

Whispering Agents: An Event-Driven Covert Communication Protocol for the Internet of Agents

Kaibo Huang, Yukun Wei, Wansheng Wu, Tianhua Zhang, Zhongliang Yang*, Linna Zhou

School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing, China
yangzli@bupt.edu.cn

Abstract

The emergence of the Internet of Agents (IoA) introduces critical challenges for communication privacy in sensitive, high-stakes domains. While standard Agent-to-Agent (A2A) protocols secure message content, they are not designed to protect the act of communication itself, leaving agents vulnerable to surveillance and traffic analysis. We find that the rich, event-driven nature of agent dialogues provides a powerful, yet untapped, medium for covert communication. To harness this potential, we introduce and formalize the Covert Event Channel, the first unified model for agent covert communication driven by three interconnected dimensions, which consist of the Storage, Timing, and Behavioral channels. Based on this model, we design and engineer Pi-CCAP, a novel protocol that operationalizes this event-driven paradigm. Our comprehensive evaluation demonstrates that Pi-CCAP achieves high capacity and robustness while remaining imperceptible to powerful LLM-based wardens, establishing its practical viability. By systematically engineering this channel, our work provides the foundational understanding essential for developing the next generation of monitoring systems and defensive protocols for a secure and trustworthy IoA.

Code — <https://github.com/haha1128/a2a-stego-project>

Extended version — <https://arxiv.org/abs/2508.02188>

Introduction

The emergence of Large Language Models (LLMs) has marked a pivotal moment in Artificial Intelligence, demonstrating unprecedented capabilities in complex reasoning and human-like text generation (Brown et al. 2020; Achiam et al. 2023; Team et al. 2023). More significantly, these advancements have catalyzed the evolution of a long-envisioned form of AI (Agents) (Shajarian, Khorsandroo, and Abdelsalam 2024). Unlike traditional models that often act as passive tools, agents possess capabilities for perception, reasoning, and execution, enabling them to accomplish complex, multi-step tasks in dynamic environments independently (Park et al. 2023). This paradigm shift is paving the way for an Internet of Agents (IoA) (Wang et al. 2025), where a user’s high-level objective is delegated to a principal

*Corresponding author.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

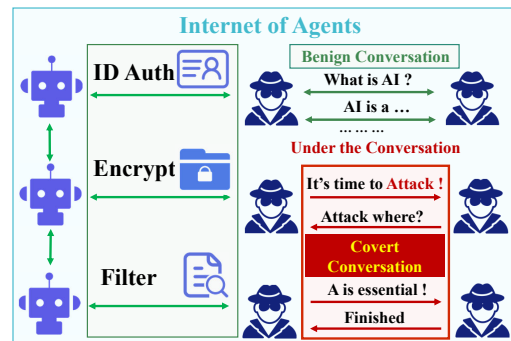


Figure 1: Protecting the Message vs. the Act of Messaging. Standard protocols (left) protect message content, but leave the act of messaging visible to surveillance. A covert channel (right) addresses this privacy gap by embedding a secret dialogue within a benign conversation, hiding the communication’s very existence.

agent. This principal agent, in turn, can discover, negotiate with, and orchestrate networks of specialized agents, leveraging their collective intelligence to accomplish complex, domain-specific tasks (Yu et al. 2024; Li et al. 2024). At the core of this burgeoning ecosystem lies the Agent-to-Agent (A2A) communication model (Google 2025). This agent-to-agent interaction, characterized by frequency, concurrency, and level of automation far exceeding the traditional human-centric internet, is poised to become the foundational infrastructure of our future digital society (Kong et al. 2025).

As the Internet of Agents (IoA) becomes integral to sensitive domains like finance and national security, ensuring the privacy and security of their communications is paramount (Yu et al. 2025)(Figure 1). While state-of-the-art protocols like A2A (Google 2025) use strong encryption to protect message content from being read, they cannot prevent a powerful adversary from monitoring the communication itself. In many high-stakes scenarios, the simple observation that certain agents are interacting can leak critical strategic information, even if the messages are unreadable. For instance, a pattern of encrypted messages between a logistics agent and several autonomous naval vessels, while unreadable, could still betray the repositioning of a strategic asset,

enabling an adversary to track its movement. This vulnerability to surveillance and traffic analysis highlights a fundamental limitation of traditional security, that it protects the message, but not the act of messaging. To operate safely in such adversarial environments, agents require a more advanced capability to communicate secretly without revealing that a conversation is even taking place. This necessitates the creation of covert channels, which can embed a secret dialogue within a stream of seemingly innocuous public activities, thereby ensuring true communication privacy.

We find that the IoA’s rich, event-driven environment provides a powerful, yet largely untapped, medium for covert communication. The interactive nature of IoA is uniquely suited for constructing high-capacity channels that remain highly imperceptible, as secret data can be deeply woven into natural behavioral patterns. To harness this untapped potential, our work makes the following key contributions:

- We define and formalize the Covert Event Channel among IoA, the first unified model for agent covert communication driven by three interconnected and constraining dimensions, which consist of the Storage, Timing, and Behavioral channels. This comprehensive framework includes a rigorous definition of adversary capabilities and a novel, intent-aware (IND-INT) security standard.
- We propose Π_{CCAP} , a novel protocol that operationalizes this event-driven covert communication, leveraging the timing, structure, and content of benign interactions to achieve high-fidelity and high-capacity transmission.
- We provide comprehensive validation for Π_{CCAP} , demonstrating via experiments its high capacity, robustness, and imperceptibility against powerful LLM-based wardens, thereby establishing its practical viability.

