Skip to main content

[![Cornell University](/static/browse/0.3.4/images/icons/cu/cornell-reduced-white-SMALL.svg)](https://www.cornell.edu/)

We gratefully acknowledge support from the Simons Foundation, [member institutions](https://info.arxiv.org/about/ourmembers.html), and all contributors. [Donate](https://info.arxiv.org/about/donate.html)

[](/IgnoreMe)

[![arxiv logo](/static/browse/0.3.4/images/arxiv-logo-one-color-white.svg)](/) > [cs](/list/cs/recent) > arXiv:2306.04634

[Help](https://info.arxiv.org/help) | [Advanced Search](https://arxiv.org/search/advanced)

All fields Title Author Abstract Comments Journal reference ACM classification MSC classification Report number arXiv identifier DOI ORCID arXiv author ID Help pages Full text

Search

[![arXiv logo](/static/browse/0.3.4/images/arxiv-logomark-small-white.svg)](https://arxiv.org/)

[ ![Cornell University Logo](/static/browse/0.3.4/images/icons/cu/cornell-reduced-white-SMALL.svg) ](https://www.cornell.edu/)

open search

GO

open navigation menu

## quick links

* [Login](https://arxiv.org/login)
* [Help Pages](https://info.arxiv.org/help)
* [About](https://info.arxiv.org/about)

# Computer Science > Machine Learning

**arXiv:2306.04634** (cs)

[Submitted on 7 Jun 2023 ([v1](https://arxiv.org/abs/2306.04634v1)), last revised 1 May 2024 (this version, v4)]

# Title:On the Reliability of Watermarks for Large Language Models

Authors:[John Kirchenbauer](https://arxiv.org/search/cs?searchtype=author&query=Kirchenbauer,+J), [Jonas Geiping](https://arxiv.org/search/cs?searchtype=author&query=Geiping,+J), [Yuxin Wen](https://arxiv.org/search/cs?searchtype=author&query=Wen,+Y),

[Manli Shu](https://arxiv.org/search/cs?searchtype=author&query=Shu,+M),

[Khalid

Saifullah](https://arxiv.org/search/cs?searchtype=author&query=Saifullah,+K),

[Kezhi Kong](https://arxiv.org/search/cs?searchtype=author&query=Kong,+K),

[Kasun

Fernando](https://arxiv.org/search/cs?searchtype=author&query=Fernando,+K),

[Aniruddha Saha](https://arxiv.org/search/cs?searchtype=author&query=Saha,+A),

[Micah

Goldblum](https://arxiv.org/search/cs?searchtype=author&query=Goldblum,+M),

[Tom

Goldstein](https://arxiv.org/search/cs?searchtype=author&query=Goldstein,+T)

View a PDF of the paper titled On the Reliability of Watermarks for Large

Language Models, by John Kirchenbauer and 8 other authors

[View PDF](/pdf/2306.04634)

> Abstract:As LLMs become commonplace, machine-generated text has the

> potential to flood the internet with spam, social media bots, and valueless

> content. Watermarking is a simple and effective strategy for mitigating such

> harms by enabling the detection and documentation of LLM-generated text. Yet

> a crucial question remains: How reliable is watermarking in realistic

> settings in the wild? There, watermarked text may be modified to suit a

> user's needs, or entirely rewritten to avoid detection. We study the

> robustness of watermarked text after it is re-written by humans, paraphrased

> by a non-watermarked LLM, or mixed into a longer hand-written document. We

> find that watermarks remain detectable even after human and machine

> paraphrasing. While these attacks dilute the strength of the watermark,

> paraphrases are statistically likely to leak n-grams or even longer

> fragments of the original text, resulting in high-confidence detections when

> enough tokens are observed. For example, after strong human paraphrasing the

> watermark is detectable after observing 800 tokens on average, when setting

> a 1e-5 false positive rate. We also consider a range of new detection

> schemes that are sensitive to short spans of watermarked text embedded

> inside a large document, and we compare the robustness of watermarking to

> other kinds of detectors.

Comments: | 9 pages in the main body. Published at ICLR 2024. Code is available at [this https URL](https://github.com/jwkirchenbauer/lm-watermarking)

---|---

Subjects: | Machine Learning (cs.LG); Computation and Language (cs.CL); Cryptography and Security (cs.CR)

Cite as: | [arXiv:2306.04634](https://arxiv.org/abs/2306.04634) [cs.LG]

| (or [arXiv:2306.04634v4](https://arxiv.org/abs/2306.04634v4) [cs.LG] for this version)

| <https://doi.org/10.48550/arXiv.2306.04634> Focus to learn more arXiv-issued DOI via DataCite

## Submission history

From: John Kirchenbauer [[view email](/show-email/88cd89bf/2306.04634)]

**[[v1]](/abs/2306.04634v1)** Wed, 7 Jun 2023 17:58:48 UTC (14,947 KB)

**[[v2]](/abs/2306.04634v2)** Fri, 9 Jun 2023 17:58:04 UTC (14,993 KB)

**[[v3]](/abs/2306.04634v3)** Fri, 30 Jun 2023 18:18:12 UTC (14,994 KB)

**[v4]** Wed, 1 May 2024 21:20:36 UTC (20,657 KB)

Full-text links:

