Position: LLM Watermarking Should Align Stakeholders' Incentives for Practical Adoption

Yepeng Liu[†], **Xuandong Zhao**[‡], **Dawn Song**[‡], **Gregory W. Wornell**[§], **Yuheng Bu**[†]
[†]UC Santa Barbara [‡]UC Berkeley [§]Massachusetts Institute of Technology
{yepengliu, buyuheng}@ucsb.edu, {xuandongzhao, dawnsong}@berkeley.edu, gww@mit.edu

Abstract

Despite progress in watermarking algorithms for large language models (LLMs), real-world deployment remains limited. We argue that this gap stems from misaligned incentives among LLM providers, platforms, and end users, which manifest as three key barriers: competitive risk, detection-tool governance, and attribution issues. We revisit three classes of watermarking through this lens. *Model watermarking* naturally aligns with LLM provider interests, yet faces new challenges in open-source ecosystems. *LLM text watermarking* offers modest provider benefit when framed solely as an anti-misuse tool, but can gain traction in narrowly scoped settings such as dataset decontamination or user-controlled provenance. *In-context watermarking* (ICW) is tailored for trusted parties, such as conference organizers or educators, who embed hidden watermarking instructions into documents. If a dishonest reviewer or student submits this text to an LLM, the output carries a detectable watermark indicating misuse. This setup aligns incentives: users experience no quality loss, trusted parties gain a detection tool, and LLM providers remain neutral by simply following watermark instructions. We advocate for a broader exploration of incentive-aligned methods, with ICW as an example, in domains where trusted parties need reliable tools to detect misuse. More broadly, we distill design principles for incentive-aligned, domain-specific watermarking and outline future research directions. Our position is that the practical adoption of LLM watermarking requires aligning stakeholder incentives in targeted application domains and fostering active community engagement.

1 Introduction

The widespread adoption of large language models (LLMs) (Grattafiori et al., 2024; Yang et al., 2024) has intensified concerns about misuse, as these models increasingly produce human-like outputs. To enhance attribution and accountability, watermarking has emerged as a key approach. This includes LLM text watermarks, which embed imperceptible signals in generated content to identify AI output, and model watermarks, which encode signatures directly into model parameters to trace unauthorized use. Together, these techniques aim to support content provenance, protect intellectual property (IP), and promote trust in AI (Zhao et al., 2024b; Liu et al., 2024b; Pan et al., 2024).

Common *LLM text watermarking* strategies include embedding watermarks by perturbing the next-token prediction distribution (Kirchenbauer et al., 2023; Zhao et al., 2023a; Liu and Bu, 2024; Liu et al., 2024a), and employing pseudo-random sampling (Aaronson, 2023; Christ et al., 2023; Kuditipudi et al., 2023; Hu et al., 2023; He et al., 2024). These methods have demonstrated effectiveness in terms of detectability, robustness, and text quality. Beyond text, popular *model watermarking* techniques include watermarking during fine-tuning (Xu et al., 2023, 2025a; Zhao et al., 2024a; Nasery et al., 2025; Xu et al., 2024a), resisting model extraction via APIs (Zhao et al., 2023b; Sander et al., 2024; Panaitescu-Liess et al., 2025; Zhao et al., 2022), and protecting IP datasets (Jovanović et al., 2024a; Liu et al., 2025c; Wei et al., 2024). *However, despite substantial research efforts and many proposed techniques, real-world adoption of watermarking remains limited.*

In this paper, we explore the key reasons hindering the broader adoption of LLM watermarking, including *competitive risks*, governance challenges of *detection tools*, and *attribution* issues. Among these, we identify the **lack of aligned incentives** as the most fundamental barrier. Without clear benefits for LLM providers, users, and other stakeholders, even well-designed watermarking algorithms backed by regulation may struggle to achieve broad real-world adoption.

This position paper advocates that, to advance the adoption of LLM watermarking, we

Findings of the Association for Computational Linguistics: ACL 2026, pages 25887–25903

should look beyond technical performance and consider the broader ecosystem of stakeholders that influence design and deployment. We believe effective watermarking systems need to be tailored to specific application domains, grounded in clear threat models, and aligned with the interests of model providers, platforms, and users. Some existing concerns can be resolved when the use case is well defined and stakeholder incentives are properly addressed. **Our position does not oppose existing or future regulatory efforts, but emphasizes that regulation alone is insufficient to drive widespread adoption.** Effective uptake of watermarking requires well-aligned incentives, not solely by mandates. Moreover, since well-designed regulations and enforcement typically progress slowly, an incentive-aligned approach can be deployed immediately without requiring international consensus or regulatory oversight, helping to bridge the gap until effective global regulations are established.

To support our position, the remainder of the paper is organized around three types of watermarking techniques applied in different use cases: *model watermarking*, *LLM text watermarking*, and the newly proposed *in-context watermarking*. For each, we analyze the specific incentive model and assess whether stakeholder incentives are aligned to support real-world adoption. Where alignment exists, we draw analogies from traditional watermarking systems to highlight relevant lessons and design principles. Where alignment is lacking, we explore how existing techniques can be adapted or repurposed for alternative application domains to better align with stakeholder interests. Specifically,

- *Model watermarking* enables LLM developers to trace unauthorized uses without affecting end-user experience. By directly protecting the provider's IP at no cost to users, it aligns naturally with provider incentives and encounters minimal adoption friction, though it may face new challenges in open-source ecosystems.
- *LLM text watermarking* provides modest direct incentive for LLM providers when used solely to mitigate misuse (e.g., academic dishonesty) and can even push users toward unwatermarked LLMs. However, its value rises when repurposed to serve provider interests, such as filtering self-generated data to prevent model collapse, or helping users safeguard confidential material.
- *In-context watermarking (ICW)* embeds a watermark via instructions in the input prompt. It aligns provider incentives by placing control of watermarking in the hands of trusted parties, such as conference organizers or educators. Existing preliminary exploration demonstrates the effectiveness of ICWs on the most advanced LLMs, highlighting the promise of this direction for broad real-world deployment.

2 Issues of Current LLM Text Watermarking System

Despite steady progress in LLM watermarking, real-world adoption remains limited. To date, only Google's SynthID has been deployed on Gemini Web and mobile endpoints (Gloaguen et al., 2024). As Scott Aaronson noted at the ICLR 2025 Workshop on GenAI Watermarking, his proposal to integrate Gumbel-max watermarking into OpenAI's models was ultimately not adopted (Aaronson, 2023). We next outline three key barriers hindering real-world adoption.

① **Competitive Risks.** Implementing LLM text watermarking exposes early adopters to immediate competitive risks. If a single company adopts watermarking, users who fear being labeled as AI-generated, or who dislike the change, can switch to other LLM providers without watermarking or use open-source models locally. This dynamic places responsible companies at a disadvantage when prioritizing AI safety. In short, with the current usage of LLM text watermarking, the market discourages, rather than rewards, its adoption.

② **Governance of Detection Tools.** Current LLM text watermarking schemes typically rely on a secret key during detection, but distributing and managing this key in practice presents significant challenges. In most cases, LLM providers are expected to pair a watermarked LLM API with a detector API[1]. However, if the detection API or the key is made fully public, it opens the door to adversarial probing and spoofing, introducing additional security risks (Pang et al., 2024; Jovanović et al., 2024b). On the other hand, restricting access, such as limiting it to educational institutions or select third parties, requires costly centralized infrastructure and raises concerns about fairness, transparency, and antitrust issues.