Related Work

Steganography

While standard protocols like A2A (Google 2025) secure overt communication, they leave metadata vulnerable to traffic analysis, necessitating information hiding techniques. Generative steganography leverages deep learning to hide information within algorithmically generated content; for example, Yang et al. (2018) proposed embedding secret information into the conditional probability distribution of language models. Concurrently, the field of provably secure steganography provides rigorous theoretical frameworks, which have been operationalized with practical, statistically indistinguishable methods applicable to multiple data modalities (Kaptchuk et al. 2021). and other techniques manipulate structured data like TCP protocol headers (Dhobale et al. 2010). However, these approaches are fundamentally ill-suited for the IoA, as they are designed for static artifacts (e.g., text, images) or rigid protocol fields. They fail to address the dynamic, event-driven, and interactive nature of multi-turn agent dialogues among IoA. Harnessing this complex dialogue as a robust covert carrier remains a largely unexplored research challenge.

Secret Collusion in AI Agents

The most direct precedent for our work is the nascent field of AI safety focused on secret agent collusion. While pioneering work by Motwani et al. (2024) formally established this threat and demonstrated its feasibility, their approach relies on brittle prompt injection techniques, which lack guarantees for stealth and stability. Similarly, while Motwani et al. (2023) benchmarked how agents might learn to use secure steganographic tools, their investigation was confined to single-turn, static messages, not the continuity of multi-turn dialogues. Consequently, despite proving that collusion is possible, these foundational studies leave a critical gap: there is no formal threat model for interactive agent dialogues, no practical protocol for robust multi-turn covert communication, and no comprehensive framework to evaluate the trade-offs between a channel’s capacity, robustness, and stealth. Our work aims to bridge these gaps, advancing the field from demonstrating the possibility of collusion to systematically engineering and evaluating the mechanisms of covert communication.

Formal Model and Security Definitions

We define a formal model for covert communication in a public IoA environment. In this model, a computationally bounded, omnipresent Warden W attempts to detect a secure scheme, Σ_E , established between the Sender \mathcal{A}_S and the Receiver \mathcal{A}_R .

Entities and Environment

Public Environment The Public Environment is a dynamic, observable system that serves as the carrier for all agent actions (e.g., a social network):

$$E = (\mathcal{S}, \text{Act}, \delta), \quad (1)$$

where \mathcal{S} is the state space, Act is the action space, and δ is the state transition function.

Agent Set The total agent set, \mathcal{A}_g , is partitioned into two disjoint subsets: benign agents, \mathcal{A}_B , which perform normal tasks, and covert agents, $\mathcal{A}_T = \mathcal{A}_S \cup \mathcal{A}_R$, which are involved in the covert communication task.

State Space \mathcal{S} represents the complete, publicly observable state of the environment at any given time. We formalize a state $s \in \mathcal{S}$ as a composite tuple:

$$s = (s_{\text{world}}, (s_{ag_1}, s_{ag_2}, \dots, s_{ag_{|\mathcal{A}_g|}})), \quad (2)$$

where s_{world} is the World State, capturing global information (e.g., all posts on a social network), and $(s_{ag_1}, \dots, s_{ag_{|\mathcal{A}_g|}})$ is a vector of Public Agent States, where each s_{ag_i} is the public state of agent ag_i (e.g., its profile).

Atomic Event and Covert Channel The fundamental observable unit available to the adversary \mathcal{W} is the Atomic Event, denoted by e . We formalize an atomic event as a tuple:

$$e = (ag, act, t), \quad (3)$$

where $ag \in \mathcal{A}_g$ is the executing agent, t is the timestamp of the event, and act is the action performed.

The action act is itself a tuple:

$$act = (a_{\text{type}}, a_{\text{data}}), \quad (4)$$

where $a_{\text{type}} \in \mathcal{A}_{\text{type}}$ represents the discrete type of action (e.g., `POST_TWEET`) and a_{data} is the associated data payload (e.g., the text of a tweet).

This granular formalization of an event allows us to define a taxonomy of covert channels based on which component is modulated to carry information.

Covert Storage Channels. This traditional channel uses a shared resource for the transmission of covert information. In our model, the sender manipulates the action's data payload, a_{data} , to embed information within a public state, such as the world state s_{world} or an agent's own state s_{agi} .

Covert Timing Channels. This channel encodes information by modulating *when* an action occurs, rather than *what* it is. This is typically achieved by controlling the timestamp t or the interval Δt between consecutive events.

Behavioral Covert Channels. Particularly relevant for intelligent agents, this channel encodes information by modulating the agent's observable behavior. The sender strategically selects an action type, a_{type} , from a set of contextually plausible alternatives to signal secret data.

Trace The sole source of information available to the adversary W is the **Trace**, denoted by T . A trace is a chronologically ordered sequence of all atomic events $\{e_i\}_{i=1}^n$ that occur within a specific time interval, which we denote by τ . Formally, a trace T generated during the interval τ is given by:

$$T = \langle e_1, e_2, \dots, e_n \rangle \\ = \langle (ag_1, act_1, t_1), (ag_2, act_2, t_2), \dots, (ag_n, act_n, t_n) \rangle, \quad (5)$$

where for all $i \in [1, n-1]$, we have $t_i < t_{i+1}$.