## Access Paper:

View a PDF of the paper titled On the Reliability of Watermarks for Large Language Models, by John Kirchenbauer and 8 other authors

* [View PDF](/pdf/2306.04634)
* [TeX Source](/src/2306.04634)
* [Other Formats](/format/2306.04634)



Current browse context:

cs.LG

[cs](/abs/2306.04634?context=cs)

[cs.CL](/abs/2306.04634?context=cs.CL)

[cs.CR](/abs/2306.04634?context=cs.CR)

### References & Citations

* [NASA ADS](https://ui.adsabs.harvard.edu/abs/arXiv:2306.04634)
* [Google Scholar](https://scholar.google.com/scholar_lookup?arxiv_id=2306.04634)
* [Semantic Scholar](https://api.semanticscholar.org/arXiv:2306.04634)

[a](/static/browse/0.3.4/css/cite.css) export BibTeX citation Loading...

## BibTeX formatted citation

×

loading...

Data provided by:

### Bookmark

[ ![BibSonomy logo](/static/browse/0.3.4/images/icons/social/bibsonomy.png)

](http://www.bibsonomy.org/BibtexHandler?requTask=upload&url=https://arxiv.org/abs/2306.04634&

description=On

the Reliability of Watermarks for Large Language Models "Bookmark on

BibSonomy") [ ![Reddit

logo](/static/browse/0.3.4/images/icons/social/reddit.png)

](https://reddit.com/submit?url=https://arxiv.org/abs/2306.04634&title=On the Reliability of Watermarks for Large Language Models "Bookmark on Reddit")

Bibliographic Tools

# Bibliographic and Citation Tools

Bibliographic Explorer Toggle

Bibliographic Explorer _([What is the Explorer?](https://info.arxiv.org/labs/showcase.html#arxiv-bibliographic-explorer))_

Connected Papers Toggle

Connected Papers _([What is Connected Papers?](https://www.connectedpapers.com/about))_

Litmaps Toggle

Litmaps _([What is Litmaps?](https://www.litmaps.co/))_

scite.ai Toggle

scite Smart Citations _([What are Smart Citations?](https://www.scite.ai/))_

Code, Data, Media

# Code, Data and Media Associated with this Article

alphaXiv Toggle

alphaXiv _([What is alphaXiv?](https://alphaxiv.org/))_

Links to Code Toggle

CatalyzeX Code Finder for Papers _([What is

CatalyzeX?](https://www.catalyzex.com))_

DagsHub Toggle

DagsHub _([What is DagsHub?](https://dagshub.com/))_

GotitPub Toggle

Gotit.pub _([What is GotitPub?](http://gotit.pub/faq))_

Huggingface Toggle

Hugging Face _([What is Huggingface?](https://huggingface.co/huggingface))_

Links to Code Toggle

Papers with Code _([What is Papers with Code?](https://paperswithcode.com/))_

ScienceCast Toggle

ScienceCast _([What is ScienceCast?](https://sciencecast.org/welcome))_

Demos

# Demos

Replicate Toggle

Replicate _([What is Replicate?](https://replicate.com/docs/arxiv/about))_

Spaces Toggle

Hugging Face Spaces _([What is

Spaces?](https://huggingface.co/docs/hub/spaces))_

Spaces Toggle

TXYZ.AI _([What is TXYZ.AI?](https://txyz.ai))_

Related Papers

# Recommenders and Search Tools

Link to Influence Flower

Influence Flower _([What are Influence Flowers?](https://influencemap.cmlab.dev/))_

Core recommender toggle

CORE Recommender _([What is CORE?](https://core.ac.uk/services/recommender))_

IArxiv recommender toggle

IArxiv Recommender _([What is IArxiv?](https://iarxiv.org/about))_

* Author
* Venue
* Institution
* Topic

About arXivLabs

# arXivLabs: experimental projects with community collaborators

arXivLabs is a framework that allows collaborators to develop and share new arXiv features directly on our website.

Both individuals and organizations that work with arXivLabs have embraced and accepted our values of openness, community, excellence, and user data privacy.

arXiv is committed to these values and only works with partners that adhere to them.

Have an idea for a project that will add value for arXiv's community? [**Learn more about arXivLabs**](https://info.arxiv.org/labs/index.html).

[Which authors of this paper are endorsers?](/auth/show-endorsers/2306.04634) | [Disable MathJax](javascript:setMathjaxCookie\(\)) ([What is MathJax?](https://info.arxiv.org/help/mathjax.html))

* [About](https://info.arxiv.org/about)
* [Help](https://info.arxiv.org/help)

* contact arXivClick here to contact arXiv [ Contact](https://info.arxiv.org/help/contact.html)
* subscribe to arXiv mailingsClick here to subscribe [ Subscribe](https://info.arxiv.org/help/subscribe)

* [Copyright](https://info.arxiv.org/help/license/index.html)
* [Privacy Policy](https://info.arxiv.org/help/policies/privacy_policy.html)

* [Web Accessibility Assistance](https://info.arxiv.org/help/web_accessibility.html)
* [arXiv Operational Status ](https://status.arxiv.org)

Get status notifications via [email](https://subscribe.sorryapp.com/24846f03/email/new) or [slack](https://subscribe.sorryapp.com/24846f03/slack/new)