Projet de Recherche : Whisper Leak

Comparaison d'articles scientifiques

Contexte et Objectifs

Problème :

Whisper Leak

Metadata from encrypted LLM streaming responses can be used to identify specific sensitive topics discussed by users.

Input Snatch

Le trafic généré par les LLMs réutiliserait le cache et cela permettrait de prédire l'input de l'utilisateur.

What Prompt

Encrypted AI responses leak information via packet sizes.

Packet payload size reveals token lengths during streaming.

-> enables inference of sensitive content.

Time Will Tell

Le trafic crypté généré par les LLMs révèleraient des informations sur la langue parlée par l'utilisateur ainsi que des données confidentielles le concernant.

Remote Timing

Vulnerabilities in consumer-grade (via API or front-end) LLMs to data-dependent timing attacks

Hypothèse :

Whisper Leak

(focus : topic inference)

Sequences of encrypted packet sizes and the timing intervals between their arrivals contain enough information to classify the topic of the initial prompt.

Input Snatch

Le recyclage d'états ou de réponses mises en cache affecte les temps de réponse et peut être utilisé en tant qu'indice pour une attaque basée sur le timing.

Des requêtes différentes sur les mêmes noeuds computationnels utilisent le même cache.

What Prompt

(focus : token-length leakage)

Size of the encrypted network packets is directly correlated to the length of each token.

-> Because LLMs stream tokens sequentially, differences in encrypted packet payload sizes reveal each token's length.

Time Will Tell

- Tâches de traduction :

Chaque langue a une densité de token et des ratios input/output différents qui permettent d'identifier la langue de l'utilisateur.

- Tâches de classification :

Les LLMS exhibent des biais pour certaines classes spécifiques (qui marquent encore plus la différence du nombre d'output tokens).

Remote Timing

Due to the nature of LLMs, it is possible to execute multiple variants of remote timing attacks in the inference of language models.

Objectif(s) :

Whisper Leak

- Demonstrate a systemic vulnerability across the industry, evaluate its success rate, and test various defensive measures.
- Model realistic adversaries and quantify precision under extreme class imbalance.

Input Snatch

- Mise en lumière des vulnérabilités associées aux optimisations de performance
- Identification de patterns dans le cache afin de reconstruire l'input.

What Prompt

- Accurately reconstruct plaintext responses from token-length sequences.
- Provide a comprehensive framework for understanding and mitigating the risks associated with the token-length side-channel

Time Will Tell

Identification et classification des réponses du trafic crypté en utilisant les longueurs des tokens et le nombre de tokens émis.

Remote Timing

The main goals are formalizing the threat model, demonstrate that the models of the main industrial actors (OpenAI, Claude, ...) are vulnerable and reproduce them, using a variety of elements (mostly inspection of packets and inference).

Type d’attaque :

Whisper Leak

Passive side-channel attack based on the analysis of network traffic metadata

- based on packet sizes and timing patterns

Input Snatch

Attaque par canal auxiliaire basée sur le timing en utilisant

- Le caching du préfixe
- Le caching sémantique

What Prompt

Passive network side-channel attack (remote keylogging–style) that reconstructs responses via token-length inference.

Time Will Tell

Principalement attaque par canal auxiliaire et analyse du trafic et du timing des output tokens

- Tâches de traduction :

Tokenizers basés sur l’encodage d’octets par paire (BPE - Byte Pair Encoding)

- Attaque temporalisé par canal auxiliaire end-to-end :

Obtention de la longueur et du timing des paquets d’input et d’output en étant placé sur le réseau qui se trouve entre le serveur de l’application et la victime

Remote Timing

Traffic analysis using a classifier that runs on the response timings.

Approche et Contributions

Contributions clés :

Whisper Leak

- Identifies a new class of topic inference attacks.
- Performs industry-wide risk assessment.
- Demonstrate the attack across 28 popular LLMs from major providers.

* mitigations currently deployed provide only partial protection.

