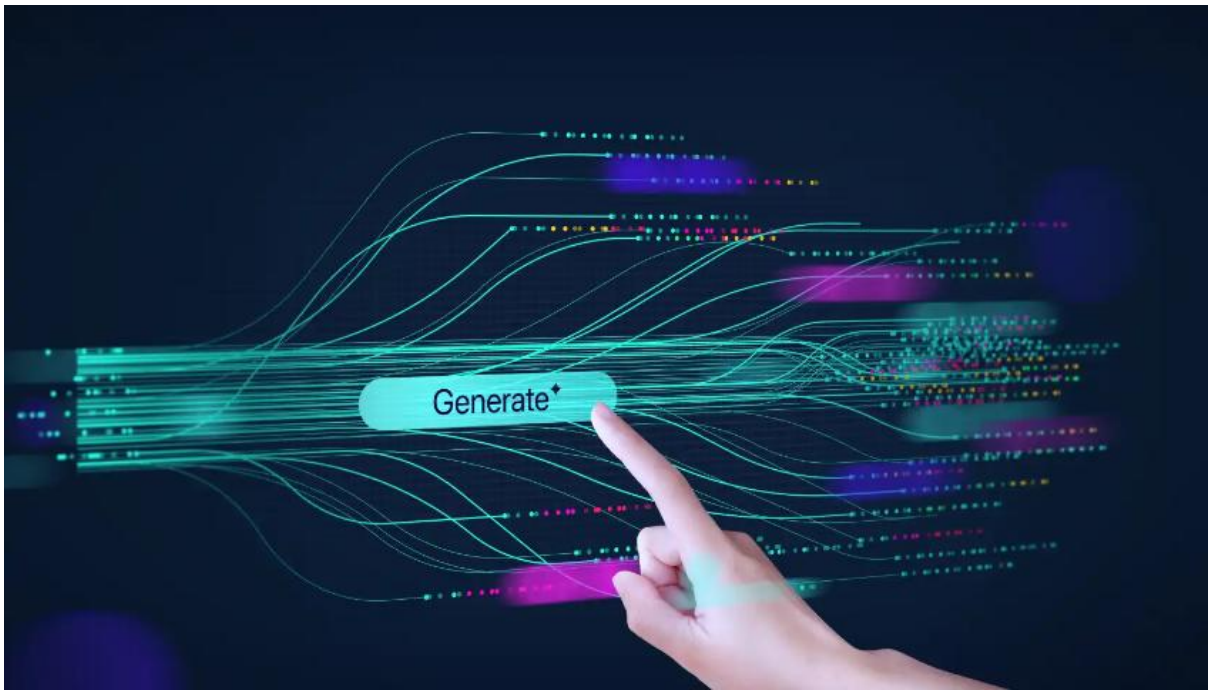
PRIMUS PARTNERS®

Quote by Devroop Dhar, Co-founder & CEO, Primus Partners

Published in CSO
November 10, 2025

Whisper Leak uses a side channel attack to eavesdrop on encrypted AI conversations

Encryption alone is no longer sufficient to protect privacy in LLM interactions, as metadata patterns can be exploited to infer sensitive subjects and corporate intent.



Read on: https://www.csoonline.com/article/4087335/whisper-leak-uses-a-side-channel-attack-to-eavesdrop-on-encrypted-ai-conversations.html

Article Content:

Researchers at Microsoft have revealed a new side channel attack named Whisper Leak that can reveal the topic of encrypted conversations between users and language models, even without access to the underlying text.

The discovery highlights a growing blind spot in AI security where encryption alone no longer guarantees privacy in model interactions.

Microsoft's Security Defender Security Research team said that attackers are in a position to exploit large language models that use metadata such as network packet sizes and timings. For instance, a nation-state actor at the internet service provider layer,

someone on the local network, or someone connected to the same Wi-Fi router can observe the encrypted traffic and use it to infer if the user's prompt is on a specific topic.