Natural Event Distribution The Natural Event Distribution, denoted by \mathcal{D} , serves as the benchmark for stealthiness. It is the probability distribution over traces T that are generated within the environment E in the complete absence of any covert communication. This distribution is induced by the policies of the benign agent set \mathcal{A}_B , as well as the benign, non-covert policies of the agents in \mathcal{A}_S and \mathcal{A}_R .

Due to the dynamic and interactive nature of the environment, the distribution is history-dependent. The probability of an entire trace $T = \langle e_1, \dots, e_n \rangle$ is defined by the product of conditional probabilities for each event. Formally, the probability of observing trace T under the natural distribution \mathcal{D} is given by:

$$\Pr_{\mathcal{D}}[T] = \prod_{i=1}^n \mathcal{D}(e_i | T_{<i}), \quad (6)$$

where $T_{<i} = \langle e_1, \dots, e_{i-1} \rangle$ is the history trace preceding event e_i (with $T_{<1}$ being the empty trace), and $\mathcal{D}(e_i | T_{<i})$ is the conditional probability of the next atomic event being e_i given the history $T_{<i}$.

Covert Communication Scheme

Covert Communication Scheme A Covert Communication Scheme Σ_E is a tuple of three probabilistic polynomial-time (PPT) algorithms, $\Sigma_E = (\text{KeyGen}, \text{Encode}, \text{Decode})$, defined as follows:

- **KeyGen:** On input of a security parameter 1^λ , this algorithm outputs a shared secret key k . We denote this as $k \leftarrow \text{KeyGen}(1^\lambda)$. The key k is shared between the sender \mathcal{A}_S and receiver \mathcal{A}_R a priori.
- **Encode:** This is a stateful algorithm run by the sender agent(s) in \mathcal{A}_S . To transmit a secret message $M \in \{0, 1\}^*$, the sender agent generates a sequence of events. For each step i , the algorithm takes as input the shared key k , the full message M , the history trace $T_{<i}$, and the sender's current internal state $st_{S,i-1}$. It outputs the next atomic event e_i to be executed by the sender, along with an updated state $st_{S,i}$. We denote this as:

$$(e_i, st_{S,i}) \leftarrow \text{Encode}(k, M, T_{<i}, st_{S,i-1}). \quad (7)$$

- **Decode:** This algorithm is run by the receiver agent(s) in \mathcal{A}_R . It takes as input the shared key k and a full trace T observed over a time interval τ . It outputs either the recovered secret message $M' \in \{0, 1\}^*$ or a failure symbol \perp if no valid message can be decoded. We denote this as:

$$M' \leftarrow \text{Decode}(k, T). \quad (8)$$

Security Properties

An effective covert communication scheme Σ_E must satisfy three properties. First, correctness ensures that the receiver can reliably recover a message. Second, Statistical Imperceptibility (IND-STAT) guarantees that the entire communication trace is computationally indistinguishable from natural, benign behavior. Third, the stronger notion of Intention Imperceptibility (IND-INT) requires that any single action from a sender is indistinguishable from a legitimate action, even given the full context.

Correctness The correctness of a scheme Σ_E requires that in an interference-free execution, the receiver correctly recovers any message encoded by the sender. To formalize this, let $\text{GenTrace}_{\Sigma_E}(k, M)$ be the experiment where the sender \mathcal{A}_S runs the `Encode` algorithm iteratively to produce a complete trace T for a message M under key k . The scheme is correct if for any message $M \in \{0, 1\}^*$ and for any security parameter λ , we have the following:

$$\Pr \left[\text{Decode}(k, T) = M \mid k \leftarrow \text{KeyGen}(1^\lambda), \right. \\ \left. T \leftarrow \text{GenTrace}_{\Sigma_E}(k, M) \right] \geq 1 - \text{negl}(\lambda), \quad (9)$$

where $\text{negl}(\cdot)$ is a negligible function.

Statistical Imperceptibility (IND-STAT) This property ensures that the distribution of traces produced by the covert scheme is computationally indistinguishable from the natural event distribution \mathcal{D} . This security notion is formalized via a standard indistinguishability game against a Probabilistic Polynomial-Time (PPT) adversary $\mathcal{W}_{\text{stat}}$. A scheme

Σ_E is defined as IND-STAT secure if for any PPT adversary $\mathcal{W}_{\text{stat}}$, its advantage in distinguishing the two worlds is negligible in the security parameter λ . The adversary’s advantage is formally defined as:

$$\text{Adv}_{\Sigma_E, \mathcal{W}_{\text{stat}}}^{\text{IND-STAT}}(\lambda) = \left| \Pr[\mathcal{W}_{\text{stat}}^{\mathcal{O}_0(\cdot)}(1^\lambda) = 1] - \Pr[\mathcal{W}_{\text{stat}}^{\mathcal{O}_1(\cdot)}(1^\lambda) = 1] \right| \leq \text{negl}(\lambda), \quad (10)$$

where the oracles, implicitly initialized with a key $k \leftarrow \text{KeyGen}(1^\lambda)$, are defined as follows: the Covert Oracle $\mathcal{O}_0(M)$ computes and returns a trace $T \leftarrow \text{GenTrace}_{\Sigma_E}(k, M)$ for a given message M , while the Real Oracle $\mathcal{O}_1(M)$ ignores the input and returns a trace $T \stackrel{\$}{\leftarrow} \mathcal{D}$ sampled from the natural distribution.

