ANALYSIS

When Encryption Isn't Enough

Understanding AI Side-Channel Exposure and How to Eliminate It

Introduction

Artificial intelligence systems are rapidly becoming embedded in mission-critical environments ranging from national security operations and intelligence analysis to enterprise decision-making and automated workflows. As adoption accelerates, so does the assumption that these systems are secure if their data is encrypted.

That assumption is now demonstrably false.

Recent advances in side-channel analysis (most notably the class of attacks exemplified by the Microsoft “Whisper Leak” attack: https://arxiv.org/abs/2511.03675) have revealed that sensitive information can be inferred from AI systems without decrypting a single byte of data. Instead, adversaries observe how systems communicate: the timing, size, and structure of encrypted traffic. From these signals, they can reconstruct meaning.

This represents a fundamental shift in cybersecurity. The risk is no longer limited to data exposure. It now includes **exposure of intent, context, and decision-making behavior**.

For organizations relying on AI (particularly those in defense, intelligence, and critical infrastructure) this changes the threat model entirely.

The Nature of the Vulnerability

To understand the problem, it is important to recognize how modern AI systems communicate.

Large Language Models (LLMs) and related systems typically generate responses incrementally. As tokens are produced, they are transmitted in a continuous stream over encrypted channels such as TLS. While the content is protected, the transmission itself retains structure. Each response exhibits a distinctive pattern:

- The cadence of output as tokens are generated
- The size and frequency of packets
- The overall “shape” of a session over time

These patterns are not random. They are a direct byproduct of how models process and generate language. Side-channel attacks exploit this structure. By observing encrypted traffic and applying statistical and machine learning techniques, an adversary can correlate these patterns with known behaviors.

Over time, this enables inference of:

- The subject matter of a query
- The type of response being generated
- The operational context of the user

In effect, the system is leaking meaning through its behavior. Furthermore, encrypted channels still expose source and destination IP address making targeting simple. This is not a flaw in encryption. It is a consequence of **observable consistency in how data is transmitted**.

Why This Matters

The implications extend far beyond academic concern. In a commercial setting, this form of exposure enables competitors or adversaries to infer proprietary activities (e.g., product development, strategic planning, or internal analysis) without ever breaching a system. In government and defense environments, the stakes are significantly higher. Observing AI-assisted workflows could reveal:

- Intelligence priorities and areas of focus
- Targeting or investigative activity
- Operational tempo and intent
- Location of field agents and clandestine operations

Even partial inference can be sufficient to provide adversaries with actionable intelligence. Importantly, these attacks are passive. They do not require system compromise, malware, or user interaction. They rely only on the ability to observe network traffic (something that is often assumed to be safe under encryption).

The Expansion of Risk in Agentic AI

The rise of Agentic AI amplifies this problem.

Unlike traditional AI interactions, which involve a single prompt and response, Agentic systems operate through sequences of actions. They may call external tools, access databases, interact with APIs, and iterate toward a goal. Each step generates additional communication, often across multiple systems. This creates a continuous, multi-stage signal that can be observed and analyzed.

Recent incidents involving AI-driven systems have demonstrated that these workflows can be manipulated or inferred through indirect means. When combined with side-channel techniques, the exposure increases significantly. An observer is no longer limited to a single interaction but can track an evolving process (effectively watching the system "think" over time). This transforms AI from a discrete risk into a persistent one.

What Is Actually Being Exploited

At its core, this class of attacks is not exploiting software bugs or cryptographic weaknesses. It is exploiting structural properties of modern communications:

- **Continuity**: Data flows as a coherent, observable stream
- **Consistency**: Similar inputs produce similar transmission patterns
- **Correlation**: Timing and size reflect internal processing
- **Stability**: Sessions maintain identifiable endpoints and structure

These properties allow adversaries to build models that map observable signals to underlying meaning. Traditional defenses do not address these characteristics. Encryption protects the content of communication but does not conceal its form. Network security architectures assume that observing encrypted traffic is acceptable. Techniques such as padding or traffic shaping attempt to obscure patterns but cannot eliminate them entirely.

As a result, the fundamental signal remains available for analysis.

