AYI NEDJIMI Consultants

LLM Model Extraction 2026 : Voler un GPT

16 mai 2026 • Mis à jour le 17 mai 2026 • 20 min de lecture • 3757 mots

Carlini et al. (2024) extraient la matrice de projection finale de GPT-3.5 pou
l'extraction par distillation atteint 92% de fidelity pour 50000 USD.

À RETENIR

A retenir — LLM Model Extraction

Carlini et al. (2024) extraient la matrice de projection embeddings → logits
200 USD en queries API.

Black-box distillation Llama 4 vers proxy 7B : 91% accuracy retention sur M
50000 USD, 200M tokens.

Defenses 2026 : *logit truncation* (top-k visible), *differential privacy* sur outp
Kirchenbauer.

Impact business : vol de proprite intellectuell
modeles concurrents, contournement de lice

> AI Act article 53 : les fournisseurs GPAI doivent documenter les contre-mes
> Litiges en cours US (NYT vs OpenAI, Getty vs Stability).

Le **model extraction llm**, ou *model stealing*, est l'attaque qui menace directement des laboratoires d'IA. L'idee : un attaquant qui ne connait pas les poids d'un LLM p a partir d'interrogations API. Trois familles existent : (1) **fonctionnal extraction** via (2) **parametre extraction** via logit-based attacks (Carlini et al., 2024), (3) **capabilit** capacites specifiques (fine-tuning specialise). En 2026, le cout d'extraction d'un L (50k-500k USD selon la fidelity attendue), mais le ROI attaquant peut etre enorme gouvernement adversaire, ou un acteur cherchant a republier le modele sans licen fournisseurs de LLM proprietaires, la documentation des contre-mesures (AI Act a exigence reglementaire incontournable, doublee d'une realite competitive ou Byte plusieurs startups asiatiques ont prouve la viabilite de l'extraction industrielle. Cet **model extraction llm** sur les frontiers 2026.

1. Genese et etat de l'art

L'attaque de model extraction remonte a Tramer et al. (2016) *Stealing Machine Le Prediction APIs*, applique a l'epoque aux SVM et logistic regression. Pour les NN, F demontrent la possibilite de cloner via queries crafted. Pour les LLM, les premiers sont :

> **Wallace et al. (2020)** : *Imitation Attacks and Defenses for Black-box Machine T* retention sur Google Translate avec ~100k queries.

> **Krishna et al. (2020)** : extraction de BERT-base fi s random, ~

Réponse sous 24h

Devis gratuit →