Input Snatch

- Analyse du risque de leaks de données avec les compromis performance/ confidentialité
- Implémentation d'une méthodologie d'attaque
- Reconstruction de l'input au lieu de l'output

What Prompt

- Identifies a novel side-channel inherent in all LLM models.
- Develops a framework for token extraction.
- Use of LLMs to reconstruct encrypted plaintext.

  (context-aware multi-sentence inference.)
- Known-plaintext fine-tuning to improve accuracy.

Time Will Tell

- Identification d'une nouvelle attaque par canal auxiliaire due à la variation du nombre d'output tokens
- Démonstration qu'un biais de tokenizer et qu'un biais de nombre de tokens existe dans la classification de texte et pose un risque de confidentialité

Remote Timing

- Formalizes threat model of timing attacks efficient language model inference
- Demonstrates that production language models today use efficient inference techniques and that they are vulnerable
- Constructed timing attacks on both passive and active settings

Pipeline :

Whisper Leak

1. Passive network adversary observes encrypted TLS traffic between user and LLM service.

2. Extraction of packet sizes and timing sequences.

3. Use of trained classifiers to infer whether the conversation topic matches a sensitive category.

Input Snatch

Établissement de patterns des targets en utilisant une méthode d'échantillonage stratégique et mitigation du bruit avec des algorithmes "proposed point processing". Mitigation du grand espace de recherche avec des techniques avancées de machine-learning pour extraire les données contextuelles et les relations sémantiques à partir de datasets open-sources (pour améliorer l'efficacité de construction).

Incorporation d'une évaluation en plusieurs étapes pour prioriser les candidats ayant de plus fortes probabilités de trouver le bon cache

What Prompt

1. Intercept encrypted traffic.

2. Extract token-length sequence from packet sizes.

3. Segment sequence into sentence-like units.

4. Use two LLMs to reconstruct text sequentially.

Time Will Tell

Attaque en 2 phases :

-Profilage : Utilisation du nombre de tokens d'output ainsi que la longueur totale en octets de l'input pour établir un seuil de distinction entre les deux classes de prédiction.

-Attaque : L'attaquant monitore la longueur en octets de l'input et le nombre de tokens d'output de la requête de classification d'un utilisateur. Selon le seuil de distinction, l'attaquant peut déterminer (par prédiction grâce au profilage) la classe prédite et s'en servir.

Remote Timing

Multiple attacks tested on the paper, but overall,

Phase of building a prompt set with two differentiate topics -> generate human-like responses and query the victim LLM with them during multiple rounds -> Use Gaussian Mixture Models to infer the topic

Side-channel Exploité:

Whisper Leak

Metadata of packet sizes and the time intervals between packets that remain visible despite encryption

- modern TLS preserves size relationship between plaintext and ciphertext

-> reveals patterns about tokens being generated

Input Snatch

Réutilisation du cache par les LLMs pour répondre à des queries similaires

Cette réutilisation affecte le temps d' émission des paquets et permet en cas d'attaque d'identifier des patterns dans le cache et dans le timing des paquets.

What Prompt

Token-length side-channel

-> packet sizes leak the number of characters in each generated token

Time Will Tell

- Attaque sur tâches de traduction :

Utilisation du BPE pour connaître la fréquence des tokens

Attaquant envoie 1000 requêtes par langue au LLM pour qu'il les traduise et mesure la densité de l'output token ainsi que le ratio Output-Input

- Attaque sur tâche de classification :

Observation ou estimation du nombre d'output tokens

- Attaque temporalisé par canal auxiliaire end-to-end :

Récupération des paquets et leurs tokens

Remote Timing

Inter-packet delay and a pre-trained classificator

Cadre d’Application

Techniques d'inférence :

Whisper Leak

Combination of binary classifiers (trained on packet sizes and timing sequences)

- gradient boosting (LightGBM), recurrent neural networks (Bi-LSTM), and transformers (BERT)

Input Snatch

Exploration d'un espace d'input inconnu et extensif pour trouver l'input mis en cache de l'utilisateur (parcours d'arbre binaire)

Focus sur les centroïdes des clusters et exploration près des clusters également.