A Different Approach: Eliminating Observability

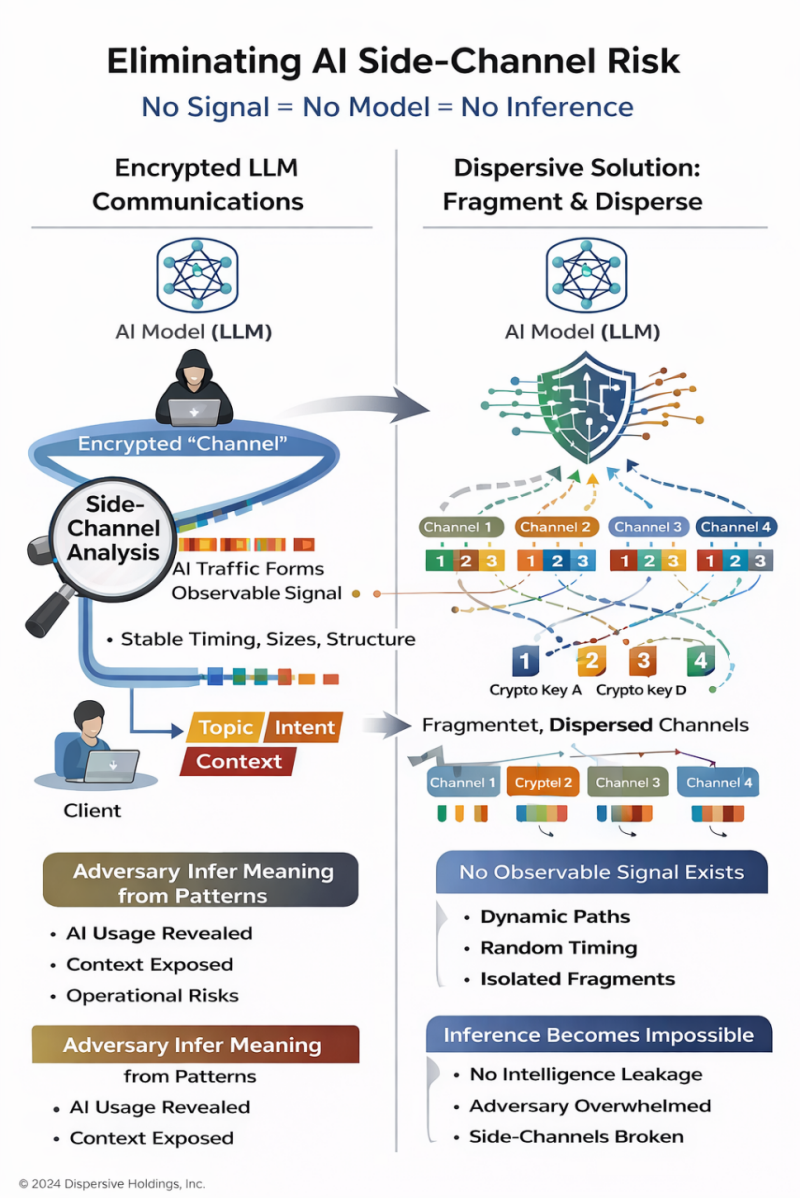
Addressing this problem requires a shift in perspective.

If the vulnerability arises from observable structure, then the solution is not to better hide that structure; it is to **eliminate it altogether**.

This is the principle behind **Dispersive® Stealth Networking's** approach.

Rather than transmitting data as a single, continuous stream, Dispersive transforms the communication process itself. Data is segmented into multiple fragments and distributed across independent network paths. These fragments are transmitted with varying timing, sequencing, and routing characteristics, and are only reassembled at the authorized endpoint. Dispersive also encrypts the payload and headers of the fragments independently per channel making unauthorized reassembly practically impossible while making the sources and destinations invisible and untargetable.

From an external perspective, the original session no longer exists as a coherent entity. There is no stable stream to observe, no consistent timing pattern, and no reliable way to associate fragments with one another.



Why This Changes the Outcome

This architectural approach directly targets the conditions required for side-channel inference. When data is split across multiple paths, an observer cannot reconstruct the full communication without simultaneous visibility into all paths, which is an increasingly impractical requirement at scale.

- When timing is randomized and influenced by independent network conditions, the relationship between transmission patterns and model behavior breaks down.
- When packet sizes are decoupled from the underlying content, statistical analysis loses its primary signal.
- When sessions do not maintain stable identifiers or endpoints, tracking and correlation become unreliable.

The result is not simply reduced risk; it is the collapse of the adversary's ability to form a usable model. Without consistent signals, machine learning classifiers cannot converge. Without convergence, inference fails. This represents a shift from attempting to defend against specific attack techniques to removing the underlying opportunity for those techniques to function.

Implications for AI Security

For organizations deploying AI, this has several important implications. First, it reframes the definition of secure communication. Protecting data in transit is no longer sufficient. It is necessary to protect the **observable characteristics of that data**.

Second, it highlights a gap in current architectures. Zero Trust, encryption standards, and endpoint protections all assume that the network can be observed without consequence. That assumption does not hold in the context of AI-driven systems.

Third, it introduces a path forward. By addressing the structural conditions that enable inference, it becomes possible to secure AI systems against a class of attacks that would otherwise remain persistent and evolving.

Conclusion

The emergence of side-channel attacks against AI systems marks a turning point in cybersecurity. It demonstrates that sensitive information can be exposed not through failure of encryption, but through the inherent properties of how systems communicate. As AI becomes more deeply integrated into operational and decision-making processes, this form of exposure will only grow in significance. The critical question is no longer whether encrypted traffic can be intercepted, but whether it can be understood. Architectures that preserve observable structure will continue to leak meaning, regardless of how strong their encryption is. Architectures that eliminate that structure fundamentally change what is possible for an adversary.

Dispersive represents one such approach. By removing the continuity and consistency required for inference, it addresses the problem at its source. Furthermore, Dispersive increases resiliency and typically increases throughput while reducing latency. You can see Dispersive defend against the Whisper Leak side-channel attack in this video: https://www.loom.com/share/d86bc3225a684dd8bcf62d8221c1cdb3.

For organizations that depend on AI—particularly in environments where intent and context are as sensitive as data—this distinction is not theoretical. It is operational.

Dispersive Holdings, Inc.
300 Colonial Center Pkwy Suite 100 | Roswell, GA 30076 USA
dispersive.io | 1.844.403.5850

dispersive