Article	**Whisper Leak**	**Input Snatch**	**Remote Timing**	**Time Will Tell**	**What Prompt**
Whisper Leak		Similarités : Différences :	Similarités : Différences :	Similarités : - But d'identification sur les paquets - Utilisation de la taille des oaqyets - Utilisation du timing entre les paquets - Accuracy-based - Bases de données initiales réelles (Réponses générées par LLMs pour les deux) - Méthode de mitigation proposée de padding Différences : - Modèles utilisés	Similarités : Cf. Partie en dessous Différences : Cf. Partie en dessous
Input Snatch	Similarités :		Similarités :	Similarités :	Similarités :

	Différences :		Différences :	- Analyse du temps - Utilisation du KV Cache - RAG (Retrieval Augmented Generation) : Amélioration de la performance du modèle en choisissant des exemples dans le contexte - Utilisation du TTFT Différences : - Construction dans IS d'input alors que dans TWT on les observe seulement	Différences :
Remote Timing	Similarités : Différences :	Similarités : Différences :		Similarités : - Basé sur l'inspection des paquets - Gaussian	Similarités : Différences :

				Mixture Models pour inférence/évaluation - End-to-end attaque sur le site de ChatGPT dans RT et temporalisé dans TWT Différences : - Type d'attaques utilisées parfois différentes (celles non citées plus haut)	
Time Will Tell	Similarités : - But d'identification sur les paquets - Utilisation de la taille des paquets - Utilisation du timing entre les paquets - Accuracy-based	Similarités : - Analyse du temps - Utilisation du KV Cache - RAG (Retrieval Augmented Generation) : Amélioration de la performanc	Similarités : - Basé sur l'inspection des paquets - Gaussian Mixture Models pour inférence/évaluation - End-to-end attaque sur le site de ChatGPT dans RT et		Similarités : - Utilisation de la longueur des tokens - Transformer-based models - Datasets utilisés publics, pas de queries générées pour les

	- Bases de données initiales réelles (Réponses générées par LLMs pour les deux) - Méthode de mitigation proposée de padding Différences : - Modèles utilisés	e du modèle en choisissant des exemples dans le contexte - Utilisation du TTFT Différences : - Constructio n dans IS d'input alors que dans TWT on les observe seulement	temporalisé dans TWT Différences : - Type d'attaques utilisées parfois différentes (celles non citées plus haut)		prompts Différences : - Test en anglais uniquement pour WP alors que TWT teste sur plusieurs langues
What Prompt	Similarités : Cf. Partie en dessous Différences : Cf. Partie en dessous	Similarités : Différences :	Similarités : Différences :	Similarités : - Utilisation de la longueur des tokens - Transformer -based models - Datasets utilisés publics, pas de queries générées pour les prompts Différences : - Test en anglais	

				uniquement pour WP alors que TWT teste sur plusieurs langues	

What Prompt vs Whisper Leak :

Similarités :

Accès : Utilisation légitime via des requêtes normales, capture passive du trafic
Interception du trafic : interception de trafic chiffré (TLS, QUIC)
Données d'entraînement : Réponses collectées ou générées utilisées pour entraîner les modèles
Robustesse : Les attaques restent efficaces malgré du bruit ou une perte partielle de données
Transférabilité : Fonctionne sur différents modèles ou plateformes (ex. GPT, Copilot)
Contre-mesures : Batching, padding, obfuscation proposés comme défenses
Efficacité des défenses : Réduisent la fuite d'information mais ne l'éliminent pas
Limites : Dépendent du style prévisible et de la stabilité des modèles LLM

Différences :

	Whisper Leak	**What Prompt**
Problème	Détecter si un sujet sensible (ex. blanchiment) est abordé à partir du trafic chiffré vocal	Reconstituer les réponses textuelles complètes à partir du trafic chiffré
Hypothèse	Les tailles de paquets et temps entre paquets révèlent le contenu	Les longueurs de tokens inférées révèlent le texte de la réponse
Objectif	Classer la présence d'un sujet cible (ex. sécurité,	Reconstituer le texte exact des réponses d'IA

	crime)	
Type d'attaque	Canal auxiliaire via analyse de taille et timing TLS	Canal auxiliaire via inférence des longueurs de tokens
Contributions clés	Évaluation de 28 modèles, classification binaire, test de 3 stratégies de défense	Attaque par LLM génératif, inférence multi-phrase, exploitation du style des LLMs
Pipeline	Capture TLS → encodage taille/timing → classification	Message sizes → longueurs de tokens → segmentation → inférence avec deux LLMs
Canal auxiliaire	Taille des paquets + temps entre paquets	Différences de longueur des paquets pour déduire la longueur des tokens
Dialogue	Questions uniques + questions issues de Quora (one-shot)	Dialogue multi-turn, séquentiel avec dépendance contextuelle
Cibles	ChatGPT app, Edge, assistants vocaux Android/iOS	ChatGPT-4, Copilot (navigateur et API)
Modèles	LightGBM, Bi-LSTM, classifieur BERT	T5 encoder–decoder, fine-tuné pour prédire à partir de séquences de longueurs
Features	Séquences de tailles de paquets, intervalles temporels	Séquences de longueurs de tokens
Prétraitement	Encodage par bins, padding, vecteur d'entrée	Segmentation heuristique des phrases à partir des longueurs de tokens
Métriques	Accuracy, taux de faux positifs	Cosine similarity, ROUGE, distance d'édition
Baselines	Comparaison entre types de modèles (pas de baseline naïf)	Markov, HMM, prompting direct avec GPT-4
Résultats	99.9% de précision sur 17/28 modèles, fonctionne avec seulement 5–20% des données	29% de correspondance exacte, 55% d'inférence thématique, très bons débuts de réponse