What Prompt

Fine-tuned T5 transformer (pre-trained)

- has an expanded token vocabulary
- provided with context of previously inferred sentences
- ranking mechanism for multiple generated options

Time Will Tell

- Classifieur binaire des textes cryptés avec un LLM en utilisant le Predict then Explain (P-E).
- Variation des échantillons de l'attaquant pour augmenter la prédiction.
- Modèles de mixture gaussien (GMM)

Remote Timing

Gaussian Mixture Models as a classifier for the goodness-of-fit.

Dialogue :

Whisper Leak

Targets real-time token streaming, where the model sends pieces of a response as they are generated.

* LLMs compute and produce tokens one by one

-> sent immediately or in-batch

training dialogue :

- target topic (variants of a sensitive topic)

- noise/background (unrelated questions from diverse topics-Quora)

Input Snatch

L'attaquant envoie ses propres requêtes (sans voir celle de la victime)

What Prompt

- Streaming responses in conversational AI.

- Multi-turn dialogue context.

- Alternating prompts and responses.

Time Will Tell

Cible le nombre de tokens de sortie produits par les LLMs.

Remote Timing

Pre-generated generalistic prompts or random, disjoin, prompts.

Cibles :

Whisper Leak

28 commercially available LLMs from major providers

- OpenAI (gpt-4o-mini, o1-mini...),
- Microsoft (deepseek-r1), DeepSeek, Mistral, X.AI, Alibaba

Input Snatch

- OpenAI
- DeepSeek
- vLLM 0.6.2
- LLaMA-2 70B

What Prompt

- OpenAI's ChatGPT-4
- Microsoft's Copilot

(both browser and API traffic)

Time Will Tell

- Gemma 2-2B
- Gemma 2-9B
- Gemma 2-27B
- Llama 3.1-8B
- Llama3.2-3B
- GPT-4o
- GPT-4o mini

Remote Timing

Simulated environment with an exposed FastChat API and a "victim vs adversary setup"

- Gpt-3.5
- Gpt-4
- Claude 3 Sonnet
- Claude 3 Haiku

For the production models, both API and front-end are targets of the attack.

Accès :

Whisper Leak

Passive observation of the network traffic between the user and LLM provider.

Input Snatch

Canaux API cryptés routés pour partager le cache avec l'attaquant.

What Prompt

Network-level access

(within the Local Area Network (LAN) of user or somewhere in the internet infrastructures)

Time Will Tell

Données récoltées de manière passives en utilisant des datasets à disposition pour la requête et réponse générée par LLMs testées (utilisation d'un ping request pour obtenir les temps d'aller-retour).

Remote Timing

Multiple use cases, legit user, black-box access, white-box access, no access at all (just to packets obtained from a MITM)

Interception :

Whisper Leak

Cloud hosted-hosted Ubuntu machines running the packet capture tool tcpdump

-> to record TLS traffic generated during the LLM response.

Input Snatch

Pas d'interception, l'attaquant envoie ses propres requêtes et mesure les temps de réponse pour trouver l'input de l'utilisateur.

What Prompt

Eavesdropping on protocols like :

- QUIC over UDP (chatgpt)
- WebSockets over TCP (copilot)

(often by filtering for known server IP addresses)

Time Will Tell

Implémentation du côté client sur une instance large t3 d'AWS (Amazon Web Services -cloud-) qui envoie des requêtes de traduction en une langue target. Le côté hôte hébergent une LLM sur une g6.xlarge instance avec NVIDIA L4 Tensor Core GPUs. Permet de calculer la charge de travail/quantité de traitement (Plus de précision en 6.4).

On estime le temps d'aller retour avec un ping request et on y soustrait la latence mesurée afin d'obtenir le temps estimé d'exécution du LLM.

Remote Timing

The type of interception is not part of the scope of the paper, the only precision is that the packets with encrypted content are recovered and the payload is discarded.

