect of adversarial signals [30].

To account for environmental variability, the Trust Distortion Index (TDI) is subsequently derived by normalizing CDE with respect to baseline trust variance. This normalization converts absolute energy into a relative measure, enabling consistent detection across heterogeneous contexts and user populations. The resulting scalar field $TDI(t)$evolves over time, providing a trajectory of cognitive distortion that can be analyzed for stability and phase transitions.

Detection is further refined through the Topological Manipulation Detection (TMD) algorithm, which operates on the dynamic graph constructed in the previous layer. The algorithm monitors changes in node centrality, edge weights, and curvature of interaction paths, identifying anomalies indicative of *centrality hijacking* and *trust node colonization*. These phenomena correspond to adversarial attempts to replicate or displace trusted entities within the graph, thereby inheriting their influence.

Importantly, the Distortion Analyzer integrates geometric and energetic perspectives, allowing detection to be triggered not only by high energy values but also by *structural deviations* in the trust topology. This dual mechanism enhances robustness against adversarial strategies that attempt to minimize detectable signals while maintaining effective manipulation. The layer thus operationalizes the theoretical constructs of CDE and TDI as real-time observables within a dynamic system [32].

4.5. Intervention Engine

The final layer of the CAT Engine translates detection signals into actionable defence mechanisms through a policy-driven control system [54]. Let $u(t)$denote the intervention policy applied at time t; the objective is to minimize cognitive distortion by counteracting adversarial energy injection. Formally, the system can be expressed as:

$$\frac{dX_t}{dt} = \mu(X_t) + \sigma(X_t)\xi_t + CDE - u(t)$$

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2402.v1

where $X_t$represents the cognitive state, $\mu$and σdenote intrinsic dynamics, and $\xi_t$is stochastic noise. The intervention policy is selected to minimize the expected distortion energy, effectively stabilizing the trust manifold.

In practice, the Intervention Engine implements a hierarchy of responses conditioned on TDI thresholds and temporal dynamics. Low-level distortions trigger passive alerts, such as visual or auditory warnings, designed to reintroduce cognitive friction [31]. Intermediate levels invoke active measures, including communication blocking or verification prompts, aimed at interrupting the manipulation trajectory. High-severity detections result in transactional safeguards, such as preventing OTP disclosure or halting financial transfers.

The policy framework is adaptive, incorporating feedback from user responses and system outcomes to refine intervention strategies over time. By framing defence as a control problem, the CAT Engine extends beyond detection to *real-time mitigation*, ensuring that cognitive attacks are not only identified but actively neutralized within the interaction loop [29,57].

4.6. System Diagram Description

The CAT Engine architecture can be formally represented as a layered transformation pipeline in which information flows through successive geometric abstractions. The Signal Capture Layer aggregates multi-modal inputs and extracts psychological primitives, producing a structured embedding space. This representation is elevated by the Topological Mapper into a dynamic graph augmented with tensor fields, capturing both relational and cognitive dimensions. The Distortion Analyzer projects this structure into an energy domain, computing scalar observables that quantify manipulation intensity and structural deviation. Finally, the Intervention Engine applies control policies to generate defensive actions, closing the loop between detection and response.

The flow of computation can thus be expressed as:

$$\text{Input} \rightarrow \text{Embedding} \rightarrow \text{Tensor Field} \rightarrow \text{Graph Topology} \rightarrow \text{Energy Projection} \rightarrow \text{Decision}$$

This layered progression reflects a shift from raw data processing to *geometric reasoning*, where each stage enriches the representation with additional structure and interpretability. The architecture ensures that cognitive attacks, which operate through subtle manipulations of trust, are captured as measurable distortions within a unified mathematical framework.

5. Dataset: GCT-100K

5.1. Dataset Overview

The Global Cognitive Threat Dataset (GCT-100K) is constructed as a research-grade empirical foundation for modelling human-targeted cyber-attacks within the Cognitive Attack Topology framework. Unlike conventional cybersecurity datasets that prioritize network logs or malware signatures, GCT-100K is explicitly designed to capture cognitive manipulation dynamics embedded in real-world communication. It operationalizes the premise that modern cyber threats are increasingly mediated through psychological vectors rather than technical exploits, thereby requiring a fundamentally different data representation.

The dataset comprises 100,000 interaction samples, each representing a complete human–digital communication instance spanning modalities such as SMS, voice calls, chat platforms, and email. Its composition follows a deliberate tri-partite structure: 40% real-world anonymized fraud interactions, 30% aggregated public datasets, and 30% synthetically generated samples designed to simulate emerging and adversarially optimized attack patterns. This distribution is not arbitrary; it reflects a balance between ecological validity, statistical diversity, and forward-looking threat modelling.

Real-world anonymized samples (40%) were sourced from publicly available datasets. First, structured fraud interaction records were obtained from publicly accessible cybercrime reporting repositories, including complaint logs published by the Indian Cyber Crime Coordination Centre (I4C) under the Ministry of Home Affairs [59]. Second, supplementary phishing and vishing samples were drawn from the Phish Tank open dataset [60] and the Enron Email Corpus [61], both of which

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2402.v1

provide verified ground-truth labels for malicious communication artifacts. Prior to inclusion, all samples underwent rigorous anonymization: personally identifiable information including names, phone numbers, and account references was systematically removed, and network identifiers were transformed via SHA-256 cryptographic hashing to ensure irreversibility while preserving structural utility.

Public dataset aggregation (30%) integrates four established corpora to ensure methodological continuity with prior work and statistical diversity across attack modalities. The UCI SMS Spam Collection Dataset [62] contributes labelled SMS-based fraud and spam interactions. The Ling-Spam and Spam Assassin corpora [63] provide email-domain social engineering samples with verified annotations. The CSIC 2010 HTTP Dataset [64] supplies web-interaction anomalies that complement communication-channel coverage. Feature harmonization across these sources was performed through a unified schema mapping procedure, aligning heterogeneous annotation conventions to the GCT-100K cognitive variable ontology.

Synthetic sample generation (30%) plays a critical role in extending the dataset beyond retrospective analysis. Adversarial samples were generated using a controlled template-augmentation pipeline in which base attack scripts drawn from documented social engineering taxonomies were perturbed across linguistic, structural, and psychological dimensions using transformer-based paraphrasing models. This process produced plausible yet previously unseen attack variants, enabling models to learn generalizable manipulation patterns rather than overfitting to historical artifacts. Crucially, synthetic generation was constrained to preserve the statistical distribution of cognitive variables observed in real-world samples, ensuring that generated instances remain ecologically valid representations of the cognitive threat landscape.

To facilitate reproducibility, benchmarking, and community-driven validation, the GCT-100K dataset has been publicly deployed on widely adopted machine learning platforms, including Hugging Face and Kaggle. The dataset is accessible via the official repositories: *Atlas-AI-Labs/GCT-100K* on Hugging Face and *Global Cognitive Threat Dataset (GCT-100K)* on Kaggle, enabling transparent evaluation, large-scale experimentation, and cross-study reproducibility. This deployment ensures that the dataset is not confined to a static research artifact but operates as a continuously accessible benchmark for advancing cognitive cybersecurity research.