③ **Lack of Attribution Issues.** Many high-

[1]Google recently released a public API for detecting SynthID watermarking in AI-generated images. See: https://blog.google/technology/ai/google-synthid-ai-content-detector/

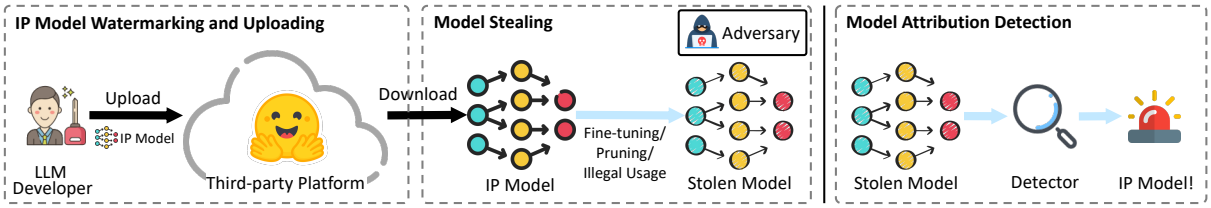

Figure 1: Example of model watermarking: an adversary fine-tunes, prunes, or illegally uses a protected model, and the LLM developers detect the unauthorized model.

quality LLM outputs reflect substantial human effort, including careful prompt engineering and post-editing, not just AI's contribution. Labeling such text as "AI-generated" (or worse, "AI-authored") using current watermarking tools overlooks the user's input and unfairly penalizes those, especially non-native speakers, who rely on LLMs for language assistance and refinement (Liang et al., 2023; Cooper et al., 2023). A multi-bit watermark that enables fine-grained attribution, rather than a simple binary indicator, may offer an effective solution.

3 Model Watermarking

We begin by examining model watermarking setups (Figure 1) and their incentive model, showing how stakeholder interests may align to support wider adoption. Drawing on lessons from traditional digital watermarking systems such as iTunes, we highlight key design principles while acknowledging the challenges unique to open-source platforms.

Related Work. Our focus is on model watermarking to protect model weights against misuse, such as theft, pruning, and unauthorized fine-tuning. Watermarking against other threats, like model stealing via distillation, typically requires distinct strategies and is discussed in Appendix A. In open-source LLMs, developers distribute weights via platforms like Hugging Face or GitHub. Developers or platforms can embed watermarks in model weights for later verification. One method is model backdooring (Xu et al., 2023, 2025a; Zhao et al., 2024a; Lou et al., 2023; Xue et al., 2023), where the LLM developer fine-tunes the model to associate a secret trigger (e.g., a specific token sequence) with a predefined output (e.g., a fixed phrase) that would not occur naturally. In this way, the model behaves as intended under normal usage but 'reveals' the watermark (abnormal behavior) with the trigger presented. Another method is model fingerprinting (Nasery et al., 2025; Xu et al., 2024a; Russinovich and Salem, 2024; Zhang et al., 2024a; Cai et al., 2024; Yamabe et al., 2024), where the model is fine-tuned on key-response pairs. Unlike backdoors, the responses are not a single secret phrase or overtly abnormal behavior, but subtle preference patterns across many queries that enable model identification.

3.1 Incentive Model

The incentive model, illustrated in Figure 2, involves three entities: IP owners, platforms, and end users, whose interests can align through model watermarking. In the open-source scenario, developers upload pretrained models to platforms such as Hugging Face to gain visibility and community adoption. Yet permissive licenses leave them exposed to adversaries who might rebrand, resell, or redistribute the weights without credit. In this context, model watermarking acts as a lightweight attribution mechanism, allowing developers to trace usage, assert ownership, and deter misuse without disrupting legitimate adoption by normal users.

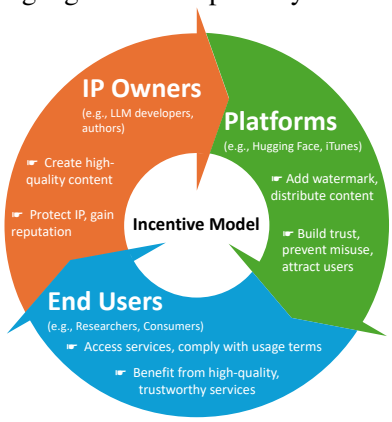

Figure 2: Incentive model for model watermarking among IP Owners, platforms, and users.

In the Model-as-a-Service (MaaS) setting (e.g., ChatGPT), LLM developers host their own APIs without an intermediary platform. Growing usage boosts their visibility, reputation, and subscription revenue. Adversaries, however, can erode this value by extracting large volumes of data, distilling the models, or deploying unauthorized replicas. And LLM developers have the incentive to adopt model watermarking (detailed in Appendix A) to trace such misuse and safeguard their IP.

From this perspective, model watermarking closely parallels traditional digital watermarking

Table 1: Analogy between model watermark and traditional digital watermark (e.g., iTunes).

	IP Owner	Platform	Adversary	Goal	Technique	Detector Owner
Model Weights	LLM Developer	Hugging Face	Model Thief	IP Protection	Watermark/Fingerprint	Platform or IP Owner
Digital Goods	Content Creator	iTunes	Media Pirate	IP Protection	Watermark/Metadata	Platform

systems used in domains such as music and visual art, where watermarks may be applied either by platform (e.g., Hugging Face or iTunes) or by IP owners to protect IP and deter unauthorized use. A more detailed comparison is provided in Table 1.

Because model watermarking directly safeguards an IP owner's core asset, it offers incentive for adoption and may see broad uptake. Potential deployment scenarios include tracing unauthorized model distillations and flagging research misconduct, as illustrated by recent plagiarism incidents (Xinhua News, 2024), where proprietary data was used as watermark to support claims of unauthorized model use.

3.2 What iTunes Taught Us

Given the similarity in incentive structures, the evolution of iTunes offers valuable insight into how model watermarking could develop over time. The iTunes Store, launched in 2003, illustrates a shift from restrictive enforcement toward user-centric, forensic watermarking. During the FairPlay era (2003-2007), tracks were sold as encrypted .m4p files playable only on Apple devices, with embedded metadata identifying the purchaser. While DRM (Digital Rights Management) restricted playback, the watermark itself was visible with any MP4 parser. In 2007, Apple removed DRM from newly purchased songs, but retained the user-specific forensic tag in .m4a file. Users gained the freedom to copy and play music anywhere, while a persistent identifier continued to link leaked files to the original purchaser. By 2009, the entire iTunes catalog was DRM-free, yet every download still carried a visible watermark, ensuring accountability without compromising user experience.

iTunes succeeded not because of sophisticated watermarking technology but because the watermark was embedded in a service users already preferred. Songs were cheap, easy to buy with one click, and synced seamlessly to iPods, putting user experience first. Over time, the watermark shifted from being restrictive (blocking unauthorized playback) to forensic (providing traceability if content leaked). It was stored as visible metadata, so no special tools were needed to detect it.

The case offers a useful lesson for model watermarking on platforms like Hugging Face: watermarking added by the platform should not degrade user experience, be transparently verifiable, and be integrated into a product offering that users actively prefer over unwatermarked alternatives.

3.3 New Challenges

While threat models share similarities, platforms like Hugging Face introduce new challenges distinct from traditional watermarking contexts. Unlike the music industry's standardized MP4 format, the model ecosystem lacks a unified file structure. Frequent model transformations (fine-tuning, repackaging, quantization, pruning) and open-source model distribution complicate watermarking (Dai et al., 2025; Xu et al., 2025b; Guo et al., 2025; Zhang and Koushanfar, 2024; Fernandez et al., 2024; Li et al., 2024a, 2023). Consequently, watermarks demand high robustness to persist through format conversions and weight modifications. Moreover, watermark needs to extend beyond model weights to associated assets like datasets, embeddings, and outputs, crucial for model training and deployment. (Dataset watermarking methods are discussed in Appendix A).