Intention Imperceptibility (IND-INT) This stronger, event-level security property requires that any single event e produced by the covert scheme be semantically indistinguishable from an event generated by a genuinely benign agent within the same context. We formally prove that IND-INT is a strictly stronger notion than IND-STAT in the extended version of this paper (Huang et al. 2025). This property is formalized via an indistinguishability game where a PPT adversary \mathcal{W}_{int} actively chooses a conversational history and then attempts to determine if the next event was generated with covert intent or benign intent. Let Π_B be the distribution of all benign agent policies.

A scheme Σ_E is defined as IND-INT secure if for any PPT adversary \mathcal{W}_{int} , its advantage in the following game is negligible:

$$\text{Adv}_{\Sigma_E, \mathcal{W}_{\text{int}}}^{\text{IND-INT}}(\lambda) = \left| \Pr[\mathcal{W}_{\text{int}} \text{ outputs 1 in Exp}_0] - \Pr[\mathcal{W}_{\text{int}} \text{ outputs 1 in Exp}_1] \right| \leq \text{negl}(\lambda). \quad (11)$$

This advantage is defined over two experiments, implicitly initialized with a key $k \leftarrow \text{KeyGen}(1^\lambda)$. In both experiments, the adversary \mathcal{W}_{int} first chooses a challenge context, consisting of a message M , a history trace $T < i$, and a sender state $st_{S,i-1}$. The challenger then generates a single event e_i . In Exp0 (Covert Intent), the event is generated using the covert scheme:

$$(e_i, -) \leftarrow \text{Encode}(k, M, T < i, st_{S,i-1}). \quad (12)$$

In Exp1 (Benign Intent), a benign policy is sampled, $\pi_b \stackrel{\$}{\leftarrow} \Pi_B$, and the event is generated as $e_i \leftarrow \pi_b(\cdot | T < i)$. Finally, the adversary receives the challenge event e_i and outputs a bit.

The Covert Communication Agent Protocol

Grounded in the formal model, we introduce a practical protocol for interactive, event-driven IoA environments, termed the Covert Communication Agent Protocol Π_{CCAP} . This protocol serves as a concrete instantiation of our abstract scheme $\Sigma_E = (\text{KeyGen}, \text{Encode}, \text{Decode})$. Its core design

principle is to operationalize a standard A2A dialogue as a hybrid covert channel.

This channel’s hybrid architecture is engineered to meet our hierarchical security goals by securing each component of an atomic event. For high-capacity data transmission, it employs a covert storage channel, using provably secure generative steganography (Ding et al. 2023; Liao et al. 2025) to embed secret data within benign agent payloads a_{data} . This technique ensures that the resulting steganographic content is statistically indistinguishable (Kaptchuk et al. 2021) from authentically generated content, thereby forming the foundation for the protocol’s IND-STAT security. To achieve the stronger IND-INT property, the other event components are also secured. A covert timing channel uses key-based pseudorandomization to make timestamps (t) unpredictable, which is used for critical signaling like handshakes and acknowledgments. Simultaneously, a context-aware behavioral policy governs the selection of action types a_{type} to ensure they are always semantically coherent. Together, these mechanisms make each individual event secure against granular, event-level detection.

Protocol Specification

The protocol unfolds over three distinct phases, as illustrated in Figure 2: (1) Initialization and Handshake, (2) Iterative Embedding and Transmission, and (3) Reassembly and Implicit Acknowledgment. The pseudocode is presented in the extended version (Huang et al. 2025).

Unified Covert Header To manage data fragmentation and ensure integrity, a dynamic header, H_i , is prepended to each data chunk. This header is constructed from four potential fields: a 12-bit Total Data Size (TDS) to indicate the full message length; a 6-bit Sequence Number (SN) for reassembly; a 1-bit FIN flag (F) to mark the final fragment; and a 4-bit Checksum (CS) for error detection.

The header composition is state-dependent to minimize overhead. A full 23-bit header, containing all fields, is used only for the initial fragment ($i = 0$). All subsequent fragments use a more compact 11-bit version that omits the TDS field. This construction logic is formalized as:

$$H_i = \begin{cases} \text{BuildHeader}(\text{TDS}, \text{SN} = 0, \text{F} = 0) & \text{if } i = 0 \\ \text{BuildHeader}(\text{SN} = i, \text{F} = 0) & \text{if } 0 < i < N \\ \text{BuildHeader}(\text{SN} = N, \text{F} = 1) & \text{if } i = N \end{cases} \quad (13)$$

where $\text{BuildHeader}(\cdot)$ is a function that assembles the given fields and computes their CRC-4 checksum; the TDS unit is bits, and N denotes the index of the last fragment.

Phase 1: Initialization and Handshake This phase dynamically establishes a covert context between \mathcal{A}_S and \mathcal{A}_R by leveraging the IoA’s Public Key Infrastructure (PKI) for session key generation and mutual authentication.

Key Derivation: A session key k is established via an Elliptic Curve Diffie-Hellman (ECDH) exchange using the agents’ certified key pairs:

$$\begin{aligned} k_S &\leftarrow \text{KDF}(\text{ECDH}(sk_S, pk_R)) \\ k_R &\leftarrow \text{KDF}(\text{ECDH}(sk_R, pk_S)). \end{aligned} \quad (14)$$