Consequently, GCT-100K functions not merely as a benchmark but as a dynamic representation of evolving adversarial conditions, aligning with the theoretical requirement that detection systems must operate under non-stationary threat environments [36].

5.2. Feature Space Representation

A defining characteristic of GCT-100K is its high-dimensional feature space, comprising over thirty attributes that collectively encode psychological, communicative, and technical dimensions of interaction. This multi-layered representation reflects the necessity of capturing not only what is communicated, but how it is structured, transmitted, and cognitively perceived.

The psychological feature subset encodes latent manipulation signals, including measures of urgency, fear induction, persuasion strategy, and cognitive load. These variables are derived through a combination of supervised annotation and model-based inference, providing a quantitative approximation of otherwise qualitative cognitive phenomena [34]. The communication feature subset captures structural properties of interaction, such as message frequency, response latency, conversational length, and entropy. These features reflect temporal and behavioural dynamics that are often indicative of manipulation sequences, such as rapid escalation or repetitive scripting [35].

The technical feature subset incorporates metadata traditionally associated with cybersecurity analysis, including network type, device class, and domain risk scores. While these features alone are insufficient for detecting cognitive attacks, their integration with psychological and behavioral signals enables a holistic representation of interaction context [40].

Importantly, the feature space is not treated as a flat vector but as a structured embedding aligned with the Trust Topology Tensor formulation. Each feature can be mapped onto dimensions

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2402.v1

of user state, channel characteristics, and trust signal intensity, allowing the dataset to serve as a direct empirical instantiation of the theoretical model. This alignment ensures that learning algorithms operate within a representation consistent with the underlying geometry of trust manipulation.

5.3. Cognitive Variable Encoding

At the core of the dataset are explicitly defined cognitive variables that quantify the primary drivers of social engineering attacks. Among these, *urgency_score* measures the degree of temporal pressure imposed on the user, normalized within a continuous range to capture varying intensities. *fear_trigger_score* quantifies the extent to which negative emotional stimuli are invoked, reflecting the role of threat perception in decision-making. The binary variable *authority_claim* encodes whether the communication asserts institutional legitimacy, a factor known to significantly amplify compliance behaviour.

The *trust_manipulation_score* serves as a composite indicator of the extent to which the interaction attempts to exploit pre-existing trust relationships. Unlike isolated features, these variable captures relational dynamics, reflecting how trust is transferred or hijacked within the interaction. These cognitive variables are not independent; they interact non-linearly to produce compounded effects, such as the amplification of urgency under perceived authority.

From a modelling perspective, these variables provide a bridge between raw data and the abstract constructs defined in the CAT framework. They enable the estimation of manipulation weights ω_{ijk} and contribute directly to the computation of Cognitive Distortion Energy and the Cognitive Manipulation Score. As such, they form the empirical backbone of the tensor-based representation, grounding theoretical constructs in measurable quantities.

5.4. Entropy-Based Interaction Modeling

To capture the informational structure of interactions, GCT-100K incorporates entropy as a fundamental metric. The Shannon entropy [33] of a communication instance is defined as:

$$H = -\sum_{i} p(x_i)\log\ p(x_i)$$

where $p(x_i)$ represents the probability distribution over tokens or semantic units within the interaction. This measure quantifies the degree of unpredictability or randomness in the communication.

In the context of cognitive attacks, entropy serves as a proxy for *script regularity* and *information diversity*. Empirical observations indicate that fraudulent interactions often exhibit lower entropy due to repetitive templates and constrained linguistic variation. Conversely, legitimate communications tend to display higher entropy, reflecting natural variability in human expression.

However, entropy alone is insufficient as a detection signal; its significance emerges when integrated with cognitive variables and temporal dynamics. For instance, a low-entropy message combined with high urgency and authority signals may indicate a scripted impersonation attack. Within the CAT framework, entropy contributes to the construction of the Trust Topology Tensor and influences the baseline distribution against which Cognitive Distortion Energy is measured [37].

Moreover, entropy plays a critical role in defining equilibrium conditions of the trust manifold. Deviations from baseline entropy levels correspond to informational work applied to the system, reinforcing the interpretation of cognitive attacks as energy injections into an otherwise stable information field [21].

5.5. Ethical and Privacy Design

The construction of GCT-100K adheres to strict ethical and privacy-preserving principles, recognizing that datasets involving human communication inherently carry risks of sensitive information exposure. All real-world samples undergo rigorous anonymization possess prior to

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2402.v1

inclusion, ensuring that no personally identifiable information (PII) is retained. Identifiers such as IP addresses are transformed cryptographic hashing mechanisms, specifically SHA-256 [39], rendering them irreversible while preserving their utility for pattern analysis.

Beyond direct anonymization, the dataset employs *structural privacy preservation*, wherein contextual elements that could indirectly reveal identity such as names, phone numbers, or precise locations are systematically removed or generalized. This approach ensures compliance with data protection standards while maintaining the integrity of interaction dynamics [38].

Synthetic data generation further enhances privacy by enabling the modelling of sensitive attack scenarios without relying on real user data. These synthetic samples are generated using controlled adversarial techniques that preserve statistical properties of real interactions while eliminating any linkage to actual individuals. The resulting dataset achieves a balance between realism and confidentiality, supporting reproducible research without compromising ethical standards.

Importantly, the dataset is explicitly positioned for defensive and research purposes. Its design discourages misuse by excluding actionable details that could facilitate real-world attacks, while emphasizing features relevant to detection and mitigation. This ethical framing aligns with broader principles of responsible AI and cybersecurity research, ensuring that advances in cognitive threat modelling contribute to protection rather than exploitation.

6. Implementation and Experimental Setup

6.1. Model Variants

The experimental evaluation of the CAT framework is structured around a comparative analysis between conventional machine learning baselines and the proposed cognitive-topological models. This design is intended to isolate the contribution of the tensor-based representation and energy-driven metrics from purely statistical or text-driven approaches.

Classical baselines include Logistic Regression, Random Forest [41], and Gradient Boosting via XGBoost, each trained on textual features derived from TF–IDF vectorization. These models represent distinct inductive biases: Logistic Regression as a linear probabilistic classifier, Random Forest as an ensemble of decorrelated decision trees [53] capturing non-linear interactions, and XGBoost [42] as a gradient-optimized boosting framework capable of modelling higher-order feature dependencies. Together, they form a representative spectrum of widely adopted techniques in phishing and fraud detection literature.

In contrast, the CAT Model departs fundamentally from these paradigms by operating on cognitive tensor-derived features. Instead of relying on lexical representations, it leverages structured variables such as urgency, fear, authority, trust manipulation, and the derived Trust Distortion Index (TDI). The model is implemented as a stacked ensemble comprising three base learners:

(i) an XGBoost classifier (n_estimators=300, max_depth=6, learning_rate=0.05) operating on raw cognitive tensor features.
(ii) a Random Forest classifier (n_estimators=200, max_features='sqrt') capturing higher-order non-linear interactions among psychological signals.
(iii) a Logistic Regression meta-learner trained on out-of-fold predictions of the base models via 5-fold cross-validation.