In open-source ecosystems, the potential for malicious platforms incentivizes developers to embed watermarks themselves, contrasting with the iTunes model, where the platform, not creators, typically applies watermarks. Additionally, while model watermarking aims to protect IP owners, it is susceptible to misuse. A dishonest model owner could fabricate infringement claims by withholding or falsifying watermark keys, falsely asserting that their watermark appears in another model. To mitigate such abuse, platforms like Hugging Face could serve as neutral arbiters. This role involves collecting and storing verified watermark metadata and maintaining clear ownership records. These governance needs highlight a crucial future direction: designing model watermarking systems for technical robustness, accountability, and trust within collaborative ecosystems. Moreover, to prevent platforms from becoming "trust bottlenecks", their neutrality could be secured in two ways. Economically, a platform's value as a two-sided market depends on its reputation; biased verification would drive develop-

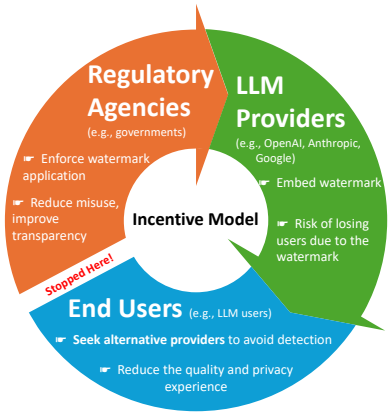

Figure 3: Broken Incentive Model for LLM Text Watermarking: Users may switch to unwatermarked models, undermining both the LLM provider's interests and the intended goal of reducing misuse.

ers to competitors, creating a self-correcting market incentive for fairness. Technically, platforms should adopt auditable, open-standard protocols to ensure the community can independently verify outcomes and prevent black-box governance.

4 LLM Text Watermarking

We now shift focus to watermarking the text generated by LLMs. In particular, we explore how the incentives of LLM providers may not align with the broader goals of AI safety, preventing the widespread adoption of text watermarking.

Related Work. LLM text watermarking typically embeds a watermark by manipulating the decoding process of LLMs, including logits perturbation and pseudo-random sampling. The detailed discussion can be found in Appendix A.

4.1 Misaligned Incentive

We analyze the incentive model (Figure 3) for the LLM text watermarking when used to prevent LLM misuse. In this setting, LLM providers embed watermarks in all generated text to prioritize AI safety. While this approach can help identify some adversarial uses, it also introduces a significant trade-off: users who object to having their generated content labeled or traceable may simply switch to unwatermarked models offered by other providers. As a result, while watermarking may help mitigate certain types of misuse and serve the public interest, it provides little direct benefit to LLM providers. This underscores a fundamental misalignment between the incentives of LLM developers and the goals of broader AI safety efforts.

One may expect that regulatory agencies may play a critical role in curbing AI misuse and internalizing its externalities. However, such efforts often fall short due to *jurisdictional limitations*, *competition from unregulated regions*, and the *widespread availability of locally deployed open-source models*.

Moreover, current LLM text watermarking techniques are limited to establishing the provenance of AI content, and they do not directly detect misuse. Hence, it should not be viewed as a universal fix for AI abuse. Instead, watermarking should be deployed in targeted settings where the incentives of stakeholders naturally align. Below, we present two use cases where existing techniques offer clear benefits to the stakeholders, showing how a shift in application domain can mitigate several of the challenges noted in Section 2.

4.2 Use Case: Watermarking Benefits LLM Developers

LLM text watermarking can be used by LLM developers to filter out texts generated by their models when collecting training data, as shown in Figure 4 Left. This helps prevent model collapse caused by training on synthetic data, as highlighted in (Shumailov et al., 2024). In this use case, watermarking serves as a data-curation tool, improving corpus quality rather than detecting misuse, which directly benefits developers by enhancing the performance of future LLMs.

In this case, all three issues discussed in Section 2 can be effectively resolved. ① Users are unlikely to be aware of the watermark, and it does not impact their experience, e.g., using the undetectable watermark (Christ et al., 2023), eliminating competitive risk. ② Since this watermark is used exclusively by developers, there is no need to make the detection tool public. ③ The watermark can be extended to encode multi-bit information, such as the model version and timestamp, to support fine-grained attribution.

Some text watermarks subtly shift the LLM token distribution. Because text watermark persists after unauthorized distillation (Gu et al., 2023; Sander et al., 2024), they provide a covert, model-level signature for ownership verification and theft tracing. Therefore, LLM text watermarking can also serve as model watermarking, deterring proprietary model extraction as discussed in Appendix A.

We note that these applications of LLM text watermarking benefit developers but rest on the assumption that providers act honestly. In settings like Chatbot Arena (Chiang et al., 2024), where

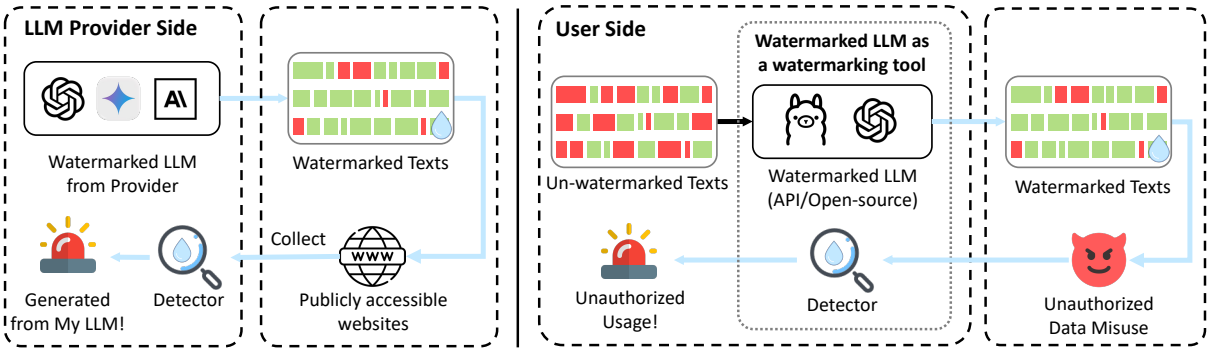

Figure 4: Illustration of two LLM text watermarking use cases. Left: Watermarking implemented by LLM Provider to detect self-generated data; Right: Watermarking implemented by the users to safeguard the user's document.

users evaluate model outputs blindly, a dishonest provider could exploit watermarking to identify their own model's responses, effectively bypassing the blind evaluation and unfairly boosting their leaderboard position. (Min et al., 2025; Singh et al., 2025). Future research and incentive design are needed to ensure watermarking serves as a tool for data integrity, not a means of deception.

4.3 Use Case: Watermarking Benefits LLM Users

LLM text watermarking can also protect end users' confidential material, as shown in Figure 4 Right. In one scenario, the users run an open-source LLM locally to paraphrase passages from a sensitive report; before returning the rewritten text, the model embeds an invisible watermark tied to the source document, much like a hidden "CONFIDENTIAL" stamp in a PDF. In a second scenario, a university partners with a commercial LLM provider to deliver summaries of restricted documents via a controlled-access API. Each response is likewise stamped with a covert mark derived from a secret key held only by the research team. In both situations, the watermark works as a digital signature that lets owners trace any unauthorized sharing.

