Whisper Leak: fuite de données dans les accès à distance aux modèles de langue

Olivier Schwander <olivier.schwander@sorbonne-universite.fr>

2026

Un récent article (Geoff McDonald and Or 2025) décrit une attaque par canal auxiliaire contre les modèles de langue utilisés avec un accès à distance (donc avec l'interface web d'un chatbot ou au travers d'une API). Cette attaque permet de d'inférer le thème d'une conversation avec un chatbot en observant uniquement le trafic chiffré entre le client et le serveur. Il n'est pas nécessaire de déchiffrer les données : les features exploitées sont la taille des paquets et les intervalles de temps entre leur envoi. Même si le texte clair n'est jamais révélé, les informations extraites peuvent être exploitées de manières plus ou moins dangereuses (surveillance étatique, violations du secret médical, violences anti-LGBTQ, répression de l'IVG, publicité ciblée, etc).

Les objectifs du projet sont les suivants :

— explorer la littérature concernant ce type d'attaques ;
— construire un dataset en interceptant du trafic chiffré ;
— reproduire les expériences de l'article ;
— étudier l'impact de différents modèles ou de différentes API ;
— valider les techniques de mitigation proposées.

Il s'agit bien d'un projet relevant du machine learning, qui ne nécessite pas de connaissances particulières en cryptographie ou en cybersécurité.

Voir également l'article de blog sur le site de Microsoft [1].

Bibliographie

Carlini, Nicholas, and Milad Nasr. 2024. "Remote Timing Attacks on Efficient Language Model Inference." October 22, 2024. https://doi.org/10.48550/arXiv.2410.17175.

McDonald, Geoff, Jonathan Bar Or. 2025. " : Aa Novel Side-Channel Attack on Remote Language Models." Microsoft Security Blog. November 7, 2025. https://www.microsoft.com/en-us/security/blog/2025/11/07/whisper-leak-a-novel-side-channel-cyberattack-on-remote-language-models/.

McDonald, Geoff, and Jonathan Bar Or. 2025. "Whisper Leak : A Side-Channel Attack on Large Language Models." November 5, 2025. https://doi.org/10.48550/arXiv.2511.03675.

Weiss, Roy, Daniel Ayzenshteyn, Guy Amit, and Yisroel Mirsky. 2024. "What Was Your Prompt ? A Remote Keylogging Attack on AI Assistants." March 14, 2024. https://doi.org/10.48550/arXiv.2403.09751.

Zhang, Tianchen, Gururaj Saileshwar, and David Lie. 2024. "Time Will Tell : Timing Side Channels via Output Token Count in Large Language Models." December 19, 2024. https://doi.org/10.48550/arXiv.2412.15431.

Zheng, Xinyao, Husheng Han, Shangyi Shi, Qiyan Fang, Zidong Du, Xing Hu, and Qi Guo. 2024. "InputSnatch : Stealing Input in LLM Services via Timing Side-Channel Attacks." November 29, 2024. https://doi.org/10.48550/arXiv.2411.18191.

1. https://www.microsoft.com/en-us/security/blog/2025/11/07/whisper-leak-a-novel-side-channel-cyberattack-on-remote-language-models/