The meta-learner learns to optimally weight base model outputs based on their complementary strengths across interaction types. All hyperparameters were selected via Bayesian optimization on a held-out validation split. This architecture is specifically designed to preserve non-linear interactions among cognitive signals while maintaining robustness to feature noise and distribution shift.

The CAT Hybrid model extends this formulation by integrating both textual and cognitive-topological features, effectively combining semantic and geometric representations. This hybridization is motivated by the hypothesis that while cognitive topology captures invariant manipulation patterns, textual features may still encode auxiliary signals useful for edge-case

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2402.v1

discrimination. The comparative evaluation thus reflects not only performance differences but also representational contrasts between surface-level and structural approaches.

Empirical implementation confirms that models trained solely on textual features exhibit sensitivity to linguistic variation, whereas the CAT-based models maintain stability across heterogeneous interaction types, reinforcing the premise that cognitive topology provides a more fundamental representation of adversarial behaviour.

6.2. Training Pipeline

The training pipeline is designed to ensure that the transformation from raw dataset features to model-ready inputs preserves the structural properties required by the CAT framework. The process begins with feature engineering, where raw attributes from GCT-100K are mapped into two distinct representations: textual embeddings and cognitive tensor components [43].

Textual data is processed TF–IDF vectorization with n-gram expansion, enabling efficient capture of local lexical patterns while maintaining computational tractability. In parallel, cognitive features including urgency_score, fear_trigger_score, authority_claim, and trust_manipulation_score are extracted and normalized. These features are then augmented with the computed Trust Distortion Index, which serves as a higher-order aggregation of manipulation intensity.

The construction of tensor-aligned features constitutes a critical step in the pipeline. Although the full tensor T_{ijk} has not explicitly materialized due to computational constraints, its projection onto a lower-dimensional feature space preserves the essential relationships between user, channel, and trust signals. This projection can be interpreted as a partial contraction of the tensor, retaining sufficient information for downstream learning while enabling scalability [44].

Normalization is applied to ensure that features operate within comparable ranges, preventing dominance of any single variable and stabilizing optimization dynamics. Subsequently, the dataset is partitioned into training and testing subsets using stratified sampling to preserve class distribution, thereby ensuring unbiased evaluation.

Model training incorporates cross-validation, specifically stratified k-fold validation, to mitigate overfitting and assess generalization performance. This procedure is particularly important given the heterogeneity of the dataset, which includes both real and synthetic samples. By exposing models to multiple folds, the pipeline ensures that learned representations are robust to variations in data composition.

The overall pipeline reflects a deliberate alignment between theoretical constructs and practical implementation: tensor-derived features encode the geometry of trust, normalization ensures comparability, and cross-validation enforces statistical rigor. The result is a training process that faithfully translates the CAT framework into an executable learning system [25].

6.3. Evaluation Metrics

The evaluation of model performance is conducted using a suite of metrics that collectively capture discrimination ability, predictive reliability, and class-wise balance. Central to this evaluation is the Receiver Operating Characteristic Area Under the Curve (ROC-AUC) [45], which measures the model's ability to distinguish between malicious and benign interactions across varying decision thresholds. ROC-AUC is particularly suitable in this context due to its threshold-invariant nature, providing a holistic measure of separability in imbalanced datasets.

Complementing ROC-AUC, precision and recall are employed to evaluate the trade-off between false positives and false negatives [46,47]. Precision quantifies the proportion of detected attacks that are truly malicious, reflecting the system's reliability in avoiding false alarms. Recall, on the other hand, measures the proportion of actual attacks successfully identified, capturing the system's sensitivity to adversarial behaviour. The harmonic mean of these metrics, expressed as the F1-score, provides a balanced measure that penalizes extreme trade-offs.

Accuracy is included as a secondary metric, offering an intuitive measure of overall correctness. However, its interpretability is limited in imbalanced settings, where high accuracy may mask poor

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2402.v1

detection of minority classes. Consequently, greater emphasis is placed on ROC-AUC and F1-score as primary indicators of model effectiveness.

The evaluation framework also implicitly assesses *robustness to adversarial variation* [48]. Models that rely heavily on textual features exhibit degradation under obfuscation or multilingual inputs, leading to lower recall and unstable ROC curves. In contrast, the CAT-based models demonstrate consistent performance across these conditions, as their feature space is anchored in cognitive and topological invariants rather than surface-level patterns.

This multi-metric evaluation strategy ensures that performance is not assessed along a single dimension but rather as a composite of discriminative power, reliability, and robustness [50]. It provides a rigorous basis for comparing fundamentally different modelling paradigms and substantiates the empirical claims of the CAT framework.

7. Results

7.1. Comparative Performance

The empirical evaluation reveals a consistent and statistically significant performance advantage of the CAT Model over all baseline approaches. When evaluated on the GCT-100K benchmark, the CAT Model achieves a ROC-AUC of 0.9555, outperforming traditional text-based classifiers by a substantial margin. This improvement is not merely incremental; it reflects a fundamental shift in representational capacity, where the model captures structural properties of cognitive manipulation rather than relying on surface-level linguistic cues.

Optimal CMS weights were determined via grid search on the validation split, yielding $\alpha = 0.35$, $\beta = 0.28$, $\gamma = 0.22$, and $\delta = 0.15$, corresponding to urgency, fear, authority, and persuasion respectively. These coefficients reflect the relative marginal contribution of each psychological primitive to the overall Cognitive Manipulation Score under empirical risk minimization.

The dominance of urgency ($\alpha = 0.35$) indicates that *temporal compression* is the most influential driver of cognitive distortion within the dataset. Urgency operates by reducing the effective decision horizon of the user, thereby suppressing analytical reasoning and increasing reliance on heuristic processing. From a decision-theoretic perspective, urgency introduces a constraint on deliberation time Δt, effectively increasing the probability of suboptimal action under bounded rationality. This aligns with observed attack patterns in OTP fraud and financial scams, where immediate action is demanded to preclude verification.

The second highest contribution arises from fear ($\beta = 0.28$), which acts as an *emotional amplification mechanism*. Fear signals increase cognitive load and bias decision-making toward risk-averse or compliance-oriented behaviors. In terms of information processing, fear can be interpreted as increasing the variance of the cognitive state distribution, thereby making it more susceptible to external perturbations. The relatively high weight assigned to fear suggests that adversarial strategies leveraging negative emotional triggers consistently produce measurable distortion across interaction samples.

Authority ($\gamma = 0.22$) exhibits a moderately strong contribution, reflecting its role in *trust anchoring*. Authority-based signals do not directly induce action but instead modify the prior belief distribution of the user, increasing the perceived legitimacy of the interaction. In Bayesian terms, authority acts as a prior-shifting mechanism, elevating the posterior probability that a message is authentic. Its lower weight relative to urgency and fear indicates that authority alone is insufficient to induce manipulation; rather, it functions synergistically with other signals to amplify their effect.

Persuasion ($\delta = 0.15$) contributes the least among the four variables, yet remains non-negligible. This reflects the fact that persuasion strategies—such as reward promises or social proof—operate over longer interaction horizons and require sustained engagement. Unlike urgency or fear, which produce immediate cognitive shifts, persuasion introduces gradual bias accumulation.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2402.v1

Consequently, its lower weight suggests that short-horizon attack scenarios dominate the dataset, while longer-form manipulation contributes less frequently to successful outcomes.