Données et Modélisation

Données (avec Taille) d'entraînement :

Whisper Leak

positive samples:

100 variants of a question on a sensitive target topic ('legality of money laundering')

negative samples :

11,716 diverse questions from Quora Questions Pair dataset

Taille des données :

- up to 21,716 queries per model formed:

121,111 conversation captures

Input Snatch

- Attaque 1 : Prompt Engineering

Chatdoctor datasets dans lesquels ont été extraits des informations de 6 domaines distincts pour obtenir 16276 échantillons

- Attaque 2 : RAG

Base de données RAG (Retrieval Augmented Generation) construite en soumettant des legal corpus sur la plateforme légale d'OpenAI où ils ont mesurés et analysés des latences de récupération selon différentes queries.

Les modèles sont testés sur des datasets de 1000 échantillons qui mesurent le taux de succès de prédiction du nombre de blocs cache-hit et le taux de succès de prédiction correcte pour 4 domaines.

What Prompt

Models trained using UltraChat dataset*

(high-quality instructional conversations)

-> 1.5 million multi-turn dialogues using the GPT-4 Turbo API

*general inquiries section

Taille des données :

- 570,000 general inquiries

- average 12.57 sentences per response and 17.5 tokens per sentence (after segmentation)

- 10k : validation+testing

Time Will Tell

- Tâches de traduction :

Flores dataset

→ Traductions de la même phrase dans des langages différents. Permet d' évaluer la qualité de la traduction avec des résultats véridiques.

Test sur 50 échantillons et training sur 1000 inputs par langue target

- Tâches de classification :

Natural Instructions dataset

→ 61 tâches distinctes, instructions humaines, 193 000 inputs. 12

Tâches de classification binaire utilisées.

Remote Timing

Multiple use cases, legit user, black-box access, white-box access, no access at all (just to packets obtained from a MITM)

Disponibilité des données :

Whisper Leak

- Target questions are documented in the paper

- code available on github

Input Snatch

Github rep avec les fonctions de classification disponible. Modèle de LLMs disponibles dessus avec des fonctions de prédiction (pour le topic de crime)

Sinon dans l'article, pour attaque 1, plateforme médicale inspirée de la platerform médicale Zuoshou. Apprentissage du vocabulaire spécifique et des corrélations entre les domaines avec des datasets open-source de Chatdoctor. Extraction des échantillons de conversations contenant l'âge et utilisation de GPT-4o pour avoir un dataset formaté avec 16276 échantillons.

What Prompt

Training data borrowed from publicly available repositories

Time Will Tell

Datasets dans les références de l'article.

Pas de Git pour les classifiers etc.

Remote Timing

No references to dataset in the papers.

Nettoyage / Normalisation des Données :

Whisper Leak

In some cases, provider-side content filters caused certain noise prompt queries to fail.

-> unsuccessful queries are filtered out

Consistent lengths;

sequences of packet sizes and inter-arrival times were zero-padded to a fixed length

(corr. to the 95% percentile length observed for the target LLM)

Input Snatch

La standardisation de la taille des blocs nous permet de vérifier si on trouve le bon cache input en se basant sur la granularité du bloc.

- OpenAI utilise des blocs de 128 tokens
- DeepSeek utilise des blocs de 64 tokens.

Une fois le domaine d'input trouvé, on pourra construire le suivant selon le contexte (permet d'augmenter l'efficacité de l'attaque car champ de recherche réduit)

What Prompt

Data is partitioned into ordered segments that approximate complete sentences using specific heuristics :

1. Split sequences at token length = 1 (likely punctuation).

2. Merge short segments.

3. Approximate sentence boundaries.

Time Will Tell

- Mitigation avec une fenêtre de 5 min et prend la médiane entre 5 mesures de TTFT (Time To First Token) et de TPOT (Total Processing Time) pour régulariser les résultats pouvant être sur un serveur plus ou moins chargé
- LLM opère en mode non flux continu et s'occupe d'une seule requête à la fois

Remote Timing

No cleaning or normalization of data.

Modèles :

Whisper Leak

- gradient boosting (LightGBM)
- recurrent neural networks (Bi-LSTM)
- pre-trained transformer models (DistilBERT-uncased)

Input Snatch

Méthode Gradient Boosting : Combinaison de plusieurs arbres de décision en un seul modèle. Méthode de descente de gradient utilisé.