In both scenarios, user-side watermarking addresses three of the main obstacles that limit the broader adoption of LLM text watermarking. ① Because the watermark is user-requested, it can be offered as an optional LLM feature that attracts users rather than deterring them. ② The secret key and the lightweight detector remain under user control, removing the need for a public API and mitigating misuse or antitrust concerns. ③ Multi-bit metadata (e.g., user ID, timestamp) enables reliable attribution by authenticating the watermark and identifying its source. The incentive structure is therefore well aligned: users gain a lightweight confidentiality enforcement tool, while LLMs (open-source or API-based) preserve output quality.

5 In-context Watermarking

Most existing LLM text watermarking methods focus on determining whether a piece of text was generated by an AI model, rather than addressing the misuse of LLMs in specific, high-stakes contexts. However, many real-world scenarios, such as a conference organizer trying to detect AI-written peer reviews or a teacher seeking to identify LLM-generated homework, involve content created outside these trusted parties' control. In these cases, existing LLM text watermarking approaches, which rely on modifying the model generation process, are difficult to apply. This highlights the need for alternative strategies that can operate in user-driven workflows.

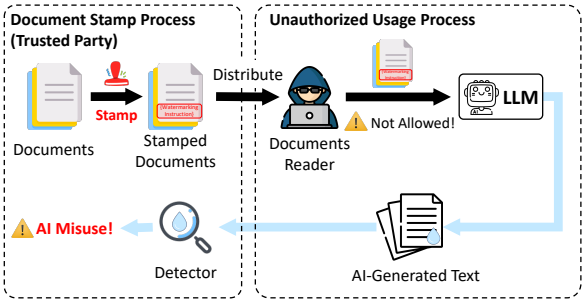

Figure 5: Overview of In-Context Watermark.

One promising approach is to modify the LLM input. Since many lazy reviewers (or students) paste documents directly into LLMs for summarization or drafting, documents can be embedded with imperceptible ICW instructions. These signals subtly influence the LLM's output, allowing downstream detection without altering the model or disrupting the reviewer's workflow.

This strategy motivates a new form of LLM text watermark, *ICWs*, and its application in sensitive settings such as peer reviews or student homework. ICWs leverage LLM's in-context learning (Dong

et al., 2022; Brown et al., 2020) and instruction-following abilities (Zhou et al., 2023; Mu et al., 2023) to embed detectable signals into generated text. By inserting carefully designed watermarking instructions into the prompt, LLMs can produce watermarked outputs, enabling reliable detection without modifying the model itself. The effectiveness of ICWs, including detection performance, robustness, and text quality, has been demonstrated empirically in existing works (Liu et al., 2025b; Zhong et al., 2024; Rao et al., 2025). Results indicate that, with well-designed watermarking instructions, ICW achieves strong performance across both proprietary and open-source models. Detailed experimental results are reported in these studies.

Related Work. The existing research on ICW is limited. Specifically, Liu et al. (2025b) investigates four ICW strategies: adding invisible characters, altering lexical choices, modifying initials, and using acrostics. The study finds that ICW effectiveness improves with LLM capabilities and shows strong performance in detection accuracy, robustness, and text quality across both in-process generation and indirect prompt injection scenarios (e.g., paper reviews). Zhong et al. (2024) proposes a method that uses a prompting LLM to generate context-aware watermarking instructions and a marking LLM to embed these watermarks into the generated text. A classifier is then trained to detect the presence of the watermark. Rao et al. (2025) designs a method specifically for detecting LLM-generated reviews. The approach injects a prompt into manuscripts that guides LLMs to include predefined patterns in the generated reviews, such as random start phrases, technical terms, or fake citations.

5.1 Exploration of Simple ICWs

The threat model (Figure 5) of ICW applied to peer review setting involves three entities: authors, reviewers, and conference organizers. Authors submit papers, and reviewers evaluate them. Conference organizers aim to maintain the integrity of the review process by identifying dishonest reviewers who violate policy by uploading submissions to LLMs for automated review. Organizers can covertly embed a watermarking instruction into the manuscript, for example, by using 'white text' (text colored the same as the background) within the PDF. If a reviewer inputs such a manuscript (containing the hidden instruction) into an LLM, the generated review may carry a detectable watermark. While authors could embed their own prompts to identify AI-generated reviews, this poses a conflict of interest, they may falsely accuse negative reviews of being AI-generated. Therefore, watermarking should be administered by conference organizers, who act as trusted parties.

A similar threat model applies to student homework, where the instructor embeds the watermarking instruction. Unlike authors in peer review, teachers do not have a conflict of interest, making the approach simpler to implement and manage in educational settings.

A simple example. As an illustrative example of ICW, we present Initials ICW, as introduced in (Liu et al., 2025b). The Initials ICW scheme embeds a watermark by encouraging LLMs to bias the initial letters of words in generated text to a subset (green letters) of English alphabet letters. An abbreviated watermarking instruction is shown below.

> ## Watermarking Instruction:
> Maximize the use of words starting with letters from {green_letter_list}.

Initials ICW increases the proportion of green initial letters in generated text, and detection is by computing the z-statistic over the frequency of green initial letters in the suspect text. It demonstrate effectiveness, especially for advanced LLMs with strong instruction-following capabilities.

In Appendix B, we present brief experimental results on advanced LLMs to illustrate the performance of ICWs and demonstrate their potential for practical deployment.

5.2 Incentive Model

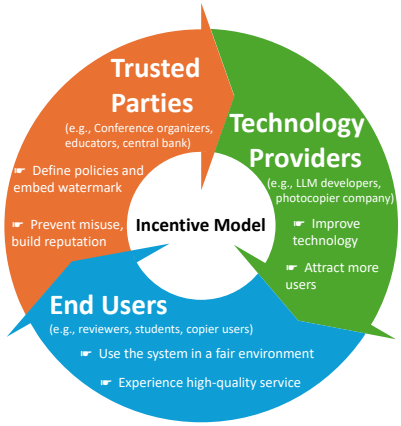

Figure 6: Incentive model for model watermarking among trusted parties, technology providers, and users.

The incentive model (Figure 6) of ICWs differs sharply from that of model and LLM text watermarking. For ICWs, the watermark is inserted not

Table 2: Analogy between ICW and EURion constellation.

	Trusted Parties	Technology Providers	Adversary	Goal	Response Mechanism
ICW	Organizer/Teacher	LLM Developer	Dishonest LLM User	Trace AI Misuse	Embed ICW
EURion	Central Bank	Photocopier Company	Counterfeiter	Prevent Money Counterfeit	Stop Service

by the LLM provider or the end user to protect their interests but by a trusted third party, such as conference organizers and educators, whose goal is to identify dishonest LLM use. The LLM provider's only requirement is to support reliable instruction following so that embedded watermarking instructions are executed reliably. This aligns incentives across stakeholders: organizers obtain higher-quality, human-authored reviews, instructors uphold academic integrity, normal users are unaffected; only dishonest behavior is flagged.

ICW aligns the incentives of different parties, thereby avoiding the usual deployment barriers. ① Because trusted parties embed the watermark in the prompt, LLM providers face no competitive risk and need not fear user loss. ② Governance of the detection tool is straightforward: trusted third parties alone hold the keys and detectors, eliminating conflicts of interest. ③ attribution is unambiguous: a detected mark directly links the LLM-generated text to a specific reviewer or student, enabling reliable enforcement of policy.