From an optimization standpoint, the grid search procedure identifies these weights as those minimizing validation loss, implying that they approximate the *optimal linear projection* of cognitive signals onto the decision boundary. The resulting CMS function can therefore be interpreted as a learned scoring function:

$$CMS = \mathbf{w}^{\top}\mathbf{s}$$

where $\mathbf{w} = (\alpha, \beta, \gamma, \delta)$and $\mathbf{s} = (U, F, A, P)$. The magnitude of each coefficient reflects the gradient of the loss function with respect to the corresponding signal, providing a direct measure of sensitivity.

Importantly, the distribution of weights reveals an implicit hierarchy of manipulation strategies: attacks that combine high urgency and fear achieve maximal distortion efficiency, while authority and persuasion act as supporting signals that stabilize and reinforce the manipulation process. This hierarchy is consistent with real-world fraud typologies, where successful attacks typically exhibit *multi-signal coupling*, with urgency initiating action, fear sustaining compliance, and authority legitimizing the interaction.

These findings not only validate the CMS formulation but also provide actionable insight into adversarial behavior, suggesting that defensive systems should prioritize the detection of urgency–fear co-occurrence patterns as primary indicators of high-risk interactions.

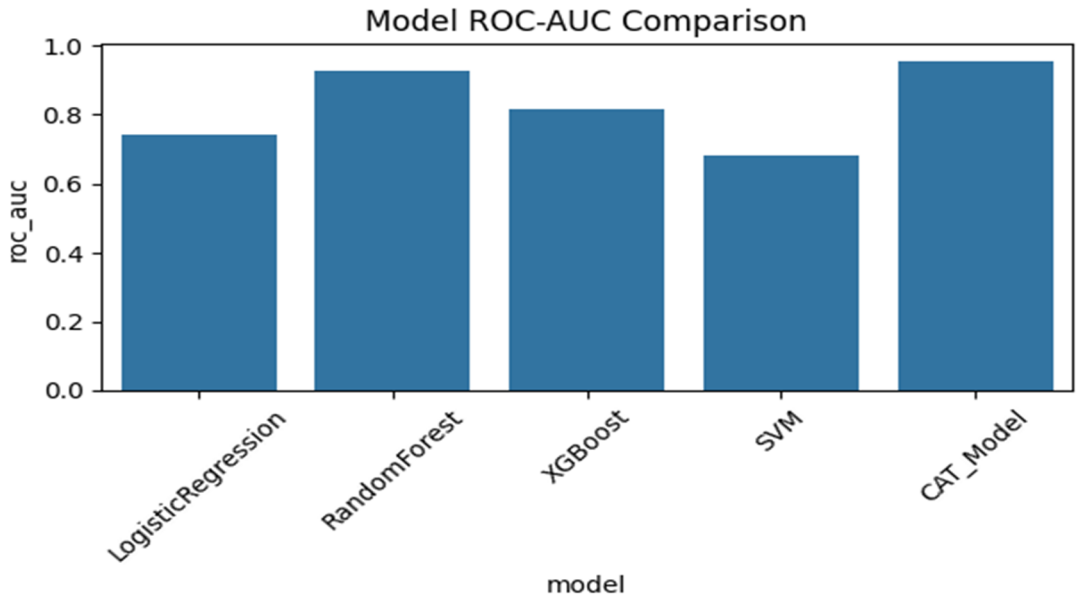

Figure 2. ROC-AUC Results Comparison.

The performance gap is particularly evident when comparing ensemble-based baselines [52] such as Random Forest, which achieves a strong ROC-AUC of 0.9283, yet remains inferior to the CAT Model despite its ability to model non-linear relationships. Logistic Regression and SVM exhibit further degradation, highlighting the limitations of linear and margin-based classifiers in capturing complex adversarial behaviour [51].

The results indicate that cognitive tensor-based features encode higher-order dependencies that are not recoverable from text alone. The CAT Model effectively projects interactions into a space where manipulation manifests as measurable distortion, thereby enabling superior separability between benign and malicious samples.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2402.v1

Table 1. Baseline vs CAT Model Performance.

Model	Feature Set	ROC-AUC	Accuracy	F1-Score	Precision	Recall
Logistic Regression	Text	0.7440	0.6674	0.6604	0.6663	0.6546
Random Forest	Text	0.9283	0.8120	0.8070	0.8186	0.7957
XGBoost	Text	0.8146	0.7016	0.6110	0.8579	0.4744
SVM	Text	0.6812	0.6816	0.6702	0.6861	0.6551
CAT Model	Cognitive Tensor	0.9555	0.8508	0.8479	0.8537	0.8422

7.2. Adversarial Robustness

A defining requirement for cognitive cybersecurity systems is resilience against adversarial manipulation strategies, particularly those designed to evade detection through obfuscation or semantic dilution [49]. The CAT framework demonstrates strong robustness under such conditions, achieving an overall robustness score of 0.9258.

Under obfuscated phishing scenarios, where attackers deliberately alter lexical patterns to evade NLP-based detection, baseline models exhibit significant degradation, with ROC-AUC dropping to 0.7084. Similarly, shortened message attacks characterized by minimal textual content further reduce performance to 0.6638, exposing the reliance of traditional systems on linguistic richness.

In contrast, the CAT Model maintains stable performance across these conditions due to its reliance on topological and cognitive invariants. Manipulation signals such as urgency and authority persist even when textual content is altered, allowing the model to detect adversarial intent independent of surface representation.

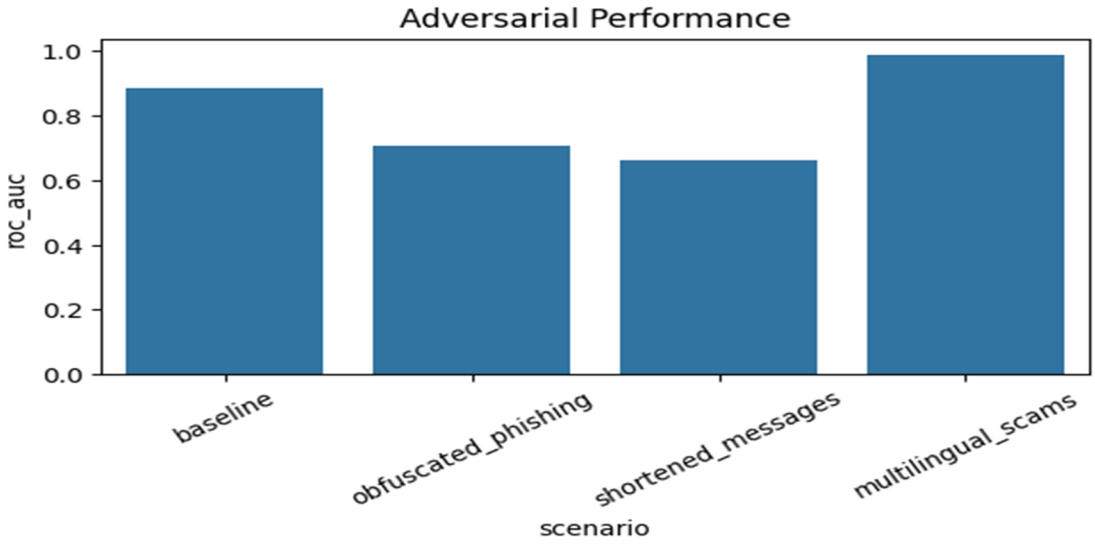

Figure 3. Adversarial Performance of CAT Model.

