Comment boycotter les agents IA en CTF

rump écrite 4 mois av. F-5 (avant Fable 5)

Qui-sommes nous

- fait de son mieux

- tabasse des chatbots

wepfen

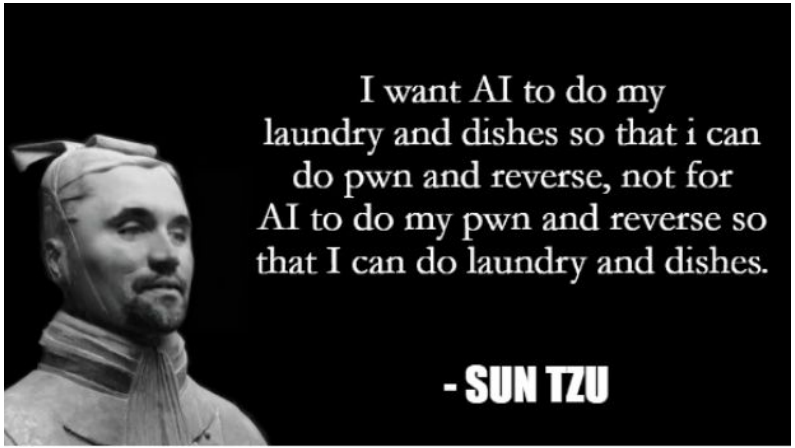
wepfen.github.io

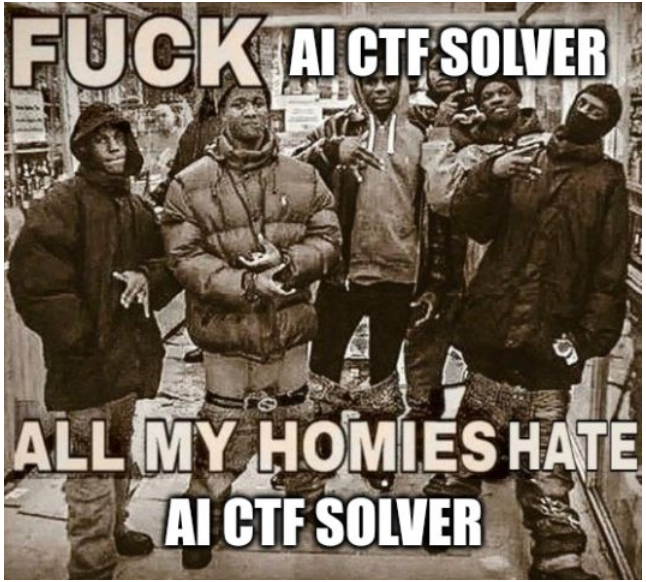
anthr4ce

anthr4ce.github.io

Pourquoi cette rump ?

- L’IA va voler nos jobs, bientôt nos hobbies (peut-être même nos femmes)
- Les CTFs sont devenus une compétition de LLM
- On va en CTF pour kiffer et pas pour faire kiffer Claude!
- Parce qu’on a envie