5.3 Rethinking the Analogy

The incentive model of ICW mirrors that of the EURion constellation used in modern banknotes. Central banks embed a subtle, machine-readable pattern in the currency; printers and photocopiers recognize the pattern to prevent money counterfeiting, while everyday users remain unaffected. Likewise, ICW lets a trusted party embed an imperceptible signal that LLMs dutifully follow, enabling reliable post-hoc detection of misuse without degrading the normal user experience.

Following this analogy (Table 2), it is conceivable that LLM providers could collaborate with conference organizers or universities to design specific patterns, similar to the EURion constellation, that elicit predefined model behaviors. For example, when a confidential document containing such a pattern is provided as input, the model could recognize its sensitivity and either embed an imperceptible watermark in the output or refuse to generate a response, while avoiding internalizing the content during training. This leads to a proactive misuse prevention method in LLM. Specifically, unlike the EURion constellation, embedding an imperceptible watermark in the output is less noticeable to users, whereas halting generation may signal the protection mechanism and invite workarounds or even denial-of-service attacks.

As an emerging paradigm, ICWs face several open challenges that warrant further investigation. These include their reliance on LLM instruction-following capabilities, robustness to sophisticated attacks, and ethical and transparency considerations. Finally, for clarity, we summarize the differences among model watermarking, LLM text watermarking, and in-context watermarking with respect to primary deployment actors, use cases, incentive alignment, and failure modes in Table 3.

6 Discussion of Future Direction

Toward Principled Multi-Bit Watermarking. Several of our use cases require richer provenance for reliable attribution, e.g., document ID, user ID, or timestamps, driving the need for multi-bit watermarking. Existing methods typically bolt on simple coding schemes in an ad hoc way (He et al., 2025; Yoo et al., 2023; Qu et al., 2024; Boroujeny et al., 2024; Wang et al., 2023). A more principled approach is to view watermarking as an information-embedding problem (Chen and Wornell, 2001; Martinian et al., 2005) and apply information-theoretic tools to establish fundamental limits. Recent efforts (He et al., 2025) have begun exploring this direction, which can inform the design of optimal coding strategies for more robust and efficient multi-bit watermarking.

Benchmarking ICW as a Measure of Instruction-Following Capability. An appealing aspect of ICW is that its effectiveness improves with more capable LLMs. As better instruction-following directly enhances watermark performance, it creates a natural incentive for LLM developers to support third-party watermarking use cases. To advance this direction, future work should establish standardized benchmarks to evaluate a model's ability to embed ICWs, positioning this as a new metric for instruction following. Building such datasets and evaluation protocols is only a starting point, but it will help guide both research and industry toward more reliable, user-controlled watermarking solutions.

Limitations

There are several alternative perspectives to the positions we present in the paper. Some researchers and policymakers advocate regulatory mandates to ensure consistent deployment and accountability. Because the harms of AI misuse are widely distributed and hard to monetize, market incentives alone are insufficient. As a result, top-down regulation is considered the most reliable path to achieving broad and timely adoption of watermarking technologies. Moreover, the emergence of anti-detection markets may challenge incentive alignment and hinder efforts to detect AI misuse, as LLM providers could also have incentives to weaken or bypass watermarking. Our framework mainly analyzes primary alignment goals and simplifies the diverse and often contested incentive structures in the real-world. In future research, we will explore more deeply how these stakeholders navigate complex strategic trade-offs across diverse regulatory environments. There is no single method that is a panacea. For those alternative perspectives and concerns, we have a more detailed discussion in Appendix C.

References

Scott Aaronson. 2023. Watermarking of large language models. https://simons.berkeley.edu/talks/scott-aaronson-ut-austin-openai-2023-08-17. Accessed: 2023-08.

Anirudh Ajith, Sameer Singh, and Danish Pruthi. 2024. Downstream trade-offs of a family of text watermarks. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14039–14053.

Li An, Yujian Liu, Yepeng Liu, Yuheng Bu, Yang Zhang, and Shiyu Chang. 2026. A reinforcement learning framework for robust and secure llm watermarking. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7181–7198.

Li An, Yujian Liu, Yepeng Liu, Yang Zhang, Yuheng Bu, and Shiyu Chang. 2025. Defending LLM watermarking against spoofing attacks with contrastive representation learning. *arXiv preprint arXiv:2504.06575*.

Maya Anderson, Guy Amit, and Abigail Goldsteen. 2024. Is my data in your retrieval database? membership inference attacks against retrieval augmented generation. *arXiv preprint arXiv:2405.20446*.

Dara Bahri, John Wieting, Dana Alon, and Donald Metzler. 2024. A watermark for black-box language models. *arXiv preprint arXiv:2410.02099*.

Massieh Kordi Boroujeny, Ya Jiang, Kai Zeng, and Brian Mark. 2024. Multi-bit distortion-free watermarking for large language models. *arXiv preprint arXiv:2402.16578*.

Jack T Brassil, Steven Low, Nicholas F Maxemchuk, and Lawrence O'Gorman. 1995. Electronic marking and identification techniques to discourage document copying. *IEEE Journal on Selected Areas in Communications*, 13(8):1495–1504.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jiacheng Cai, Jiahao Yu, Yangguang Shao, Yuhang Wu, and Xinyu Xing. 2024. Utf: Undertrained tokens as fingerprints a novel approach to LLM identification. *arXiv preprint arXiv:2410.12318*.

Souradip Chakraborty, Amrit Singh Bedi, Sicheng Zhu, Bang An, Dinesh Manocha, and Furong Huang. 2023. On the possibilities of ai-generated text detection. *arXiv preprint arXiv:2304.04736*.

Bilva Chandra, Jesse Dunietz, and Kathleen Roberts. 2024. Reducing risks posed by synthetic content: An overview of technical approaches to digital content transparency. Technical Report NIST AI 100-4, National Institute of Standards and Technology, Gaithersburg, MD.

Yapei Chang, Kalpesh Krishna, Amir Houmansadr, John Wieting, and Mohit Iyyer. 2024. Postmark: A robust blackbox watermark for large language models. *arXiv preprint arXiv:2406.14517*.

Harsh Chaudhari, Giorgio Severi, John Abascal, Matthew Jagielski, Christopher A Choquette-Choo, Milad Nasr, Cristina Nita-Rotaru, and Alina Oprea. 2024. Phantom: General trigger attacks on retrieval augmented language generation. *arXiv preprint arXiv:2405.20485*.

Brian Chen and Gregory W Wornell. 2001. Quantization index modulation: A class of provably good methods for digital watermarking and information embedding. *IEEE Transactions on Information theory*, 47(4):1423–1443.

Ruibo Chen, Yihan Wu, Junfeng Guo, and Heng Huang. 2025. Improved unbiased watermark for large language models. *arXiv preprint arXiv:2502.11268*.

Zhaorun Chen, Zhen Xiang, Chaowei Xiao, Dawn Song, and Bo Li. 2024. Agentpoison: Red-teaming LLM agents via poisoning memory or knowledge bases. *Advances in Neural Information Processing Systems*, 37:130185–130213.

Pengzhou Cheng, Yidong Ding, Tianjie Ju, Zongru Wu, Wei Du, Ping Yi, Zhuosheng Zhang, and Gongshen

Liu. 2024. Trojanrag: Retrieval-augmented generation can be backdoor driver in large language models. *arXiv preprint arXiv:2405.13401*.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, and 1 others. 2024. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Forty-first International Conference on Machine Learning*.

Miranda Christ, Sam Gunn, and Or Zamir. 2023. Undetectable watermarks for language models. *arXiv preprint arXiv:2306.09194*.