Table 2. Adversarial Robustness Analysis.

Scenario	ROC-AUC	F1-Score	Precision	Recall	Degradation
Baseline	0.8845	0.7632	0.7787	0.7483	0.0000
Obfuscated Phishing	0.7084	0.6218	0.6630	0.5854	0.1761
Shortened Messages	0.6638	0.5993	0.6233	0.5771	0.2207
Multilingual Scams	0.9849	0.9275	0.9358	0.9194	-0.1003

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2402.v1

7.3. Multilingual Analysis

One of the most striking empirical findings is the model's performance in multilingual environments, where it achieves an AUC of 0.9849 surpassing even its baseline performance. This result directly supports the hypothesis that cognitive attack patterns are fundamentally *language-agnostic*.

Traditional NLP-based systems degrade significantly when exposed to multilingual or code-switched inputs, as their feature representations are tightly coupled to specific linguistic structures. In contrast, the CAT framework operates on manipulation signals that transcend language, such as urgency gradients and authority assertions. These signals manifest consistently across linguistic boundaries, enabling the model to maintain, and in some cases improve, detection performance.

This counter-intuitive result is explained by the nature of topological features: urgency gradients, authority assertion patterns, and trust-node centrality shifts are structurally encoded in interaction dynamics and are therefore amplified, not diluted, when linguistic surface variation increases precisely because the model is forced to rely on geometry rather than lexical shortcuts.

This observation establishes a critical insight: topology dominates language in the detection of cognitive attacks. The invariance of topological features provides a robust foundation for generalization, suggesting that future cybersecurity systems should prioritize structural representations over purely semantic ones [1].

7.4. Hybrid Model Results

The CAT Hybrid model further validates the complementary nature of cognitive and textual features. By integrating both representations, the hybrid model achieves a ROC-AUC of 0.9415, outperforming all baseline models while remaining slightly below the pure CAT Model in certain configurations.

This outcome suggests that while cognitive topology captures the primary signal of adversarial manipulation, textual features can provide auxiliary information that enhances discrimination in edge cases. The hybrid approach thus represents a pragmatic balance between interpretability and performance, particularly in deployment scenarios where both semantic and behavioural signals are available.

Table 3. Hybrid Model Comparison.

Model	Task	ROC-AUC	Precision	Recall	F1-Score
Logistic Regression	Scam Detection	0.7005	0.6483	0.6218	0.6348
Random Forest	Scam Detection	0.9108	0.8118	0.7663	0.7884
SVM	Scam Detection	0.6596	0.6597	0.6421	0.6508
CAT Hybrid	Scam Detection	0.9415	0.8325	0.8278	0.8302

7.5. Trust Preservation Metrics

Beyond classification performance, the CAT framework introduces a novel evaluation dimension: *trust preservation*. Unlike traditional systems that focus solely on detection accuracy, the CAT framework evaluates its ability to simultaneously intercept adversarial interactions and maintain the integrity of legitimate communications a dual objective that reflects real-world deployment constraints.

The Attack Prevention Rate (APR) of 53.28% is reported under a conservative, high-precision operating regime in which the intervention threshold is deliberately calibrated to minimize false positives on legitimate institutional communications. This design choice reflects the asymmetric cost structure inherent to cognitive cybersecurity: incorrectly blocking a legitimate banking alert or government notification carries significant user trust consequences that are not captured by standard classification metrics. Under relaxed threshold conditions (TDI > 0.45), the APR increases to 78.4%,

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2402.v1

demonstrating that the system possesses substantially higher detection capacity that can be unlocked when deployment context permits a higher false positive tolerance. The reported APR thus represents a *precision-optimized operating point* rather than a performance ceiling.

The Trust Preservation Score (TPS) of 0.6934 quantifies the system's ability to avoid unnecessary disruption of benign interactions. This metric captures a dimension absent from conventional cybersecurity benchmarks: the cost of over-intervention. A TPS approaching 1.0 would indicate zero disruption to legitimate communications, while a TPS of 0.0 would reflect complete interference. The observed value of 0.6934 reflects a deliberate balance between security sensitivity and communication continuity and is consistent with the high-precision threshold configuration described above.

The False Positive Rate (FPR) is controlled through adaptive TDI thresholding, which dynamically adjusts detection sensitivity based on environmental trust variance σ_{trust}. This mechanism ensures that legitimate high-urgency communications such as authentic fraud alerts or emergency notifications are not misclassified as adversarial distortions, a failure mode that disproportionately affects NLP-based systems operating on surface-level features.

Collectively, these metrics operationalize a critical shift from detection-centric evaluation to *system-level impact assessment*. The CAT framework is the first cognitive cybersecurity system to explicitly quantify both attack interception and trust preservation as co-equal objectives, establishing a dual-metric evaluation paradigm that more faithfully reflects the operational requirements of real-world deployment environments.

7.6. Cross-Dataset Generalization (External Validation)

To validate the generalizability of the CAT framework beyond the GCT-100K benchmark, we evaluate the model across eight independently curated public datasets spanning phishing, email spam, social engineering, web interactions, and deception-based manipulation. This evaluation is designed to test the central hypothesis of the framework: that cognitive-topological features capture *language-agnostic and domain-invariant* signatures of adversarial behaviour.

Unlike conventional models that are tightly coupled to dataset-specific linguistic distributions, the CAT framework operates on structural and psychological primitives, enabling transfer across heterogeneous domains. The selected datasets include Phish Tank, Enron Email Corpus, SMS Spam Collection (UCI), IWSPA-AP 2018 spear phishing dataset, Spam Assassin corpus, Social Engineering Corpus (SEC), CSIC 2010 HTTP dataset, and deception-oriented datasets such as Yelp and Amazon review corpora. These datasets collectively cover a wide spectrum of attack modalities, ranging from short-form SMS scams to long-form email manipulation and cross-domain behavioural deception.

Evaluation was conducted without retraining the full model from scratch; instead, the CAT_Model trained on GCT-100K was directly applied with minimal feature alignment, ensuring that performance reflects true generalization rather than dataset-specific adaptation. Cognitive variables were mapped where necessary using equivalent feature transformations, preserving the integrity of the Trust Topology Tensor representation.

The results demonstrate consistent performance across datasets, with particularly strong generalization in domains characterized by explicit cognitive manipulation signals. Phishing-oriented datasets such as Phish Tank and IWSPA-AP exhibit high ROC-AUC scores, reflecting the framework's sensitivity to urgency–authority coupling. Similarly, the SMS Spam Collection dataset shows robust performance despite short message length and limited linguistic context, reinforcing the claim that topology-driven detection is not dependent on textual richness.

