Écrit par une IA ?

Détecter un texte généré par une intelligence artificielle

Mise en situation. Un professeur corrige des devoirs à la maison. Certaines copies lui semblent trop lisses, trop bien construites : ont-elles été écrites par une IA comme ChatGPT, Gemini ou Le Chat ? Sur Internet, de nombreux sites promettent de le dire « à coup sûr ».
Dans cette activité, vous allez calculer à quel point ces détecteurs se trompent, puis découvrir une autre méthode, le **tatouage** des textes, et la programmer en Python.

1. Les détecteurs « à l'aveugle »

Définition

Un **détecteur de texte IA** analyse un texte et estime s'il a été écrit par une IA. La plupart s'appuient sur la **prévisibilité** du texte : une IA choisit souvent des mots très probables, donc un texte dont chaque mot est facile à deviner est jugé suspect.
Un détecteur peut se tromper de deux façons :

- il **rate** un texte écrit par une IA ;
- il **accuse à tort** un texte écrit par un humain : on parle de **faux positif**.

Remarque

En janvier 2023, l'entreprise OpenAI (qui a créé ChatGPT) a publié son propre détecteur. D'après ses tests, il reconnaissait **26 %** des textes écrits par une IA, et il accusait à tort **9 %** des textes écrits par des humains. OpenAI l'a retiré le 20 juillet 2023, en raison de sa faible précision.

▶ À vous de jouer !

On fait analyser par ce détecteur 1 000 textes : 500 écrits par une IA et 500 écrits par des humains.

1. Combien de textes écrits par une IA sont reconnus ? Combien sont ratés ?
2. Combien de textes écrits par des humains sont accusés à tort ?
3. Recopier et compléter le tableau :

Détecteur	Écrit par une IA	Écrit par un humain	Total
Déclaré « IA »			
Déclaré « humain »			
Total	500	500	1 000

4. Parmi les textes déclarés « IA », quelle proportion a en réalité été écrite par un humain ? Arrondir au centième.
5. Un élève rend un devoir que ce détecteur déclare « IA ». Peut-on le sanctionner sur cette seule base ?

Coup de pouce

26 % de 500 se calcule ainsi : $0{,}26 \times 500$. Pour la question 4, on divise le nombre de textes humains déclarés « IA » par le nombre total de textes déclarés « IA ».

Remarque

En 2023, des chercheurs de l'université Stanford ont testé sept détecteurs sur des rédactions d'anglais écrites par des élèves dont ce n'est pas la langue maternelle : 61 % de ces textes, tous écrits par des humains, ont été déclarés « IA ». Avec un vocabulaire plus simple, leurs textes sont plus prévisibles… comme ceux d'une IA. Les détecteurs sont donc injustes envers certains élèves.

2. Le tatouage des textes

Définition

Tatouer (ou « filigraner ») un texte, c'est cacher dans le texte une marque invisible pour le lecteur, au moment où l'IA l'écrit. Une méthode, proposée en 2023, utilise une règle qui partage les paires de mots consécutifs en deux couleurs, **vertes** et **rouges** :

- dans un texte écrit par un humain, environ la **moitié** des paires sont vertes, par hasard ;
- une IA tatouée choisit, quand elle hésite entre plusieurs mots, celui qui forme une paire verte avec le mot précédent : son texte contient **beaucoup plus** de paires vertes.

Pour savoir si un texte est tatoué, on compte ses paires vertes.

Dans la réalité, la règle est très compliquée et reste **secrète** : seule l'entreprise qui a créé l'IA peut l'appliquer. Dans cette activité, on utilise une règle simplifiée :

Exemple

Règle de l'activité. On ne tient pas compte de la ponctuation, et on compte les lettres de chaque mot (une lettre accentuée compte pour une lettre). La paire formée par deux mots consécutifs est **verte** si le nombre total de lettres des deux mots est **pair**, et **rouge** sinon.

- « un rôle » : $2 + 4 = 6$, pair : la paire est verte.
- « vieux film » : $5 + 4 = 9$, impair : la paire est rouge.

Un texte de 21 mots contient 20 paires de mots consécutifs.

▶ À vous de jouer !

On considère deux textes de 21 mots. L'un a été écrit par un humain, l'autre par une IA tatouée avec la règle de l'activité.
Texte A : « Hier soir, nous avons regardé un vieux film avec mes cousins. La fin était triste, mais la musique restait très belle. »
Texte B : « Selon les experts, les forêts jouent un rôle majeur pour le climat, abritent de nombreux insectes et fixent le sol durablement. »

1. Compléter le tableau du texte A : la longueur de chaque mot, puis la couleur (V ou R) de la paire formée par ce mot et le mot **précédent**.

Mot	Hier	soir	nous	avons	regardé	un	vieux	film	avec	mes	cousins
Longueur											
Paire											

Mot	La	fin	était	triste	mais	la	musique	restait	très	belle
Longueur										
Paire										

2. Compléter de même la ligne « Paire » du tableau du texte B.

Mot	Selon	les	experts	les	forêts	jouent	un	rôle	majeur	pour	le
Longueur	5	3	7	3	6	6	2	4	6	4	2
Paire											

Mot	climat	abritent	de	nombreux	insectes	et	fixent	le	sol	durablement
Longueur	6	8	2	8	8	2	6	2	3	11
Paire										

3. Pour chaque texte, calculer le nombre de paires vertes, puis la proportion de paires vertes parmi les 20 paires.
4. Lequel des deux textes semble tatoué ?

Coup de pouce

La somme de deux nombres est paire quand ils sont tous les deux pairs ou tous les deux impairs. Une paire est donc verte quand les deux mots ont des longueurs de même parité : 4 et 6, ou 5 et 7, par exemple.