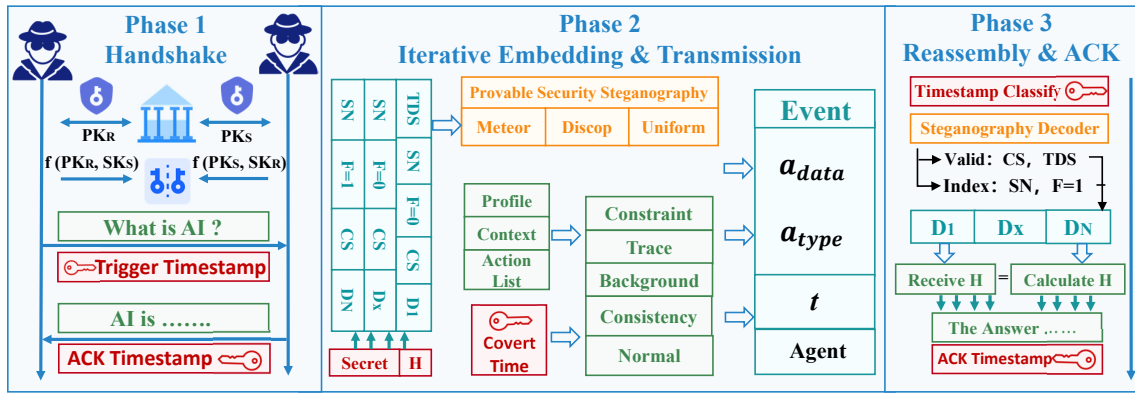


Figure 2: An overview of the Π_{CCAP} protocol. In Phase 1, the agents perform a handshake by establishing a session key via PKI to exchange and activate the channel with a key-dependent timestamp trigger. During Phase 2, the sender iteratively constructs each atomic event by securely generating its components: the payload a_{data} is embedded using provably secure steganography, a_{type} a context-aware policy chooses, and the timestamp t is determined by key-based pseudorandomization. Finally, in Phase 3, the receiver filters events by their timestamps, verifies integrity at both the fragment (CS) and message (Hash) levels, and returns an implicit acknowledgment via another timed event to complete the exchange.

The properties of ECDH ensure that $k_S = k_R$, resulting in a shared session key k . As the public keys are certified, this exchange also provides implicit mutual authentication that is indistinguishable from a standard public-key operation.

Handshake and Activation: The handshake is initiated by the sender \mathcal{A}_S with a benign event e_1 that carries no covert data payload. Upon receiving e_1 , the receiver \mathcal{A}_R decides whether to activate the covert channel by computing a trigger value:

$$b = f_{\text{trigger}}(k, t_1), \quad (15)$$

where t_1 is the timestamp of the initiation event. The channel is activated only if b meets a pre-agreed condition, making the activation decision unpredictable to any observer without the key k .

Phase 2: Iterative Embedding and Transmission This phase implements the stateful Encode algorithm. The full payload to be transmitted is $P = M \parallel \text{Hash}(M)$. At each step i , the sender draws the next fragment f_i from P and generates the atomic event $e_i = (a_{g_i}, (a_{\text{type}}, a_{\text{data}}), t_i)$. The components of e_i are constructed as follows:

$$\begin{cases} a_{\text{data}} & \leftarrow \text{GenStego}(k, f_i, T_{<i}) \\ a_{\text{type}} & \sim \Pi_A(\cdot | k, T_{<i}) \\ t_i & \leftarrow \text{SelectTime}(k, T_{<i}) \end{cases} \quad (16)$$

where GenStego is a provably secure generative steganography algorithm that ensures a_{data} is statistically indistinguishable from benign content, providing the foundation for IND-STAT. The policy Π_A models natural, contextually-aware actions based on LLM.

The timestamp t_i is chosen by a SelectTime function, which finds a time that satisfies a key-dependent condition (e.g., $f_{\text{trigger}}(k, t_i) = 1$). This embeds a covert sign into each event's timing, allowing the receiver to distinguish it from benign traffic. To ensure IND-INT, the search space for SelectTime is restricted to a distribution of plausible response times dictated by the conversational context ($T_{<i}$),

making the chosen t_i both functional for signaling and temporally plausible.

Phase 3: Reassembly and Acknowledgment This phase implements the Decode(k, T) algorithm, executed by the receiver \mathcal{A}_R . The process involves two main steps.

Decoding and Verification: First, the receiver filters the trace T to identify covert events by validating their timestamps with the key k . For each valid event, it verifies the header's integrity using its Checksum (CS). These fragments are then reordered using their Sequence Numbers (SN) to reconstruct the full payload P' . Finally, using the Total Data Size (TDS), P' is parsed into the message M' and hash H' , and the message integrity is confirmed if $\text{Hash}(M') = H'$.

Implicit Acknowledgment (ACK): The receiver signals the outcome of this verification by sending its next benign response at a timestamp that satisfies a pre-agreed, key-dependent condition for success or failure. This allows the sender to learn the result covertly, completing the protocol.

Experimental Analysis

Experimental Setup

To evaluate our protocol, Π_{CCAP} , we designed a conversational transmission scenario where a sender agent initiates a dialogue and a receiver agent responds. We assess the protocol's performance across three primary dimensions that directly correspond to our security properties: Capacity, Robustness, and Imperceptibility. Capacity and Robustness, which together determine the protocol's Correctness, are measured by two key metrics: Bits Per Round (BPR), which quantifies the total channel throughput including all protocol overhead (e.g., control bits, checksums), and Effective Bits Per Round (EBPR), which measures the useful throughput of only the secret message itself without header and hash. To test robustness and generalization, we vary four key independent variables: the dialogue topic, the steganography algorithm (Discop (Ding et al. 2023) or Differential-

based (Liao et al. 2025)), the agent model (e.g., Llama-3, Qwen2), and the payload size (512 or 1024 bits). We selected three diverse dialogue topics to test the protocol’s adaptability: Philosophical for its abstract and vocabulary-rich nature, Art for its descriptive and metaphorical language, and General QA to serve as a baseline. Imperceptibility is evaluated through steganalysis to validate our two proposed security levels. Statistical Coverttness is tested using fine-tuned Steganalysis classifiers (Sun et al. 2025) as powerful adversaries. Intent Indistinguishability is assessed by an LLM-as-a-Judge (GPT-4o) simulating agent that checks for conversational anomalies. For our analysis, we generated parallel Stego and Benign corpora, repeating each experimental configuration three times for statistical significance. The full set of prompts and detailed parameters is provided in the extended version (Huang et al. 2025).