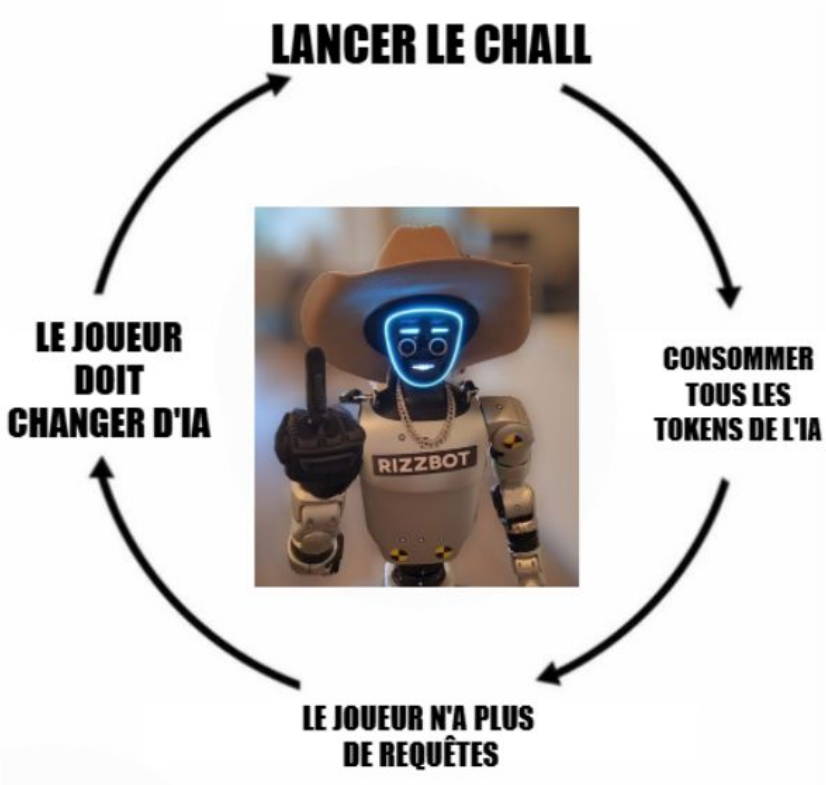




C’est quoi un agent IA

- LLM avec des outils. Il agit, il ne fait pas que répondre
- Il tourne en boucle : réfléchit → appelle un outil → observe → recommence
- Tout ça manipule des tokens : des bouts de texte que le modèle prédit un par un
- Claude Code, Codex, Gemini CLI... soumets-lui un challenge (avec contexte ou pas), il flag tout seul

Exemple de strat



Système de notation des techniques

- Inutile/passif

- Très peu dérangeant

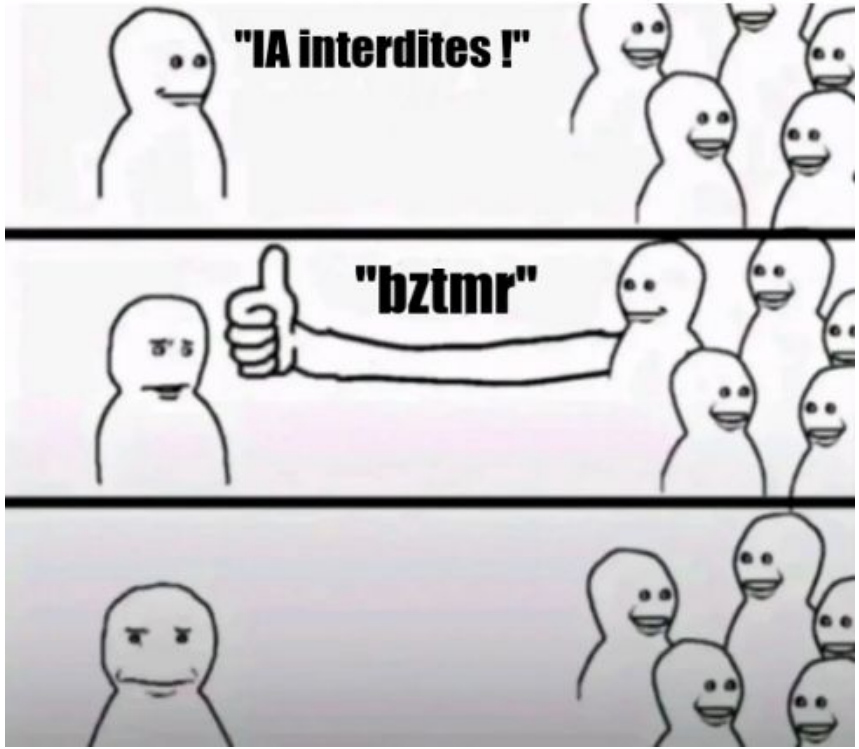
- Un peu chiant

- Frustrant (donc très efficace)