3. Décider avec un intervalle

Un texte humain peut avoir, par hasard, un peu plus ou un peu moins de la moitié de paires vertes. À partir de quelle proportion faut-il le déclarer tatoué ?

Définition

On admet que, pour un texte écrit par un humain qui contient n paires de mots (avec $n \geqslant 25$ environ), la proportion de paires vertes se trouve dans l'intervalle

$$\left[0{,}5 - \frac{1}{\sqrt{n}} \ ; \ 0{,}5 + \frac{1}{\sqrt{n}}\right]$$

dans environ 95 % des cas. C'est un **intervalle de fluctuation**.
Règle de décision : si la proportion de paires vertes d'un texte est **supérieure** à $0{,}5 + \frac{1}{\sqrt{n}}$, on déclare le texte tatoué. Sinon, on ne peut pas conclure.

▶ À vous de jouer !

1. Les textes A et B comptent $n = 20$ paires. Calculer l'intervalle correspondant, en arrondissant les bornes au millième. (On l'utilise ici même si n est un peu petit.)
2. Que peut-on conclure pour le texte A ? Pour le texte B ?
3. Un texte humain peut-il être déclaré tatoué par erreur ? Est-ce fréquent ?

4. Programmer le détecteur en Python

Exemple

Les textes sont écrits sans ponctuation. La méthode `split()` découpe un texte en une liste de mots, et `len(mot)` donne le nombre de lettres d'un mot.

```
from math import sqrt

def est_verte(m1, m2):
    return (len(m1) + len(m2)) % 2 == 0

def proportion_verte(texte):
    mots = texte.split()
    n = len(mots) - 1
    verts = 0
    for i in range(n):
        if est_verte(mots[i], mots[i + 1]):
            verts = verts + 1
    return verts / n
```

Rappel : `a % 2` est le reste de la division de `a` par 2. Il vaut 0 quand `a` est pair.

▶ À vous de jouer !

1. Que renvoie `est_verte("un", "rôle")` ? Et `est_verte("vieux", "film")` ?
2. Dans la fonction `proportion_verte`, pourquoi `n` vaut-il `len(mots) - 1` ?
3. On exécute `proportion_verte("la fin était triste")`. Recopier et compléter le tableau d'exécution de la boucle, puis donner la valeur renvoyée.

`i`	`mots[i]`	`mots[i + 1]`	Somme des longueurs	Paire verte ?	`verts`
0					
1					
2					

4. Écrire une fonction `est_tatoue(texte)` qui renvoie `True` si la proportion de paires vertes du texte est supérieure à $0{,}5 + \dfrac{1}{\sqrt{n}}$, et `False` sinon.

5. Les limites du tatouage

Remarque

- En octobre 2024, des chercheurs de Google DeepMind ont présenté dans la revue *Nature* SynthID-Text, une méthode de tatouage de ce type. Elle a été testée sur près de 20 millions de réponses de l'IA Gemini, sans que les utilisateurs remarquent de différence de qualité.
- Depuis le 2 août 2026, le règlement européen sur l'IA (AI Act) impose aux fournisseurs d'IA générative de marquer les contenus qu'elles produisent, textes compris, d'une façon lisible par une machine. Les IA déjà commercialisées avant cette date ont jusqu'au 2 décembre 2026 pour s'y conformer.
- Mais chaque entreprise a sa propre règle secrète : un texte tatoué par une IA ne peut être reconnu que par l'outil de la même entreprise.

▶ À vous de jouer !

Un élève a obtenu le texte B avec une IA tatouée. Pour échapper au détecteur, il remplace quelques mots par des synonymes.

1. Il remplace « majeur » par « important ». Quelles paires changent de couleur ? Quelle est la nouvelle proportion de paires vertes ? Le texte est-il encore déclaré tatoué ?
2. Il remplace ensuite « nombreux » par « multiples ». Mêmes questions.
3. Pourquoi un texte très court (un tweet, une réponse d'une ligne) ne peut-il pas être reconnu comme tatoué ?
4. Le tatouage permet-il de prouver qu'un texte a été écrit par un humain ?

Coup de pouce

Quand on remplace un mot, seules deux paires peuvent changer : celle qu'il forme avec le mot précédent et celle qu'il forme avec le mot suivant.

6. Pour aller plus loin

★ Question étoilée

Pour que ses textes restent naturels, une IA est tatouée « légèrement » : ses textes contiennent environ 60 % de paires vertes.

1. Pour un texte de $n = 25$ paires, calculer l'intervalle de fluctuation. Le texte est-il déclaré tatoué ?
2. Même question pour $n = 400$ paires.
3. À partir de combien de paires un texte qui contient 60 % de paires vertes est-il déclaré tatoué ?

À retenir

- Les détecteurs « à l'aveugle » se trompent souvent : ils ratent des textes d'IA et accusent à tort des humains (**faux positifs**), en particulier ceux qui écrivent simplement. Leur verdict n'est jamais une preuve.
- Le **tatouage** cache dans un texte d'IA une marque invisible : l'IA favorise les paires de mots « vertes » d'une règle secrète.
- Pour décider, on compare la proportion de paires vertes à un **intervalle de fluctuation** : au-dessus de $0{,}5 + \dfrac{1}{\sqrt{n}}$, le texte est déclaré tatoué.
- Le tatouage a des limites : il faut un texte assez long, il disparaît quand on reformule, et chaque entreprise a sa propre règle. Depuis 2026, l'AI Act oblige les fournisseurs d'IA à marquer leurs contenus.