Random Forest: Méthode de bagging (Technique d' échantillonnage permettant de prendre plusieurs fois la même instance -tirage aléatoire avec remise-) étendue en combinant avec l'incertitude des caractéristiques pour créer une forêt d'arbres de décisions non corrélé (grâce à une génération d'un sous ensemble aléatoire de caractéristiques).

XGBoost (Extreme Gradient Boosting) : BIbliothèque de ML qui utilise des des arbres de décision et le Gradient Boosting, la seule diff est qu'il additionne les valeurs résiduels.

Modèle gaussien Naive Bayes pour des constructeurs d'Age et de Genre et permettre une prédiction du domaine par le constructeur.

What Prompt

Two T5-based Large Language Models :

- LLMA → generates the initial segment
- LLMB → generates all the following segments (using context)

Time Will Tell

- Langage : Modèles multilingues, tests sur M2M100, MBart5 et Tower (Variante de LLama-2)
- Modèle de mixture gaussien (GMM - Gaussian Mixture Model-) sur le dataset Flores-200

Remote Timing

- Gaussian Mixture Models (SpecDecode, SpecINfer, Medusa, Lookahead Decoding, CLLMs)

Features :

Whisper Leak

for each response stream :

application data record sizes (derived from TLS record lengths)

+

inter-arrival time between these records

Input Snatch

- Mécanismes de validation post-processing pour maintenir une intégrité sémantique et une consistance structurelle
-

What Prompt

Token-length sequence

-> from changes in packet payload sizes over time

Time Will Tell

Utilisation de :

- La taille des paquets
- Le temps entre les réponses (TTFT - Time To First Token- et TPOT -Total Processing Time-)

Remote Timing

The inter-packet response time

Prétraitement :

Whisper Leak

BERT :

numerical data (packet sizes and inter-arrival times) discretized into 50 bins each.

-> new vocab. : each bin is mapped to a unique token (ex. [TIME_5], [LEN_12])

LSTM :

padding sequences and then passing them through dedicated embedding layers

LightGBM :

each sequence pair (size, time) was flattened into a single feature vector

Input Snatch

Pas de prétraitement des données.

What Prompt

Filtering out control packets & protocol artifacts (like the "saw-tooth" pattern in QUIC).

Time Will Tell

Pas d'informations concernant le pré-traitement dans l'article.

Remote Timing

No pretreatment is realized.

Motifs (patterns) Exploités :

Whisper Leak

Autoregressive* generation produces structured timing and size patterns.

*each output is generated based on the previous outputs

- encryption of data results in ciphertext size being directly proportional to the original plaintext size

Input Snatch

- Différences de latence
- Délais de réponses observés

What Prompt

AI responses are marked by a degree of predictability in style and a tendency to reuse phrases.

- Warnings (consistent safety responses ex. “Sorry, I can’t help with that…”)

- Templates (beginning answers with standard phrasing ex. “Here are some tips…”)

- Unique Token Sequences (rare phrases with distinctive token lengths)

- Structure (recognizable formatting tokens ex. lists, bullet points)

Time Will Tell

Nombre d’output tokens qui ne varient pas selon le sujet donc facile à exploiter pour identifier le contenu des outputs.

Densité des tokens

Les deux combinés

Remote Timing

The pattern used gears towards the building of a classifier (either a simpler one if a prompt is faster than the other to be tested or a more complex GMM / neural network) using the inter-token response time.

Évaluation et Performance

Métriques :

Whisper Leak

Area Under the Precision-Recall Curve (AUPRC)

-> used to evaluate binary classifiers with highly imbalanced datasets

Precision even at high noise levels :

10,000:1 noise-to-target ratio

Input Snatch

- Taux de succès de prédiction du nombre de blocs cache-hit (SR Block)
- Attack Success Rate sur les maladies, les symptomes etc. (ASR)

What Prompt

Primary success metric:

Cosine similarity for topic alignment

(semantic similarity between original and reconstructed response)

-> from sentence embeddings

textual accuracy :

- Edit distance at the character level (for lexical similarity)

- ROUGE score at the word level (exact reconstruction accuracy)