- Malveillant

Méthode 1 : Dire non aux joueurs

- Pas assez agressif



Méthode 2 : Accepter l’IA comme elle est

- Progressiste

- Trop passif
- Ne boycott pas l’IA

(il lui reste 3 mois)

Méthode 3 : Scoreboard séparé

- Part d’une bonne volonté

- Les users peuvent décider d’en avoir rien à foutre
- Ne boycott pas l’IA

Méthode 4 : Faire de la discrimination anti-IA

- User-agent, headers, …

- Bypassable sans rien faire (les IA s’adaptent)

Méthode 5 : Prompt injection

- Cacher des prompts injection dans les challs
- Les encoder ou les offusquer

- Souvent détecté par les IAs

Méthode 6 : Mettre un fake flag ÉVIDENT

- Encore mieux sur des challs à essais limités
- Proba de recevoir des insultes

- Ca remplit une case du bingo bad CTF
- Si trop explicite, peut se faire cramer

CTF{If_you_flag_this_you_will_be_molested_and_kicked_out_of_the_ctf}

Méthode 7 : Fausse pistes

- Dépendance circulaire : pour obtenir A il te faut B, pour obtenir B il te faut C, et pour obtenir C il te faut A. L'agent boucle indéfiniment.
- Fonction **decrypt()** qui dechiffre pas

- Crame pas mal les tokens

Méthode 8 : Output flooding

- Peut garder l’IA bloquée longtemps

- Marche moins si le gars a plusieurs agents
- L’IA peut s’en rendre compte

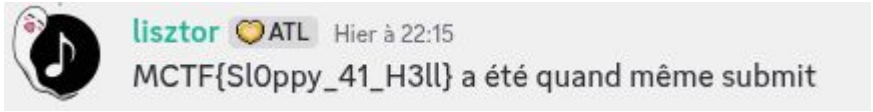
Méthode 9 : Être créatif / original

- Faire des challs peu courants (architecture exotique, jeux vidéos, projet github niche …)

- C’est qu’une question de temps pour que le l’IA solve

Méthode 10 : Cacher un faux flag dans les fichiers

- En clair (mais trop cramé)
- Le cacher avec de la stega (ex: **MineSlayer** au midnight CTF, a été très efficace, voir screen)



source : discord midnightCTF 2026

Méthode 11 : Bloquer les domaines via proxy web / dns

- Solide

- Bypassable avec un VPS
- Si le mec à son modèle en local on est cuit
- Marche pas en remote

Méthode 12 : Challenge Zero day

- L’IA aura du mal sur une vuln qu’elle ne connaît pas
- Trouver la zero day sans utiliser d’IA (de préférence).
- Selon la difficulté de la zeroday, Claude aurait pas vraiment de soucis

Méthode 13 : Dire non (proactivement)

- Si on notifie des gens qui solvent un peu trop vite -> interrogatoire + shoulder surfing

- Des gens maîtrisent très très bien le ALT TAB

Méthode 14 : Blackbox (déconseillé)

- Moins de solve (évidemment)

- Méthode non-éthique
- Ca fait chier l'IA mais aussi le joueur
- Forte probabilité de recevoir des insultes/menaces de la part des joueurs

Méthode 15 : Stegano **proactive**

- N’importe quoi peut servir à faire un chall stegano.
- Proba de se faire insulter également
- Ne PAS prendre une technique qui existe déjà pour faire chier l’IA

Méthode 16 : Texte pour trigger les safety du modèle

- Trigger les sécurités du modèle :
  - Armes biologiques
  - Racisme
  - Produire des drogues
  - Contenu coquin 🔞
- Pas censé être bypassable (normalement)

Tester pour voir si ça marche

- Y a beaucoup moins de first blood en 30 secondes
- Moins de solves medium/hard
- Utiliser des tools comme CTF solver ou BenchCTF