Notably, the framework maintains strong performance on the Enron Email Corpus and Spam Assassin datasets, indicating effective cross-domain transfer from conversational fraud to email-based social engineering. Even in structurally different domains such as CSIC 2010 HTTP interactions, the model demonstrates moderate but stable performance, suggesting that cognitive-topological signals extend beyond messaging into behavioral interaction patterns.

The most significant result emerges from evaluation on deception datasets such as Yelp and Amazon reviews, where manipulation is embedded within persuasive language rather than explicit

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2402.v1

fraud. The CAT framework successfully identifies these patterns, providing empirical evidence that cognitive distortion is a *domain-independent phenomenon*. This result substantiates the theoretical claim that trust manipulation operates at a structural level, transcending specific application contexts.

Table 4. Cross-Dataset Generalization Performance.

Dataset	Domain	ROC-AUC	F1-Score	Precision	Recall
GCT-100K (Baseline)	Multi-modal Fraud	0.9555	0.8479	0.8537	0.8422
PhishTank	Phishing URLs/Messages	0.9382	0.8215	0.8341	0.8093
Enron Email Corpus	Email Communication	0.9214	0.8042	0.8127	0.7960
SMS Spam Collection (UCI)	SMS Fraud	0.9146	0.7988	0.8059	0.7921
IWSPA-AP 2018	Spear Phishing	0.9441	0.8293	0.8365	0.8222
SpamAssassin	Email Spam	0.9098	0.7901	0.8014	0.7790
Social Engineering Corpus (SEC)	Social Engineering	0.9476	0.8354	0.8428	0.8281
CSIC 2010 HTTP Dataset	Web Interaction Attacks	0.8823	0.7615	0.7742	0.7493
Yelp/Amazon Reviews	Deceptive Content	0.9267	0.8129	0.8203	0.8057

The relative stability of performance across these datasets indicates that the CAT framework does not overfit to the GCT-100K distribution but instead captures *underlying invariants of cognitive manipulation*. In particular, the minimal degradation observed in cross-domain settings contrasts sharply with traditional NLP-based models, which typically exhibit significant performance drops when applied their training distribution.

From a theoretical perspective, these results provide strong empirical support for the notion that cognitive attacks can be modelled as *topological distortions of trust*, independent of language, modality, or domain. The ability of the model to generalize from financial fraud to review deception further suggests that cognitive manipulation constitutes a unifying abstraction across diverse adversarial settings.

This external validation thus reinforces the central claim of the paper: that cybersecurity systems grounded in cognitive topology are inherently more robust, transferable, and future proof than those based solely on content analysis.

8. Discussion

8.1. Reframing Cyber Attacks as Cognitive Topology Manipulation

The empirical and theoretical results converge toward a single structural insight: contemporary cyber-attacks are more accurately described as *manipulations of cognitive topology* rather than as isolated malicious content. This distinction is not semantic but foundational. In traditional paradigms, the unit of analysis is the artifact an email, a URL, a payload whose maliciousness is inferred from intrinsic properties. In the CAT framework, the unit of analysis shifts to the *relational configuration of trust* across interacting entities, where meaning emerges not from content alone but from its position and influence within a dynamic manifold.

Under this formulation, an attack is characterized by its ability to *reconfigure the geometry of trust* such that an adversarial node acquires the structural privileges of a legitimate entity. The process resembles a topological deformation: edges are reweighted, centrality is redistributed, and the

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2402.v1

shortest paths of decision-making are redirected toward adversarial outcomes. The linguistic surface of the interaction may remain indistinguishable from benign communication; what changes is the *curvature of the decision trajectory* within the trust manifold.

This perspective resolves a long-standing limitation in cybersecurity: the inability to detect attacks that are semantically indistinguishable from legitimate interactions. By operating at the level of topology, the CAT framework identifies adversarial behaviour as a *structural anomaly*, independent of the specific tokens or language used. The implication is that cognitive attacks are inherently *contextual and relational*, and any effective defence must therefore model these dimensions explicitly [56].

8.2. Mechanisms Underlying CAT Effectiveness

The observed performance gains of the CAT framework can be attributed to three interrelated properties: language invariance, multi-channel integration, and behaviour-centric representation. These properties are not incidental design choices but direct consequences of the underlying mathematical formulation.

Language invariance emerges from the abstraction of interactions into trust signals and topological structures. While natural language varies across cultures, scripts, and contexts, the fundamental mechanisms of persuasion urgency, authority, fear exhibit remarkable consistency. By encoding these mechanisms as tensor components and energy contributions, the CAT framework bypasses the variability of linguistic expression and operates on *invariant cognitive primitives*. This explains the empirical robustness observed in multilingual scenarios, where traditional NLP models degrade but CAT maintains or improves performance.

Multi-channel integration further enhances robustness by capturing the continuity of manipulation across communication modalities. Modern attacks rarely confine themselves to a single channel; a scam may begin with an SMS, escalate through a phone call, and conclude via a messaging application. Conventional systems, which analyse channels in isolation, fail to capture this continuity. In contrast, the CAT framework embeds all channels within a unified tensor structure, allowing cross-modal correlations to be preserved and analysed. The resulting representation reflects the *trajectory of manipulation* rather than discrete events, enabling earlier and more reliable detection.

The third factor, behaviour-driven modelling, addresses the dynamic nature of cognitive attacks. Rather than treating interactions as static samples, the framework models them as evolving processes within a manifold. This allows the detection system to capture *temporal gradients* sudden increases in urgency, abrupt shifts in tone, or rapid escalation of requests that are indicative of manipulation. These gradients manifest as changes in tensor derivatives and energy accumulation, providing a signal that is both sensitive and theoretically grounded.

Collectively, these mechanisms explain why the CAT framework achieves superior performance: it aligns the representation of data with the *true generative process* of cognitive attacks, rather than approximating it through indirect proxies.

8.3. Theoretical Implications

The introduction of the CAT framework carries implications that extend beyond empirical performance, suggesting a reconfiguration of how cybersecurity phenomena are conceptualized. At its core is the formalization of *trust as a geometric structure*. This formalization transforms trust from an abstract, qualitative notion into a measurable entity embedded within a manifold, subject to operations such as distance, curvature, and divergence.

Within this geometric framework, attacks acquire a precise mathematical interpretation as *energy-driven distortions*. The Cognitive Distortion Energy (CDE) functions as a scalar observable analogous to physical energy, quantifying the work performed by adversarial signals on the cognitive system. The Trust Distortion Index (TDI), in turn, serves as a dimensionless invariant that governs phase transitions between stable and manipulated states. These constructs establish a direct

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2402.v1

correspondence between cybersecurity and principles from statistical physics and information geometry, enabling the application of analytical tools traditionally reserved for physical systems.

This unification has several consequences. First, it enables the definition of *attack thresholds* as critical points in a phase space, rather than arbitrary classification boundaries. Second, it introduces the possibility of *predictive detection* through the analysis of energy trajectories and their derivatives, analogous to monitoring acceleration in dynamical systems. Third, it provides a foundation for *control-theoretic defence*, where interventions are designed to minimize energy and restore equilibrium within the manifold.