A Feder Cooper, Katherine Lee, James Grimmelmann, Daphne Ippolito, Christopher Callison-Burch, Christopher A Choquette-Choo, Niloofar Mireshghallah, Miles Brundage, David Mimno, Madiha Zahrah Choksi, and 1 others. 2023. Report of the 1st workshop on generative ai and law. *arXiv preprint arXiv:2311.06477*.

Xinyue Cui, Johnny Tian-Zheng Wei, Swabha Swayamdipta, and Robin Jia. 2025. Robust data watermarking in language models by injecting fictitious knowledge. *arXiv preprint arXiv:2503.04036*.

Cyberspace Administration of China. 2023. Interim administrative measures for generative artificial intelligence services. Issued 10 Jul 2023, effective 15 Aug 2023; requires explicit labels and implicit watermarks.

Yanbo Dai, Zongjie Li, Zhenlan Ji, and Shuai Wang. 2025. Seal: Subspace-anchored watermarks for llm ownership. *arXiv preprint arXiv:2511.11356*.

Sumanth Dathathri, Abigail See, Sumedh Ghaisas, Po-Sen Huang, Rob McAdam, Johannes Welbl, Vandana Bachani, Alex Kaskasoli, Robert Stanforth, Tatiana Matejovicova, and 1 others. 2024. Scalable watermarking for identifying large language model outputs. *Nature*, 634(8035):818–823.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, and 1 others. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

European Union. 2024. Regulation (eu) 2024/1689 of the european parliament and of the council of 13 march 2024 laying down harmonised rules on artificial intelligence and amending certain union legislative acts (artificial intelligence act).

Pierre Fernandez, Antoine Chaffin, Karim Tit, Vivien Chappelier, and Teddy Furon. 2023. Three bricks to consolidate watermarks for large language models. In *2023 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE.

Pierre Fernandez, Guillaume Couairon, Teddy Furon, and Matthijs Douze. 2024. Functional invariants to watermark large transformers. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4815–4819. IEEE.

Jiayi Fu, Xuandong Zhao, Ruihan Yang, Yuansen Zhang, Jiangjie Chen, and Yanghua Xiao. 2024. Gumbelsoft: Diversified language model watermarking via the gumbelmax-trick. *arXiv preprint arXiv:2402.12948*.

Soumya Suvra Ghosal, Souradip Chakraborty, Jonas Geiping, Furong Huang, Dinesh Manocha, and Amrit Singh Bedi. 2023. Towards possibilities & impossibilities of ai-generated text detection: A survey. *arXiv preprint arXiv:2310.15264*.

Eva Giboulot and Teddy Furon. 2024. Watermax: breaking the LLM watermark detectability-robustness-quality trade-off. *arXiv preprint arXiv:2403.04808*.

Thibaud Gloaguen, Nikola Jovanović, Robin Staab, and Martin Vechev. 2024. Black-box detection of language model watermarks. *arXiv preprint arXiv:2405.20777*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Chenchen Gu, Xiang Lisa Li, Percy Liang, and Tatsunori Hashimoto. 2023. On the learnability of watermarks for language models. *arXiv preprint arXiv:2312.04469*.

Qingxiao Guo, Xinjie Zhu, Yilong Ma, Hui Jin, Yunhao Wang, Weifeng Zhang, and Xiaobing Guo. 2025. Invariant-based robust weights watermark for large language models. *arXiv preprint arXiv:2507.08288*.

Haiyun He, Yepeng Liu, Ziqiao Wang, Yongyi Mao, and Yuheng Bu. 2024. Universally optimal watermarking schemes for LLMs: from theory to practice. *arXiv preprint arXiv:2410.02890*.

Haiyun He, Yepeng Liu, Ziqiao Wang, Yongyi Mao, and Yuheng Bu. 2025. Distributional information embedding: A framework for multi-bit watermarking. *arXiv preprint arXiv:2501.16558*.

Abe Bohan Hou, Jingyu Zhang, Tianxing He, Yichen Wang, Yung-Sung Chuang, Hongwei Wang, Lingfeng Shen, Benjamin Van Durme, Daniel Khashabi, and Yulia Tsvetkov. 2023. Semstamp: A semantic watermark with paraphrastic robustness for text generation. *arXiv preprint arXiv:2310.03991*.

Zhengmian Hu, Lichang Chen, Xidong Wu, Yihan Wu, Hongyang Zhang, and Heng Huang. 2023. Unbiased watermark for large language models. *arXiv preprint arXiv:2310.10669*.

Baihe Huang, Hanlin Zhu, Julien Piet, Banghua Zhu, Jason D. Lee, Kannan Ramchandran, Michael Jordan, and Jiantao Jiao. 2025. Watermarking using semantic-aware speculative sampling: from theory to practice.

Mingjia Huo, Sai Ashish Somayajula, Youwei Liang, Ruisi Zhang, Farinaz Koushanfar, and Pengtao Xie. 2024. Token-specific watermarking with enhanced detectability and semantic coherence for large language models. *arXiv preprint arXiv:2402.18059*.

Wenlong Ji, Weizhe Yuan, Emily Getzen, Kyunghyun Cho, Michael I Jordan, Song Mei, Jason E Weston, Weijie J Su, Jing Xu, and Linjun Zhang. 2025. An overview of large language models for statisticians. *arXiv preprint arXiv:2502.17814*.

Nikola Jovanović, Robin Staab, Maximilian Baader, and Martin Vechev. 2024a. Ward: Provable rag dataset inference via LLM watermarks. *arXiv preprint arXiv:2410.03537*.

Nikola Jovanović, Robin Staab, and Martin Vechev. 2024b. Watermark stealing in large language models. *arXiv preprint arXiv:2402.19361*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pages 6769–6781.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. In *International Conference on Machine Learning*, pages 17061–17084. PMLR.

Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. 2023. Robust distortion-free watermarks for language models. *arXiv preprint arXiv:2307.15593*.

Taehyun Lee, Seokhee Hong, Jaewoo Ahn, Ilgee Hong, Hwaran Lee, Sangdoo Yun, Jamin Shin, and Gunhee Kim. 2023. Who wrote this code? watermarking for code generation. *arXiv preprint arXiv:2305.15060*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Peixuan Li, Pengzhou Cheng, Fangqi Li, Wei Du, Haodong Zhao, and Gongshen Liu. 2023. Plmmark: a secure and robust black-box watermarking framework for pre-trained language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14991–14999.

Shen Li, Liuyi Yao, Jinyang Gao, Lan Zhang, and Yaliang Li. 2024a. Double-i watermark: Protecting model copyright for llm fine-tuning. *arXiv preprint arXiv:2402.14883*.

Xiang Li, Feng Ruan, Huiyuan Wang, Qi Long, and Weijie J Su. 2024b. Robust detection of watermarks for large language models under human edits. *arXiv preprint arXiv:2411.13868*.

Xiang Li, Feng Ruan, Huiyuan Wang, Qi Long, and Weijie J Su. 2025. A statistical framework of watermarks for large language models: Pivot, detection efficiency and optimal rules. *The Annals of Statistics*, 53(1):322–351.

Yuying Li, Gaoyang Liu, Yang Yang, and Chen Wang. 2024c. Seeing is believing: Black-box membership inference attacks against retrieval augmented generation. *arXiv e-prints*, pages arXiv–2406.

Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. 2023. Gpt detectors are biased against non-native english writers. *Patterns*, 4(7).