Time Will Tell

- Accuracy du classifieur qui détermine la langue
- Attack Success Rate (ASR) sur chacun des prompts entre les différents LLMs testés
- Coefficient de corrélation de Pearson

Remote Timing

- Attack Success Rate (ASR)
- Accuracy

Baselines :

Whisper Leak

Performance compared across :

- 3 attack architectures (BERT, LSTM, LightGBM)
- 3 feature configurations (Both, Size Only, Time Only)

Input Snatch

Attaque 1

Comparasion entre Gradient Boosting, Random Forest et XGBoost

Attaque 2

Comparaison de performance entre différentes méthodes

- Baseline
- GaussianNB
- Prob_vocabulary
- Finetuned LLM

What Prompt

- Markov/HMM models. *did not scale well

- GPT-4 baseline reconstruction. *underperformed significantly

- Generic text-trained model (C4 dataset). (trained on general data instead of target outputs) *better than GPT-4

- Victim-trained; target assistant responses. *large accuracy gains

-> upper bound

Time Will Tell

Comparaison entre les différents modèles et avec différentes langues

- Modèles (Tower, M2M100, MBart50, Gemma2, GPT-4o)
- Langues (Anglais, français, espagnol, russe, coréen, chinois etc.)

Remote Timing

Comparison between different LLMs (Claude,GPT) and different languages

Résultats :

Whisper Leak

Highly successful, most models reached >98% AUPRC (despite high noise)

-> most models achieved near perfect performance using packet sizes alone

17/28 models enable 100% precision at 5-20% recall

-> minimal false positives

Input Snatch

Reconstruction jusqu'à 90% des caractères de l'input de l'utilisateur

What Prompt

Accurately reconstructed 29% of an AI assistant's responses and successfully inferred the topic from 55% of them.

-> Attack success rate of over 54.5% on the first segment and 37.96% for entire responses.

Time Will Tell

Utilisation du coefficient de corrélation de Pearson pour déterminer si le temps de génération est corrélé de façon linéaire avec le nombre de tokens générés selon les LLMs.

⇒ Oui corrélation très forte (0.987 de coeff) sur OpenAI mais pas sur GPT-4o (0.370)

⇒ Probablement dû à un décodage spéculatif + des améliorations générées par la LLM ce qui affecte le nb de tokens et le temps de génération

Remote Timing

Language prediction was fairly accurate with GPT-4.

Attack Success Rate increases as models increase in size.

Robustesse :

Whisper Leak

Attack performance improves substantially with more training data

No clear trend in attack effectiveness vs temperature is observed

-> remains effective across different temps.

Input Snatch

- Interférence liée au bruit qui peut perturber la mesure des timings
- Limites de taux imposées par les API et mémoire GPU limitée également (qui peuvent éjecter la target request de la mémoire)
- Durée de l"attaque par canal auxiliaire limitée par le TTL du cache

What Prompt

Attack remains effective even under imperfect network conditions ex.

- lost tokens, grouped / paired tokens

- noisy, compressed, or incomplete token-stream

Time Will Tell

Expériences avec un batch size de 1 mais dans des systèmes réels, on veut une latence faible et un haut débit.

On veut que même en présence de bruit, la relation avec le nombre d'output tokens reste linéaire..

Remote Timing

The behavior of the attack is inconsistent in time on production models if a single training in the attack is done and then repeated at later dates as the behavior and implementation of every LLM changes, including in the same releases.

Transférabilité :

Whisper Leak

Yes, the results reveal an industry-wide vulnerability (inferred from the effect on 28 models)

-> systemic issue rooted in fundamental architecture choices

Input Snatch

Tests sur plusieurs LLMs mais la robustesse de l'attaque baisse lorsqu'on optimise l'inférence

What Prompt

Yes, a model trained on one assistant's data can successfully attack a different assistant.

-> uniformity of LLM writing styles

Time Will Tell

Les résultats révèlent une similitude dans les ASRs (Attack Success Rate) obtenus entre les différents modèles testés.

Remote Timing

The attack is transferable but needs a classifier training more or less specific to each model.