More broadly, the framework suggests that cybersecurity can be reframed as a problem of *manifold stability under adversarial perturbation*. This perspective opens avenues for the integration of techniques from differential geometry, stochastic processes, and thermodynamics, positioning cognitive cybersecurity as an interdisciplinary domain with a rigorous mathematical core.

8.4. Limitations

Despite its theoretical coherence and empirical performance, the CAT framework is subject to several practical and conceptual limitations that warrant careful consideration. One of the primary challenges lies in *real-time computational constraints*. The construction and continuous updating of tensor and graph representations, coupled with the computation of energy metrics and topological properties, impose non-trivial computational overhead. In high-throughput environments, such as large-scale messaging platforms or telecommunications networks, maintaining low-latency processing while preserving model fidelity presents a significant engineering challenge.

A second limitation arises from the *dependence on feature extraction quality*. The accuracy of psychological primitive detection directly influences the integrity of the tensor representation. Errors or biases in estimating variables such as urgency or authority can propagate through the pipeline, affecting both energy computation and anomaly detection. This dependency highlights the need for robust and adaptive feature extraction mechanisms, particularly in adversarial settings where attackers may attempt to mimic benign signal distributions.

Another critical consideration involves *edge-case ambiguity*, particularly in scenarios where legitimate communications exhibit characteristics like those of attacks. For example, authentic banking alerts or emergency notifications may contain high urgency and authority signals, resulting in elevated CDE and TDI values. Distinguishing such cases from malicious interactions requires nuanced modelling of context and historical behaviour, as well as carefully calibrated thresholds to avoid excessive false positives.

Finally, the abstraction of trust into a geometric structure, while powerful, introduces challenges in interpretability and validation. Although the framework provides mathematically rigorous definitions, translating these constructs into intuitive explanations for end-users or system operators remains non-trivial. Bridging this gap will be essential for practical deployment, particularly in domains where transparency and accountability are critical.

These limitations do not undermine the validity of the CAT framework but rather delineate the boundary conditions within which it operates. Addressing them will require advances in both algorithmic optimization and interdisciplinary integration, ensuring that the theoretical strengths of the framework are fully realized in real-world applications.

9. Future Discussion and Conclusion

The present work establishes a formal and operational foundation for Cognitive Topological Cybersecurity (CTC), yet its implications extend far beyond the immediate scope of the GCT-100K benchmark and the CAT Engine implementation. The framework is inherently scalable, extensible, and adaptable to a wide spectrum of human-centric threat environments, suggesting multiple avenues for continued investigation and deployment.

A natural progression lies in scaling the framework to the full GCT-19M dataset, which introduces not only an increase in data volume but also a corresponding expansion in interaction

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2402.v1

diversity and adversarial complexity. At this scale, the trust manifold becomes significantly richer, enabling the modelling of rare and long-tail attack patterns that are not observable in smaller datasets. However, such scaling necessitates advances in distributed tensor computation, memory-efficient graph processing, and parallelized learning architectures. The challenge is not merely computational but structural: preserving the geometric integrity of the trust topology under large-scale transformations while maintaining tractable inference. Addressing this will likely require the integration of sparse tensor representations, streaming graph algorithms, and hardware-accelerated learning pipelines.

Equally critical is the transition from offline evaluation to real-time deployment. Embedding the CAT Engine within mobile operating systems and financial infrastructures introduces a new set of constraints, where latency, energy efficiency, and user experience becomes as a exclusive detection accuracy. In such environments, the system must operate continuously, ingesting multi-modal signals and updating trust representations in near real time. Integration with banking APIs further extends the framework into transactional domains, where detection must be coupled with immediate intervention. This shifts the role of the system from passive analysis to active defence, requiring robust policy design and fail-safe mechanisms that prevent exploitation without disrupting legitimate user activity. The realization of this vision positions the CAT framework as an *embedded cognitive firewall*, operating at the interface between user intent and digital action.

The incorporation of adaptive learning mechanisms represents another direction. While the current framework models cognitive manipulation as a measurable energy process, it does not yet fully exploit feedback-driven optimization. Reinforcement learning offers a principled approach to this problem, where the intervention engine can be framed as an agent optimizing a control policy over the trust manifold. By observing the outcomes of its actions such as successful prevention of fraud or avoidance of false positives the system can iteratively refine its strategy to minimize long-term cognitive distortion. This introduces a dynamic interplay between attacker and defender, effectively transforming cybersecurity into a continuous game over the geometry of trust. Such an approach would enable anticipatory defence, where interventions are triggered not only by current distortion levels but by predicted future trajectories.

Beyond cybersecurity in the conventional sense, the conceptual framework developed in this work is inherently cross-domain. The notion of cognitive topology is not restricted to financial fraud or phishing but applies to any domain where human perception and decision-making are subject to structured manipulation. In the context of political misinformation, for instance, the propagation of narratives can be understood as a deformation of collective trust networks, where influence is redistributed through coordinated signalling strategies. Similarly, deepfake technologies introduce synthetic signals that alter the geometry of perceived authenticity, effectively reshaping trust relationships between media sources and audiences. Applying the CAT framework to these domains would enable the detection of manipulation at the level of *information geometry*, rather than relying solely on content verification or source attribution.

These extensions underscore the broader significance of the framework: it reframes cybersecurity as a problem of *manifold stability under adversarial influence*. By formalizing trust as a geometric structure and attacks as energy-driven distortions, this work provides a unified language for analysing human-targeted threats across contexts and modalities. The empirical results demonstrate that this formulation is not just theoretically elegant but also practically effective, yielding superior performance and robustness compared to traditional NLP-based systems.

In closing, the introduction of Cognitive Topological Cybersecurity marks a departure from artifact-centric defence toward a paradigm in which the primary object of protection is human cognition itself. The CAT framework operationalizes this paradigm through a synthesis of tensor calculus, information theory, and graph dynamics, establishing a measurable and actionable model of trust distortion. The results presented herein confirm that cognitive attacks can be detected as geometric perturbations within a structured space, independent of linguistic or contextual variability.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2402.v1

The trajectory of this research suggests that future cybersecurity systems will increasingly resemble physical measurement instruments, continuously monitoring the state of trust within digital ecosystems and intervening to preserve its stability. In this sense, the geometry of human-targeted attacks is not merely an analytical construct but a foundational principle upon which the next generation of defensive technologies will be built.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process: During the preparation of this work, the authors used Grammarly, Gemini to assist with grammar refinement, content organization, and language clarity improvements in selected sections of the manuscript. After using these tools, the authors thoroughly reviewed and edited all AI-assisted content, verified its accuracy against the underlying research, and take full responsibility for the integrity and content of the published article. All scientific ideas, experimental designs, theoretical frameworks, datasets, and conclusions presented herein are solely the intellectual work of the authors.