Aiwei Liu, Leyi Pan, Xuming Hu, Shu'ang Li, Lijie Wen, Irwin King, and Philip S Yu. 2023a. An unforgeable publicly verifiable watermark for large language models. *arXiv preprint arXiv:2307.16230*.

Aiwei Liu, Leyi Pan, Xuming Hu, Shiao Meng, and Lijie Wen. 2024a. A semantic invariant robust watermark for large language models. In *The Twelfth International Conference on Learning Representations*.

Aiwei Liu, Leyi Pan, Yijian Lu, Jingjing Li, Xuming Hu, Xi Zhang, Lijie Wen, Irwin King, Hui Xiong, and Philip Yu. 2024b. A survey of text watermarking in the era of large language models. *ACM Computing Surveys*, 57(2):1–36.

Mingrui Liu, Sixiao Zhang, and Cheng Long. 2025a. Mask-based membership inference attacks for retrieval-augmented generation. In *Proceedings of the ACM on Web Conference 2025*, pages 2894–2907.

Yepeng Liu and Yuheng Bu. 2024. Adaptive text watermark for large language models. *arXiv preprint arXiv:2401.13927*.

Yepeng Liu, Yiren Song, Hai Ci, Yu Zhang, Haofan Wang, Mike Zheng Shou, and Yuheng Bu. 2024c. Image watermarks are removable using controllable regeneration from clean noise. *arXiv preprint arXiv:2410.05470*.

Yepeng Liu, Xuandong Zhao, Christopher Kruegel, Dawn Song, and Yuheng Bu. 2025b. In-context watermarks for large language models. *arXiv preprint arXiv:2505.16934*.

Yepeng Liu, Xuandong Zhao, Dawn Song, and Yuheng Bu. 2025c. Dataset protection via watermarked canaries in retrieval-augmented LLMs. *arXiv preprint arXiv:2502.10673*.

Yixin Liu, Hongsheng Hu, Xun Chen, Xuyun Zhang, and Lichao Sun. 2023b. Watermarking text data on large language models for dataset copyright. *arXiv preprint arXiv:2305.13257*.

Qian Lou, Yepeng Liu, and Bo Feng. 2023. Trojtext: Test-time invisible textual trojan insertion. *arXiv preprint arXiv:2303.02242*.

Yueyuan Ma. 2022. Specialization in a knowledge economy. *Available at SSRN*, 4052990.

Emin Martinian, Gregory W Wornell, and Brian Chen. 2005. Authentication with distortion criteria. *IEEE Transactions on Information Theory*, 51(7):2523–2542.

Hasan Mesut Meral, Bülent Sankur, A Sumru Özsoy, Tunga Güngör, and Emre Sevinç. 2009. Natural language watermarking via morphosyntactic alterations. *Computer Speech & Language*, 23(1):107–125.

Rui Min, Tianyu Pang, Chao Du, Qian Liu, Minhao Cheng, and Min Lin. 2025. Improving your model ranking on chatbot arena by vote rigging. *arXiv preprint arXiv:2501.17858*.

Piotr Molenda, Adian Liusie, and Mark Gales. 2024. Waterjudge: Quality-detection trade-off when watermarking large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3515–3525.

Norman Mu, Sarah Chen, Zifan Wang, Sizhe Chen, David Karamardian, Lulwa Aljeraisy, Basel Alomair, Dan Hendrycks, and David Wagner. 2023. Can LLMs follow simple rules? *arXiv preprint arXiv:2311.04235*.

Anshul Nasery, Jonathan Hayase, Creston Brooks, Peiyao Sheng, Himanshu Tyagi, Pramod Viswanath, and Sewoong Oh. 2025. Scalable fingerprinting of large language models. *arXiv preprint arXiv:2502.07760*.

Leyi Pan, Aiwei Liu, Zhiwei He, Zitian Gao, Xuandong Zhao, Yijian Lu, Binglin Zhou, Shuliang Liu, Xuming Hu, Lijie Wen, and 1 others. 2024. MarkLLM: An open-source toolkit for LLM watermarking. *arXiv preprint arXiv:2405.10051*.

Michael-Andrei Panaitescu-Liess, Zora Che, Bang An, Yuancheng Xu, Pankayaraj Pathmanathan, Souradip Chakraborty, Sicheng Zhu, Tom Goldstein, and Furong Huang. 2025. Can watermarking large language models prevent copyrighted text generation and hide training data? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25002–25009.

Qi Pang, Shengyuan Hu, Wenting Zheng, and Virginia Smith. 2024. No free lunch in LLM watermarking: Trade-offs in watermarking design choices. *arXiv preprint arXiv:2402.16187*.

Lip Yee Por, KokSheik Wong, and Kok Onn Chee. 2012. Unispach: A text-based data hiding method using unicode space characters. *Journal of Systems and Software*, 85(5):1075–1082.

Jipeng Qiang, Shiyu Zhu, Yun Li, Yi Zhu, Yunhao Yuan, and Xindong Wu. 2023. Natural language watermarking via paraphraser-based lexical substitution. *Artificial Intelligence*, 317:103859.

Wenjie Qu, Dong Yin, Zixin He, Wei Zou, Tianyang Tao, Jinyuan Jia, and Jiaheng Zhang. 2024. Provably robust multi-bit watermarking for ai-generated text via error correction code. *arXiv e-prints*, pages arXiv–2401.

Vishisht Rao, Aounon Kumar, Himabindu Lakkaraju, and Nihar B Shah. 2025. Detecting LLM-written peer reviews. *arXiv preprint arXiv:2503.15772*.

Jie Ren, Han Xu, Yiding Liu, Yingqian Cui, Shuaiqiang Wang, Dawei Yin, and Jiliang Tang. 2023. A robust semantics-based watermark for large language model against paraphrasing. *arXiv preprint arXiv:2311.08721*.

Stefano Giovanni Rizzo, Flavio Bertini, and Danilo Montesi. 2016. Content-preserving text watermarking through unicode homoglyph substitution. In *Proceedings of the 20th International Database Engineering & Applications Symposium*, pages 97–104.

Mark Russinovich and Ahmed Salem. 2024. Hey, that's my model! introducing chain & hash, an LLM fingerprinting technique. *arXiv preprint arXiv:2407.10887*.

Tom Sander, Pierre Fernandez, Alain Durmus, Matthijs Douze, and Teddy Furon. 2024. Watermarking makes language models radioactive. *Advances in Neural Information Processing Systems*, 37:21079–21113.

Ryoma Sato, Yuki Takezawa, Han Bao, Kenta Niwa, and Makoto Yamada. 2023. Embarrassingly simple text watermarks. *arXiv preprint arXiv:2310.08920*.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759.

Shivalika Singh, Yiyang Nan, Alex Wang, Daniel D'Souza, Sayash Kapoor, Ahmet Üstün, Sanmi Koyejo, Yuntian Deng, Shayne Longpre, Noah Smith, and 1 others. 2025. The leaderboard illusion. *arXiv preprint arXiv:2504.20879*.

Mercan Topkara, Umut Topkara, and Mikhail J Atallah. 2006. Words are not enough: sentence level natural language watermarking. In *Proceedings of the 4th ACM international workshop on Contents protection and security*, pages 37–46.

Apurv Verma, NhatHai Phan, and Shubhendu Trivedi. 2025. Watermarking degrades alignment in language models: Analysis and mitigation. *arXiv preprint arXiv:2506.04462*.

Lean Wang, Wenkai Yang, Deli Chen, Hao Zhou, Yankai Lin, Fandong Meng, Jie Zhou, and Xu Sun. 2023. Towards codable watermarking for injecting multi-bits information to LLMs. *arXiv preprint arXiv:2307.15992*.