Défenses et Limites

Défenses :

Whisper Leak

Mitigation strategies:

- Random padding (extra unpredictable, non-functional data added to packets)

- Token batching (collecting multiple generated tokens and sending them together)

- Packet injection (injecting synthetic packets at random intervals)

goal : obfuscate both size and timing patterns

Input Snatch

Implémentation d'un cache distinct afin de palier au partage de cache entre utilisateurs. Cette mesure permet d'augmenter la sécurité mais cela fait baisser l'efficacité. Des LLMs comme OpenAI et DeepSeek ont implémenté des isolations pour cachetage du préfixe afin de protéger la vie privée et assurer la sécurité de l'utilisateur

Limitation du débit afin de pallier aux attaques fréquentes. En limitant la fréquence des demandes, la limitation de débit permet d'empêcher l'investigation rapide et successive de l'analyse du timing. Cela nécessite cependant de bien balancer la sécurité avec un accès légitime de l'utilisateur

Complication de l'analyse de temps. Obfuscation du timing permet de contrer efficacement l'exploitation des variations de temps de réponse.

Deux stratégies d'obfuscation proposées :

Homogénéisation du temps de réponse à travers une exécution en temps constant

Injection d'un temps de battement aléatoire etdésactivation des réponses en flux continu afin d'éliminer la mesure des timing patterns.

What Prompt

- Random padding (obscuring the actual length of tokens)

- Transmitting tokens in larger groups rather than individually

- Batching Responses (sending complete responses at once)

Time Will Tell

Fixer une limite au prompt : Dire au LLM "Réponds en 60 mots" par exemple permet de diminuer la précision sur certains modèles mais pas d'autres. (Section 7.2)

Avoir un nb de tokens uniforme permettrait de troubler la prédiction

Padding pour obtenir une longueur d'output en octets similaire entre toutes les langues

Remote Timing

Induce some randomness in the packet timing and sizes or a constant output of tokens is the most realistic approach as it allows a counter-measure without a significant impact in the service

Use of different internet protocols to deliver the content.

Efficacité :

Whisper Leak

All three defense approaches provide meaningful reductions in attack effectiveness (3.5%-4.8%), though residual vulnerabilities remain.

-> underlying information leak remains present

Input Snatch

- Equilibrer la sécurité, la performance et l'expérience utilisateur permet la protection contre des vulnérabilités side-channel

What Prompt

- Padding and the batching of responses would increase bandwidth usage.

- Grouping tokens or sending responses in batches could detract from the user experience.

Time Will Tell

Fixer une limite au prompt : Permet de diminuer la précision sur certains modèles mais pas d'autres. (Section 7.2)

Avoir un nb de tokens uniforme permettrait de troubler la prédiction -> Peut avoir un impact sur la performance du modèle (Section 7.1)

Padding à rajouter très variable selon les différentes langues, pas forcément optimal pour implémentation

Remote Timing

The output of tokens at a constant rate by the model is arguably the best solution, having only the downside that adds a non-negligeable overhead in the bandwith

Limites :

Whisper Leak

Mitigations involve trade-offs between

- security guarantees,

- bandwidth overhead (increase in amount of data transmitted over the network)

- latency impact (increase in response delay)

ex. packet injection incurs bandwidth overhead (2-3x traffic volume) but maintains streaming performance.

Input Snatch

- Déploiement du cloud inévitable pour des raisons computationnelles
- Capture des informations faisable mais ne fournit pas le contexte détaillé requis pour reconstruire correctement l'input
- Pas de liens possibles entre des utilisateurs spécifiques et l'input

What Prompt

Performance drops as the response gets longer.

(error propagation from inferred context, linguistic ambiguity, etc…)

-> attack success rate for entire responses was 37.96%

Model sometimes "cheats" by altering tokens to stay on topic rather than achieving a perfect word-for-word match.

-> always trying to find most probable sentence

Time Will Tell

Exploitation de l'allocation mémoire, des pointeurs d'instruction et des mesures de tps dans la GPU pourrait leak les infos side-channel dans le futur

Est ce que l'optimisation de la latence peut impacter les observations ?

Remote Timing

Tunneling the traffic through other protocols or applications are difficult to implement and have heavy additional costs.