

Cassandra Crossing 689- Inquinare i propri scritti



Photo by [Etienne Girardet](#) on [Unsplash](#)

(689) - L'introduzione del watermarking nelle produzioni degli LLM come Claude conferma i rischi che Cassandra descrive da tempo, e che riguardano voi e tutto quello che scrivete.

20 agosto 2026 - I 24 lettori sanno bene che Cassandra usa metodi “manuali” ed antiquati in molte situazioni, a costo di essere derisa e sottovalutata.

Ha strane abitudini come gestire siti scrivendo l'HTML a mano con un editor di testo, rifuggire dai social e dalle dialettiche via chat, nonché non utilizzare LLM, neppure come tool “secondari”, durante la scrittura.

E proprio quest'ultima abitudine, che Cassandra ha da sempre ritenuta importante per la qualità degli elaborati e per l'esercizio della mente, si è recentemente dimostrata ancora più necessaria a causa dell'inserimento forzato di tecnologie di watermarking in Claude da parte di Anthropic.

Come probabilmente avete già letto, Claude genererà, automaticamente ed in maniera non disattivabile, un watermark per tutti i testi e tutte le immagini che genera.

Il caso delle immagini oggi non ci interessa; in questo caso il watermarking di Claude è semplicemente un set di metadati firmato digitalmente, e come tutti i metadati facilmente rimovibile. Non altera quindi, nemmeno minimamente, l'immagine a cui viene applicato.

Diverso è il caso del watermark per i testi; questo è molto più sofisticato, e funziona alterando sottilmente la scelta delle parole che l'LLM opera, senza alterare in maniera sostanziale il testo prodotto.

Questa scelta, per un dato LLM, è rilevabile tramite un algoritmo, purché si conosca una particolare coppia di chiavi pubblica/privata con cui l'LLM ha “criptato” le modifiche che costituiscono il watermarking.

L'algoritmo di verifica del watermarking di Claude fornisce un valore di probabilità che il testo sia stato elaborato con quel dato modello. E così fanno tutti quelli introdotti in altri contesti fino ad oggi. Si stabilisce poi un arbitrario e sempre discutibile valore di soglia per trasformarlo in una indicazione binaria “LLM sì/ LLM no”.

Nel tipo di watermarking implementato in Claude ci sono una quantità di problemi:

- il watermarking deve essere verificato separatamente per ciascun modello esistente;
- l'utilizzo dell'algoritmo di verifica del watermarking è vincolato al possesso di una chiave, e non è un semplice algoritmo riproducibile da chiunque;
- il watermarking è probabilistico, non fornisce certezza; è sempre e solo un *indizio*, non una *prova*
- il watermarking segnala solo il “passaggio” attraverso un LLM, non l'importanza di quello che è stato fatto durante l'elaborazione; una generazione completa del testo da parte dell'LLM od una correzione sintattica sono equivalenti.

E concentriamoci su questo ultimo problema; non che gli altri non siano altrettanto importanti, ma oggi semplicemente sono fuori tema.

E' pur vero che un verificatore di watermark può anche scambiare un testo interamente scritto da una persona per un testo generato da un LLM; dato il funzionamento probabilistico dei watermark questo non si può mai escludere.

Ma di contro qualsiasi passaggio attraverso un LLM del tipo che è stato implementato su Claude “macchierà” indelebilmente il vostro testo. Anche nel caso che l'impatto su di esso sia trascurabile per il contenuto, come una modifica di stile od una correzione sistematica sul testo.

Se siete interessati all'accettazione come opera originale dei vostri testi, in giro per la Rete od in situazioni accademiche, dovete assolutamente considerare questa possibilità.

E mentre non ci sono modi sicuri di evitare o cancellare un watermarking quando il testo è generato effettivamente dall'LLM da un prompt o comunque è stato utilizzato per elaborarlo, esistono modi ben noti per correggere e verificare errori di battitura, di punteggiatura (analisi del periodo) ed in alcuni casi di stile, semplicemente tramite i correttori dei vari programmi di videoscrittura che utilizzate.

Ed ora veniamo al punto ed alla raccomandazione accorata che la vostra profetessa preferita vuole farvi; l'argomento watermark era infatti solo un pretesto per poterne parlare.

Se avete studiato l'italiano (od un'altra lingua) e lo usate per scrivere testi e trasmettere qualcosa che avete nella testa (e no, le cose scritte in chat non sono questo), il vostro stile ed il vostro modo di esprimervi sono una parte di voi; una parte che vi sembra apparentemente secondaria ma che è invece sostanziale della vostra scrittura.

Qualsiasi passaggio attraverso un LLM azzera queste caratteristiche, convertendole nella media dei testi con cui è stato addestrato, ed eventualmente nella direzione richiesta dal prompt.

E' questo che volete?

Qualsiasi autore di qualsiasi epoca, se potesse capire la questione, risponderebbe con un “No!” indignato.

Ma voi la questione potete capirla perfettamente, e quindi siete del tutto responsabili per quello che farete, o non farete.

Sì, perché voi, se non siete molto, molto giovani, fino a pochissimi anni fa scrivevate tranquillamente le vostre cose con la vostra prosa ed il vostro stile, anche se con un po' più di fatica.

Sapete benissimo che usate gli LLM per una questione di produttività, o meglio di pigrizia, non di “qualità” del testo.

La qualità siete voi ed il vostro stile; se usate gli LLM la state in gran parte cancellando.

Non è importante per voi?

Pensateci.

[Scrivere a Cassandra](#) - [Mastodon](#) - [Buttondown.com](#)

[Videorubrica “Quattro chiacchiere con Cassandra”](#)

[Lo Slog \(Static Blog\) di Cassandra](#)

[L'archivio di Cassandra: scuola, formazione e pensiero](#)

[Cassandra per i posteri: l'archivio](#) su [Internet Archive](#)

Licenza d'utilizzo: i contenuti di questo articolo, dove non diversamente indicato, sono sotto licenza Creative Commons Attribuzione—Condividi allo stesso modo 4.0 Internazionale (CC BY-SA 4.0), tutte le informazioni di utilizzo del materiale sono disponibili a [questo link](#).