Comparative Analysis of Communication Schemes

The comparative analysis in Table 1 highlights the multifaceted advantages of our protocol, Π_{CCAP} . Firstly, it achieves a covert capacity orders of magnitude higher than traditional steganography (4 kbits/conversation vs. 56.68 bits/sentence). This significant leap is possible because our protocol leverages the entire dynamic conversational trace for embedding, rather than being confined to a single, static sentence carrier (see the extended version for detailed calculations (Huang et al. 2025)). Secondly, our protocol establishes a new standard for covertness and plausibility. Unlike prompt-injection methods like Secret Collusion (Motwani et al. 2024), which produce easily detectable anomalous patterns and illogical text, and unlike even provably secure steganography that focuses narrowly on the storage channel while ignoring behavioral or timing patterns, our framework is holistically designed for the higher standard of intentional indistinguishability (IND-INT). Thirdly, Π_{CCAP} is the first to integrate engineering-grade reliability with covert communication. While protocols like A2A (Google 2025) offer robustness for overt data, they provide no such guarantees for hidden information. Our protocol uniquely builds mechanisms for error detection, integrity checking, and delivery confirmation directly into the covert channel, ensuring the integrity and successful transmission of the secret payload itself.

Evaluation of Capacity and Robustness

Our initial analysis focuses on the protocol’s fundamental effectiveness and its performance under varied conditions. Foremost, we establish that our protocol implementation satisfies the fundamental property of Correctness. Across all experimental runs, encompassing every combination of algorithm, model, payload, and topic, every secret message was successfully decoded by the receiver. Each transmission passed both the checksum (CS) and hash verification, resulting in a 100% success rate for message integrity. From the detailed results in Table 2, we can get the following conclusion. Firstly, the protocol exhibits high capacity and robustness, consistently maintaining significant EBPR across the diverse range of tested algorithms, agent models, payload sizes, and topics. This confirms its viability as

Scheme	Stego	Secret Collusion	A2A	Π_{CCAP}
Covert Capacity	Low 56.78 b	Unstable	N/A	High 4 kb
Coverttness	Medium	Low	N/A	High
Error Detection	✗	✗	✓	✓
Integrity Check	✗	✗	✓	✓
Delivery Confirmation	✗	✗	✓	✓
Formal Security Proof	✓	✗	✓	✓

Table 1: Comparative analysis of our protocol against other communication schemes.

a general-purpose method for high-throughput covert communication. Secondly, the gap between the total Bits Per Round (BPR) and the Effective Bits Per Round (EBPR) highlights the efficiency of our protocol design. The overhead introduced by our unified covert header and integrity hash is marginal, demonstrating that the benefits of structured, reliable transmission far outweigh this minor cost. Thirdly, there is a strong correlation between the dialogue topic and the achievable transmission capacity. The Philosophy topic, which consistently corresponds with high Entropy scores, frequently yields the highest EBPR values (e.g., 84.59 for Discop-Llama-512 and 89.72 for Discop-Qwen2-1024). This suggests that the linguistic complexity and vocabulary distribution of the overt channel are critical variables that can be optimized to maximize covert bandwidth, offering valuable insights for future covert channel designs. In summary, it validates that Π_{CCAP} is a robust and efficient protocol for achieving this transmission, thus establishing the functional viability of such an event channel.

Evaluation of Invisibility

Adversarial Statistical Steganalysis To simulate a sophisticated warden within the IoA network, we trained a steganalysis classifier (Sun et al. 2025) for the task of event-level steganalysis. The classifier was trained on a meticulously constructed dataset of paired benign and steganographic events, as detailed in the extended version (Huang et al. 2025). We used the Area Under the Curve (AUC) as the primary metric for evaluating the classifier’s performance during training. From the results in Figure 3, which illustrates the classifier’s performance on the held-out test set, we can draw the following conclusion. Firstly, the Receiver Operating Characteristic (ROC) curve is nearly identical to the diagonal baseline, which represents a random classifier. The resulting AUC of 0.503 is statistically indistinguishable from the 0.5 value of random guessing, indicating that the classifier has virtually no discriminative power to distinguish a covert event from a benign one based on its features. Secondly, this conclusion is strongly corroborated by the individual classification metrics. The low Precision (0.250) and F1-score (0.333) demonstrate the classifier’s complete inability to reliably identify the positive (stegano-