Metadata becomes the new attack surface

Unlike traditional data breaches or model leaks, Whisper Leak exploits a side channel in network communication rather than a flaw in encryption itself.

LLM services generate responses step by step, by producing one token at a time instead of the entire response at once. Also, the communications with AI-powered chatbots are often encrypted with HPPS over TLS (HTTPS), ensuring the authenticity of the server and security through encryption.

But while the Transport Layer Security successfully encrypts the content of communications, it leaks the size of the underlying data chunks being transmitted. For an LLM that streams responses token by token, this size information reveals patterns about the tokens being generated.

Combined with timing information between packets, these leaked patterns form the basis of the Whisper Leak attack as sufficient information is leaked to enable topic classification, explained Microsoft Defender Security Team in the technical report.

"These are not usual data breaches either. They do not steal the files directly; they observe what is happening around the data," said Devroop Dhar, co-founder & MD at Primus Partners. "They don't have to break encryption or code. What they do instead is look for small clues; timing, lags, maybe how quickly a system answers, and from that, they try to understand what's going on inside. It's very technical and tough to catch when it's happening," he added.

Inside Microsoft's proof-of-concept

Researchers at Microsoft simulated a real-world scenario in which the adversary could observe encrypted traffic but not decrypt it. They chose "legality of money laundering" as the target topic for the proof-of-concept.

For positive samples, the team used a language model to generate 100 semantically similar variants of questions about this topic. For negative noise samples, it randomly sampled 11,716 unrelated questions from the Quora Questions Pair dataset, covering a wide variety of topics.

Once done, the collected data was trained using LightGBM, Bi-LSTM, and BERT-based models, evaluated in time-only, packet-size only, or both modes.

The research team demonstrated the attack across 28 popular LLMs from major providers, and achieved near-perfect classification (often >98% Area Under the Precision-Recall Curve (AUPRC)) and high precision even at extreme class imbalance

(10,000:1 noise-to-target ratio). For many models, they achieved 100% precision in identifying sensitive topics while recovering 5-20% of target conversations, noted the report.

Plugging the leaks

The findings were shared with OpenAI, Mistral, Microsoft, and xAI, and mitigation measures were implemented to minimise the risk. To mitigate the effectiveness of cyberattacks, OpenAI, and later Microsoft Azure, added a random sequence of text of variable length to each response.

This obfuscation field masked the length of each token, reducing the attack's effectiveness. Similarly, Mistral included a new parameter called "p" that had a similar effect.

CISO's next frontier

Even if the attack doesn't expose the exact prompt or content of a conversation, it can accurately classify its subject or intent, putting enterprises at major risk.

"If an LLM is just handling public data, it is fine. But if it is processing data like client records, internal documents, financial data, etc, then even a small leak matters. The bigger worry is for companies that run their own AI models or connect them to cloud APIs. Like banks, healthcare, legal firms, defence, where data sensitivity is too high," Dhar said.

While it is the AI providers that will have to address the issue, Microsoft researchers' recommendations include avoiding discussing highly sensitive topics over AI chatbots when on untrusted networks, using VPN services for adding an additional layer of protection, opting for providers that have already implemented mitigation, and using non-streaming models of large language model providers.

Dhar pointed out that most AI security checklists do not even mention side channels yet. CISOs need to start asking their teams and vendors how they test for these kinds of probable issues.

"Also, in order to be defensive, we need to keep models isolated, add a bit of random delay so timing data is not predictable, and watch for weird or repeated queries that look like probing. Basically, we need to treat the AI pipeline the way we would treat a critical server, by following a few simple steps like logging it, segmenting it, and not assuming that it is invisible just because it is encrypted," he added. Over time, we will need proper "AI pen-testing," like what happened when cloud APIs first became mainstream. It is the same pattern, once the tech matures, attackers get creative and then security always has to catch up, he noted.