

Claude marque invisiblement tes textes : comment ça marche et comment s'en sortir

Le filigrane invisible de Claude expliqué — et les solutions pour t'en débarrasser

CE QUE TU VAS APPRENDRE

- » Comment Claude encode un filigrane directement dans les choix de mots
- » Ce que fait concrètement watermarks-remover et comment l'installer
- » Comment réécrire un texte pour qu'il sonne humain sans outil externe
- » Le prompt prêt à copier-coller pour nettoyer ton style IA

Ce qui a changé le 2 août 2026

Depuis le 2 août 2026, Claude glisse une marque invisible dans chaque texte qu'il génère. Pas un caractère spécial, pas un pixel caché. La marque est enfouie dans les choix de mots eux-mêmes.

C'est une obligation légale. L'AI Act européen (article 50) impose que les contenus d'IA soient identifiables. Google, Meta, Microsoft, Mistral et OpenAI ont signé le même engagement et déploient leurs propres systèmes.

ATTENTION

La marque apparaît même si Claude a seulement corrigé tes fautes ou traduit ton texte. Pas uniquement quand tu lui demandes d'écrire de zéro.

Comment ça fonctionne vraiment

Claude génère un texte token par token — un token, c'est un mot, un bout de mot, ou de la ponctuation. À chaque étape, plusieurs suites sont possibles et statistiquement équivalentes.

Exemple : "Le restaurant était étonnamment..." — bien, calme, abordable, élégant sont tous valides.

Normalement, le modèle choisit selon des probabilités avec une part de hasard. Le filigrane, lui, utilise une clé secrète pour légèrement favoriser certains tokens sur d'autres dans ce contexte précis. Peut-être calme plutôt qu'élégant.

Mais si la phrase change — "En entrant, j'ai immédiatement remarqué que le restaurant était étonnamment..." — le contexte précédent est différent, donc le système va peut-être favoriser spacieux cette fois.

C'est le point clé : il n'y a pas de liste secrète de mots IA. Les préférences changent en permanence selon ce qui précède.

Répète ça sur des centaines de tokens. Un choix isolé ne veut rien dire — un humain peut très bien écrire le même mot. Mais sur un texte suffisamment long, les micro-préférences forment un pattern statistique. C'est ça, le filigrane.

Un détecteur qui connaît la clé secrète d'Anthropic peut ensuite analyser le texte et demander : est-ce que ce texte a choisi l'option marquée beaucoup plus souvent que le hasard le prédirait ? Si oui, Claude est probablement impliqué.

ASTUCE

Copier-coller le texte ne retire pas la marque — elle voyage avec les mots. En revanche, réécrire en profondeur, paraphraser ou mélanger avec beaucoup d'autre contenu affaiblit le signal. Les textes très courts sont aussi plus difficiles à détecter faute de tokens suffisants.

Anthropic prépare par ailleurs une API de détection qui permettra à n'importe quelle école ou employeur de vérifier automatiquement. La méthode s'inspire de SynthID, la technologie de Google.

Les solutions concrètes

Option 1 : un outil open source

watermarks-remover (<https://github.com/guillaumemeyer/watermarks-remover>) est un outil gratuit (licence MIT) créé par Guillaume Meyer qui tente de nettoyer les marques dans un texte ou dans des fichiers (images, PDF, Word). Le repo contient aussi un fichier `skill.md` — c'est ce qui permet de l'installer directement comme compétence dans Claude, sans passer par une interface tierce.

Comment l'installer via Claude Skills :

1. Ouvre le repo GitHub et copie le contenu brut du fichier `skill.md`.
2. Dans Claude, clique sur Personnaliser dans la colonne de gauche, puis sur Compétences.
3. Clique sur Nouvelle compétence, colle le contenu du fichier, et enregistre sous le nom `watermarks-remover`.
4. Ensuite, tape simplement `/watermarks-remover` suivi de ton texte pour lancer le nettoyage directement depuis l'interface.

ATTENTION

L'API de détection d'Anthropic n'est pas encore publique. Impossible de confirmer à 100% que cet outil fonctionne sur le filigrane de Claude.

Option 2 : réécrire avec ta voix

C'est la méthode la plus fiable. Le principe : supprimer les tics d'IA et ajouter des marqueurs humains.

Les tics à supprimer :

- La formule "ce n'est pas X, c'est Y" — garde juste la conclusion
- Deux mots collés pour le rythme ("Simple. Rapide.") — remets-les dans une vraie phrase
- Les auto-félicitations ("et ça change tout") — elles n'apportent rien
- Les phrases d'échauffement ("voici le truc :") — entre directement dans l'info
- La manie de tout présenter par trois — dans la vraie vie, on a deux raisons ou cinq
- Les fourchettes vagues ("5 à 10 minutes") — un chiffre net inspire plus confiance
- La conclusion qui récapitule — termine sur ton dernier point concret, c'est tout
- Les mots qu'on n'emploie pas à l'oral ("incontournable", "révolutionner") — remplace par du simple

Les marqueurs humains à ajouter :

- Un chiffre précis et concret (43 €, 4h30, 11 mois)
- Un vrai nom propre plutôt qu'"un outil" ou "un client"
- Une parenthèse avec une opinion (et un peu d'attitude)
- Une phrase qui commence par Et, Mais ou Parce que
- Un paragraphe d'une seule ligne
- Un petit agacement ou une hésitation assumée
- Un détail vrai mais un peu inutile

ASTUCE

Pour un texte purement pratique (guide, mail, notice), dis à Claude "utilise la norme ASD-STE100". C'est la norme des manuels d'avion : mots courts, phrases claires. Ça retire d'un coup une grande partie du style robotique.

Le prompt à copier-coller

Colle ce prompt dans Claude ou ChatGPT, puis ton texte en dessous.

PROMPT / COMMANDE

Tu es mon relecteur anti « style IA ». Je te donne un texte en français. Tu le réécrits pour qu'il sonne humain, sans changer le sens ni les infos.

1. Supprime ces tics d'IA :

- la formule « ce n'est pas X, c'est Y » (garde seulement l'idée finale)
- deux mots collés pour le rythme (« Simple. Rapide. »)
- les auto-félicitations (« et ça change tout »)
- les phrases d'échauffement (« voici le truc : », « soyons clairs »)
- la manie de tout présenter par trois
- les fourchettes vagues (« 5 à 10 minutes » -> un chiffre net)
- la conclusion qui répète le texte (« en résumé », « au final »)
- les mots jamais utilisés à l'oral (« incontournable », « révolutionner »)
- les tirets longs

2. Ajoute au moins quatre marqueurs humains :

- un chiffre précis et concret
- un vrai nom propre plutôt qu'« un outil » ou « un client »
- une parenthèse avec une opinion
- une phrase qui commence par Et, Mais ou Parce que
- un paragraphe d'une seule ligne
- un petit agacement ou une hésitation``
- un détail vrai mais un peu inutile

3. Garde le vouvoiement et le ton du texte d'origine.

4. N'invente jamais de fausses infos, fausses stats ou fausses citations. Ne remplace pas les mots par des synonymes bizarres. Ne rajoute pas de fautes.

Rends-moi d'abord le texte réécrit. Puis, en trois lignes, dis-moi ce que tu as changé.

Pour l'utiliser sans recoller à chaque fois : dans Claude, clique sur Personnaliser à gauche, puis Compétences, et enregistre ce prompt sous le nom `texte-humain`. Ensuite tape simplement `/texte-humain` quand tu en as besoin.