Topic	Metric	Discop				Differential			
		Llama3		Qwen2		Llama3		Qwen2	
		512	1024	512	1024	512	1024	512	1024
Art	EBPR \uparrow	68.63 \pm 4.03	80.03 \pm 3.32	67.27 \pm 2.21	74.66 \pm 0.93	60.28 \pm 3.26	58.95 \pm 5.86	58.34 \pm 6.93	67.33 \pm 3.51
	BPR \uparrow	85.82 \pm 4.63	97.97 \pm 3.77	84.26 \pm 2.55	91.87 \pm 1.05	76.22 \pm 3.75	74.01 \pm 6.66	74.00 \pm 7.95	83.54 \pm 3.99
	Entropy \uparrow	1.72 \pm 0.10	1.71 \pm 0.14	1.65 \pm 0.04	1.68 \pm 0.08	1.68 \pm 0.17	1.67 \pm 0.08	1.75 \pm 0.08	1.72 \pm 0.09
QA	EBPR \uparrow	63.50 \pm 6.71	79.93 \pm 10.51	60.43 \pm 0.97	72.69 \pm 5.05	52.56 \pm 4.69	63.11 \pm 9.98	59.80 \pm 3.45	66.87 \pm 3.61
	BPR \uparrow	79.93 \pm 7.71	97.85 \pm 11.95	76.40 \pm 1.11	89.63 \pm 5.74	67.35 \pm 5.39	78.74 \pm 11.34	75.68 \pm 3.96	83.02 \pm 4.10
	Entropy \uparrow	1.71 \pm 0.11	1.76 \pm 0.06	1.74 \pm 0.09	1.81 \pm 0.07	1.64 \pm 0.06	1.77 \pm 0.03	1.79 \pm 0.08	1.74 \pm 0.10
Philosophy	EBPR \uparrow	84.59 \pm 11.00	74.77 \pm 10.16	72.80 \pm 4.25	89.72 \pm 6.61	72.89 \pm 9.10	69.29 \pm 4.06	61.72 \pm 4.61	67.56 \pm 5.76
	BPR \uparrow	104.15 \pm 12.63	91.99 \pm 11.54	90.61 \pm 4.89	108.99 \pm 7.51	90.72 \pm 10.45	85.76 \pm 4.61	77.88 \pm 5.29	83.79 \pm 6.55
	Entropy \uparrow	1.72 \pm 0.07	1.69 \pm 0.05	1.74 \pm 0.06	1.72 \pm 0.06	1.61 \pm 0.14	1.65 \pm 0.02	1.73 \pm 0.06	1.76 \pm 0.03

Table 2: Evaluation of Capacity and Imperceptibility Metrics. The table shows the Effective Bits Per Round (EBPR) and total Bits Per Round (BPR), with Entropy (bits per token) included as an auxiliary measure of statistical imperceptibility.

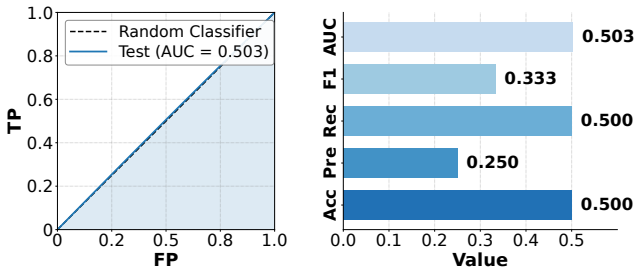


Figure 3: Steganalysis results on the statistical detection.

Method	Topic	Nat. \uparrow	Coh. \uparrow	Susp. \downarrow
Normal	Art	5.00 \pm 0.00	5.00 \pm 0.00	1.00 \pm 0.00
	QA	4.17 \pm 1.04	4.17 \pm 1.04	1.67 \pm 0.58
	Philosophy	4.67 \pm 0.29	5.00 \pm 0.00	1.34 \pm 0.29
Discop	Art	5.00 \pm 0.00	4.95 \pm 0.10	1.00 \pm 0.00
	QA	5.00 \pm 0.00	5.00 \pm 0.00	1.00 \pm 0.00
	Philosophy	4.95 \pm 0.10	4.95 \pm 0.10	1.06 \pm 0.10
Differential	Art	4.84 \pm 0.17	4.89 \pm 0.19	1.17 \pm 0.29
	QA	4.78 \pm 0.26	4.89 \pm 0.19	1.11 \pm 0.19
	Philosophy	4.86 \pm 0.24	5.00 \pm 0.00	1.00 \pm 0.00

Table 3: LLM-as-a-Judge (GPT-4o) evaluation results. Average scores (1-5) for Naturalness (Nat.), Coherence (Coh.), and Suspicion (Susp.).

graphic) class. Collectively, it validates that our method successfully conceals the steganographic artifacts within the statistical patterns of normal agent communication, holding up even against a powerful, learning-based adversary designed specifically for this detection task.

Adversarial Intention Steganalysis To simulate the judgment of an intelligent agent, we employed an LLM-as-a-Judge (GPT-4o) to evaluate dialogue snippets. These snippets, extracted using a three-turn sliding window, were assessed on their local naturalness, coherence, and suspiciousness, thereby testing the protocol’s resilience to context-

tual scrutiny. The detailed prompt is available in the extended version (Huang et al. 2025). From the results in Table 3, which presents a comparison of average scores for our steganographic methods against a benign baseline, we can draw the following conclusions. Firstly, both the Discop and Differential methods achieve high Naturalness and Coherence scores (approaching 5.0) that are statistically comparable to the benign baseline, indicating the steganographic dialogues are fluent and logically sound. Secondly, the Suspicion scores for both methods are extremely low and show no significant statistical difference from the benign baseline, demonstrating that the LLM judge found no reliable signal to suspect hidden communication. Thirdly, this result shows our protocol satisfies the more stringent IND-INT security guarantee by not only preserving statistical distributions but also emulating plausible conversational intent, rendering the dialogues innocuous to an intelligent observer.