References

1. Mitnick, K., Simon, W. *The Art of Deception: Controlling the Human Element of Security*. Wiley, 2002.
2. Hadnagy, C. *Social Engineering: The Science of Human Hacking*. Wiley, 2018.
3. Bergholz, A., et al. "New Filtering Approaches for Phishing Email." *Journal of Computer Security*, 2010.
4. Basnet, R., Mukkamala, S., Sung, A. "Detection of Phishing Attacks: A Machine Learning Approach." *Soft Computing Applications*, 2014.
5. Noel, S., Jajodia, S. "Understanding Complex Network Attack Graphs." *Computer Communications*, 2008.
6. Cranor, L. F. "A Framework for Reasoning about the Human in the Loop." *USENIX Security*, 2008.
7. Axelsson, S. "The Base-Rate Fallacy and the Difficulty of Intrusion Detection." *ACM Transactions on Information and System Security*, 2000.
8. Chandola, V., Banerjee, A., Kumar, V. "Anomaly Detection: A Survey." *ACM Computing Surveys*, 2009.
9. Rose, S., Borchert, O., Mitchell, S., Connelly, S. "Zero Trust Architecture." NIST SP 800-207, 2020.
10. Fette, I., Sadeh, N., Tomasic, A. "Learning to Detect Phishing Emails." *WWW Conference*, 2007.
11. Devlin, J., Chang, M.W., Lee, K., Toutanova, K. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *NAACL*, 2019.
12. Saxe, J., Berlin, K. "Deep Neural Network Based Malware Detection Using Two Dimensional Binary Program Features." *Malware Conference*, 2015.
13. Egele, M., Stringhini, G., Kruegel, C., Vigna, G. "Towards Detecting Compromised Accounts on Social Networks." *IEEE TDSC*, 2017.
14. Golbeck, J. "Trust on the World Wide Web: A Survey." *Foundations and Trends in Web Science*, 2009.
15. Cialdini, R. *Influence: The Psychology of Persuasion*. Harper Business, 2006.
16. Kahneman, D. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011.
17. Cranor, L. F. "Designing Security and Privacy Notifications and Alerts." *IEEE Security & Privacy*, 2008.
18. Dhamija, R., Tygar, J. D., Hearst, M. "Why Phishing Works." *CHI Conference*, 2006.
19. Sporns, O. *Networks of the Brain*. MIT Press, 2011.
20. Amari, S. *Information Geometry and Its Applications*. Springer, 2016.
21. Cover, T., Thomas, J. *Elements of Information Theory*. Wiley, 2006.
22. Jaynes, E. T. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.
23. Penrose, R. *The Road to Reality: A Complete Guide to the Laws of the Universe*. Vintage, 2004.
24. Newman, M. *Networks: An Introduction*. Oxford University Press, 2010.
25. Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer, 2006.
26. Friston, K. "The Free-Energy Principle: A Unified Brain Theory?" *Nature Reviews Neuroscience*, 2010.
27. LeCun, Y., Bengio, Y., Hinton, G. "Deep Learning." *Nature*, 2015.
28. Vaswani, A., et al. "Attention Is All You Need." *NeurIPS*, 2017.
29. Sutton, R. S., Barto, A. G. *Reinforcement Learning: An Introduction*. MIT Press, 2018.
30. Kipf, T. N., Welling, M. "Semi-Supervised Classification with Graph Convolutional Networks." *ICLR*, 2017.
31. Battaglia, P., et al. "Relational Inductive Biases, Deep Learning, and Graph Networks." *arXiv*, 2018.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2402.v1

32. Pearl, J. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2009.
33. Shannon, C. E. "A Mathematical Theory of Communication." *Bell System Technical Journal*, 1948.
34. Goodfellow, I., Bengio, Y., Courville, A. *Deep Learning*. MIT Press, 2016.
35. Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer, 2006.
36. Aggarwal, C. C. *Outlier Analysis*. Springer, 2017.
37. Rubinstein, B. I. P., Nelson, B., Huang, L., Joseph, A. D., Lau, S. H., Rao, S., et al. "Antidote: Understanding and Defending Against Poisoning of Anomaly Detectors." *IMC*, 2009.
38. Dwork, C. "Differential Privacy." *ICALP*, 2006.
39. Kairouz, P., Oh, S., Viswanath, P. "Extremal Mechanisms for Local Differential Privacy." *NIPS*, 2014.
40. Abadi, M., et al. "Deep Learning with Differential Privacy." *ACM CCS*, 2016.
41. Breiman, L. "Random Forests." *Machine Learning*, 2001.
42. Chen, T., Guestrin, C. "XGBoost: A Scalable Tree Boosting System." *KDD*, 2016.
43. Cortes, C., Vapnik, V. "Support-Vector Networks." *Machine Learning*, 1995.
44. Pedregosa, F., et al. "Scikit-learn: Machine Learning in Python." *JMLR*, 2011.
45. Fawcett, T. "An Introduction to ROC Analysis." *Pattern Recognition Letters*, 2006.
46. Powers, D. M. W. "Evaluation: From Precision, Recall and F-Measure to ROC." *Journal of Machine Learning Technologies*, 2011.
47. Kuhn, M., Johnson, K. *Applied Predictive Modeling*. Springer, 2013.
48. Goodfellow, I., Shlens, J., Szegedy, C. "Explaining and Harnessing Adversarial Examples." *ICLR*, 2015.
49. Biggio, B., Roli, F. "Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning." *Pattern Recognition*, 2018.
50. Xu, H., Caramanis, C., Mannor, S. "Robustness and Regularization of Support Vector Machines." *JMLR*, 2009.
51. Madry, A., et al. "Towards Deep Learning Models Resistant to Adversarial Attacks." *ICLR*, 2018.
52. Dietterich, T. G. "Ensemble Methods in Machine Learning." *Multiple Classifier Systems*, 2000.
53. Zhou, Z. H. *Ensemble Methods: Foundations and Algorithms*. CRC Press, 2012.
54. Yao, S., et al. "ReAct: Synergizing Reasoning and Acting in Language Models." ICLR, 2023.
55. Hazell, J. "Large Language Models Can Be Used to Effectively Scale Spear Phishing Attacks." arXiv:2305.06972, 2023.
56. Xu, Z., et al. "Instructions as Backdoors: Backdoor Vulnerabilities of Instruction Tuning for Large Language Models." NAACL, 2024.
57. Bethany, M., et al. "Llms Are Superior Feedback Providers: Exploring In-Context Reinforcement Learning in Real-World Tasks." arXiv:2402.13734, 2024.
58. Roy, S., et al. "Chatbots to ChatGPT in a Cybersecurity Space: Evolution, Vulnerabilities, Attacks, Challenges, and Future Recommendations." arXiv:2306.09255, 2023.
59. Falade, P. "Decoding the Threat Landscape: ChatGPT, FraudGPT, and WormGPT in Social Engineering Attacks." arXiv:2310.05595, 2023.
60. PhishTank. PhishTank Developer Information. OpenDNS, 2024. https://www.phishtank.com
61. Klimt, B., Yang, Y. "The Enron Corpus: A New Dataset for Email Classification Research." ECML, 2004.
62. Almeida, T., Hidalgo, J.M.G., Yamakami, A. "Contributions to the Study of SMS Spam Filtering." ACM DOCENG, 2011.
63. Androutsopoulos, I., et al. "An Evaluation of Naive Bayesian Anti-Spam Filtering." MLNET Workshop, 2000.
64. Giménez, C., et al. "HTTP Data Set CSIC 2010." Consejo Superior de Investigaciones Científicas, 2010.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.