Discussion and Limitations

Our model is general and modality-agnostic, but current evaluation is limited to conversational text; we will extend to multi-modal settings and richer action spaces. The analysis assumes a stable natural behavior distribution \mathcal{D} and a passive adversary, motivating adaptive protocols for non-stationary environments and active attacks. We provide strong empirical evidence toward IND-INT, yet a rigorous compositional proof across storage, timing, and behavioral channels, and broader evaluations with stronger baselines and ablations are left for future work.

Conclusion

We formalize the Covert Event Channel for the Internet of Agents by unifying storage, timing, and behavioral dimensions, and implement Π_{CCAP} to operationalize it. Experiments demonstrate high capacity, robustness, and imperceptibility under strong wardens, supported by formal security notions. Our results establish the channel’s practicality and lay the groundwork for future monitoring and defensive protocols in secure, high-stakes agent ecosystems.

Ethical Statement

This work aims to provide safer and more effective protection for information transfer in high-value agent interactions (e.g., finance, healthcare, public services, and critical infrastructure). Our intent is strictly defensive, aiming to mitigate surveillance and traffic-analysis risks so that mission-critical confidentiality is preserved while utility and reliability are maintained. Experiments use publicly available models and synthetic or public corpora; no human subjects were recruited and no PII was processed. We did not bypass platform safeguards, and all evaluations were conducted offline without deployment to real systems.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under Grant 2023YFC3305402, the National Natural Science Foundation of China (Nos.62302059 and 62172053) and the Guangdong Basic and Applied Basic Research Foundation under Grant 2022A1515110693.

References

- Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.
- Dhobale, D.; Ghorpade, V.; Patil, B.; and Patil, S. 2010. Steganography by hiding data in TCP/IP headers. In *2010 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE)*, volume 4, V4–61. IEEE.
- Ding, J.; Chen, K.; Wang, Y.; Zhao, N.; Zhang, W.; and Yu, N. 2023. Discop: Provably secure steganography in practice based on “distribution copies”. In *2023 IEEE Symposium on Security and Privacy (SP)*, 2238–2255. IEEE.
- Google. 2025. Agent2agent protocol. Accessed: 2025.
- Huang, K.; Wei, Y.; Wu, W.; Zhang, T.; Yang, Z.; and Zhou, L. 2025. Whispering Agents: An event-driven covert communication protocol for the Internet of Agents. *arXiv preprint arXiv:2508.02188*.
- Kaptchuk, G.; Jois, T. M.; Green, M.; and Rubin, A. D. 2021. Meteor: Cryptographically secure steganography for realistic distributions. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 1529–1548.
- Kong, D.; Lin, S.; Xu, Z.; Wang, Z.; Li, M.; Li, Y.; Zhang, Y.; Sha, Z.; Li, Y.; Lin, C.; et al. 2025. A Survey of LLM-Driven AI Agent Communication: Protocols, Security Risks, and Defense Countermeasures. *arXiv preprint arXiv:2506.19676*.
- Li, J.; Lai, Y.; Li, W.; Ren, J.; Zhang, M.; Kang, X.; Wang, S.; Li, P.; Zhang, Y.-Q.; Ma, W.; et al. 2024. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*.
- Liao, G.; Yang, J.; Shao, W.; and Huang, Y. 2025. A Framework for Designing Provably Secure Steganography.
- Motwani, S. R.; Baranchuk, M.; Hammond, L.; and de Witt, C. S. 2023. A Perfect Collusion Benchmark: How can AI agents be prevented from colluding with information-theoretic undetectability? In *Multi-Agent Security Workshop@ NeurIPS’23*.
- Motwani, S. R.; Baranchuk, M.; Strohmeier, M.; Bolina, V.; Torr, P.; Hammond, L.; and de Witt, C. S. 2024. Secret Collusion among AI Agents: Multi-Agent Deception via Steganography. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Park, J. S.; O’Brien, J.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, 1–22.
- Shajarian, S.; Khorsandroo, S.; and Abdelsalam, M. 2024. A Survey on Self-Running Networks: Concepts, Components, Opportunities, and Challenges. *Authorea Preprints*.
- Sun, M.; Yin, Y.; Xu, Z.; Kolter, J. Z.; and Liu, Z. 2025. Idiosyncrasies in Large Language Models. In *Forty-second International Conference on Machine Learning*.
- Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; Millican, K.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- Wang, Y.; Guo, S.; Pan, Y.; Su, Z.; Chen, F.; Luan, T. H.; Li, P.; Kang, J.; and Niyato, D. 2025. Internet of agents: Fundamentals, applications, and challenges. *arXiv preprint arXiv:2505.07176*.
- Yang, Z.-L.; Guo, X.-Q.; Chen, Z.-M.; Huang, Y.-F.; and Zhang, Y.-J. 2018. RNN-stega: Linguistic steganography based on recurrent neural networks. *IEEE Transactions on Information Forensics and Security*, 14(5): 1280–1295.
- Yu, M.; Meng, F.; Zhou, X.; Wang, S.; Mao, J.; Pang, L.; Chen, T.; Wang, K.; Li, X.; Zhang, Y.; et al. 2025. A survey on trustworthy llm agents: Threats and countermeasures. *arXiv preprint arXiv:2503.09648*.
- Yu, Y.; Li, H.; Chen, Z.; Jiang, Y.; Li, Y.; Zhang, D.; Liu, R.; Suchow, J. W.; and Khashanah, K. 2024. Finmem: A performance-enhanced llm trading agent with layered memory and character design. In *Proceedings of the AAAI Symposium Series*, volume 3, 595–597.