Johnny Tian-Zheng Wei, Ryan Yixiang Wang, and Robin Jia. 2024. Proving membership in LLM pretraining data via data watermarks. *arXiv preprint arXiv:2402.10892*.

Xinhua News. 2024. World insights: Stanford ai team apologizes for plagiarizing chinese university's model. Accessed: 2025-05-22.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*.

Jiashu Xu, Mingyu Derek Ma, Fei Wang, Chaowei Xiao, and Muhao Chen. 2023. Instructions as backdoors: Backdoor vulnerabilities of instruction tuning for large language models. *arXiv preprint arXiv:2305.14710*.

Jiashu Xu, Fei Wang, Mingyu Derek Ma, Pang Wei Koh, Chaowei Xiao, and Muhao Chen. 2024a. Instructional fingerprinting of large language models. *arXiv preprint arXiv:2401.12255*.

Xiaojun Xu, Yuanshun Yao, and Yang Liu. 2024b. Learning to watermark LLM-generated text via reinforcement learning. *arXiv preprint arXiv:2403.10553*.

Yijie Xu, Aiwei Liu, Xuming Hu, Lijie Wen, and Hui Xiong. 2025a. Mark your LLM: Detecting the misuse of open-source large language models via watermarking. *arXiv preprint arXiv:2503.04636*.

Zhenhua Xu, Xubin Yue, Zhebo Wang, Qichen Liu, Xixiang Zhao, Jingxuan Zhang, Wenjun Zeng, Wengpeng Xing, Dezhang Kong, Changting Lin, and 1 others. 2025b. Copyright protection for large language models: A survey of methods, challenges, and trends. *arXiv preprint arXiv:2508.11548*.

Jiaqi Xue, Mengxin Zheng, Ting Hua, Yilin Shen, Yepeng Liu, Ladislau Bölöni, and Qian Lou. 2023. TrojLLM: A black-box trojan prompt attack on large language models. *Advances in Neural Information Processing Systems*, 36:65665–65677.

Shojiro Yamabe, Tsubasa Takahashi, Futa Waseda, and Koki Wataoka. 2024. Mergeprint: Robust fingerprinting against merging large language models. *arXiv preprint arXiv:2410.08604*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Xi Yang, Kejiang Chen, Weiming Zhang, Chang Liu, Yuang Qi, Jie Zhang, Han Fang, and Nenghai Yu. 2023. Watermarking text generated by black-box language models. *arXiv preprint arXiv:2305.08883*.

Xi Yang, Jie Zhang, Kejiang Chen, Weiming Zhang, Zehua Ma, Feng Wang, and Nenghai Yu. 2022. Tracing text provenance via context-aware lexical substitution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11613–11621.

KiYoon Yoo, Wonhyuk Ahn, Jiho Jang, and Nojun Kwak. 2023. Robust multi-bit natural language watermarking through invariant features. *arXiv preprint arXiv:2305.01904*.

Eva Zhang, Arka Pal, Akilesh Potti, and Micah Goldblum. 2024a. vtune: Verifiable fine-tuning for LLMs through backdooring. *arXiv preprint arXiv:2411.06611*.

Junyan Zhang, Shuliang Liu, Aiwei Liu, Yubo Gao, Jungang Li, Xiaojie Gu, and Xuming Hu. 2025. Cohemark: A novel sentence-level watermark for enhanced text quality. *arXiv preprint arXiv:2504.17309*.

Ruisi Zhang, Shehzeen Samarah Hussain, Paarth Neekhara, and Farinaz Koushanfar. 2024b. {REMARK-LLM}: A robust and efficient watermarking framework for generative large language models. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 1813–1830.

Ruisi Zhang and Farinaz Koushanfar. 2024. Emmark: Robust watermarks for ip protection of embedded quantized large language models. In *Proceedings of the 61st ACM/IEEE Design Automation Conference*, pages 1–6.

Shuai Zhao, Meihuizi Jia, Zhongliang Guo, Leilei Gan, Xiaoyu Xu, Xiaobao Wu, Jie Fu, Yichao Feng, Fengjun Pan, and Luu Anh Tuan. 2024a. A survey of recent backdoor attacks and defenses in large language models. *arXiv preprint arXiv:2406.06852*.

Xuandong Zhao, Prabhanjan Ananth, Lei Li, and Yu-Xiang Wang. 2023a. Provable robust watermarking for ai-generated text. *arXiv preprint arXiv:2306.17439*.

Xuandong Zhao, Sam Gunn, Miranda Christ, Jaiden Fairoze, Andres Fabrega, Nicholas Carlini, Sanjam Garg, Sanghyun Hong, Milad Nasr, Florian Tramer, and 1 others. 2024b. Sok: Watermarking for ai-generated content. *arXiv preprint arXiv:2411.18479*.

Xuandong Zhao, Lei Li, and Yu-Xiang Wang. 2022. Distillation-resistant watermarking for model protection in nlp. *arXiv preprint arXiv:2210.03312*.

Xuandong Zhao, Lei Li, and Yu-Xiang Wang. 2025. Permute-and-flip: An optimally stable and watermarkable decoder for LLMs. In *The Thirteenth International Conference on Learning Representations*.

Xuandong Zhao, Yu-Xiang Wang, and Lei Li. 2023b. Protecting language generation models via invisible watermarking. In *International Conference on Machine Learning*, pages 42187–42199. PMLR.

Xin Zhong, Agnibh Dasgupta, and Abdullah Tanvir. 2024. Watermarking language models through language models. *arXiv preprint arXiv:2411.05091*.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

Chaoyi Zhu, Jeroen Galjaard, Pin-Yu Chen, and Lydia Y Chen. 2024. Duwak: Dual watermarks in large language models. *arXiv preprint arXiv:2403.13000*.

Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. 2024. Poisonedrag: Knowledge corruption attacks to retrieval-augmented generation of large language models. *arXiv preprint arXiv:2402.07867*.

A Other Related Works

LLM Text Watermarking. LLM text watermarking typically embeds a watermark by manipulating the decoding process of LLMs (He et al., 2025; Li et al., 2024b, 2025; Liu et al., 2023a; Zhang et al., 2025; Zhu et al., 2024; Fu et al., 2024; Xu et al., 2024b; Huo et al., 2024; Hou et al., 2023; Ren et al., 2023; Dathathri et al., 2024; Giboulot and Furon, 2024; Fernandez et al., 2023; Yoo et al., 2023; Qu et al., 2024; Ghosal et al., 2023; Chakraborty et al., 2023; Ji et al., 2025; Liu et al., 2024c), including logits perturbation (Liu and Bu, 2024; Liu et al., 2024a; Huang et al., 2025; Lee et al., 2023) and pseudo-random sampling (Aaronson, 2023; He et al., 2024; Chen et al., 2025; Zhao et al., 2025). Specifically, Figure 7 illustrates the green/red list watermarking method (Kirchenbauer et al., 2023), which partitions the LLM vocabulary into green and red token sets. The model is then subtly biased toward generating green tokens by modifying the output logits during sampling. (Aaronson, 2023) uses the Gumbel-Max trick to pseudo-randomly sample the next token during the text generation. Moreover, in addition to manipulating the decoding process, (Xu et al., 2024b) trains a paired LLM and detector to embed and detect watermarks collaboratively. (Bahri et al., 2024) introduces a black-box approach that, at each generation step, generates several candidate n-grams and selects the one with the highest hash-based score. Unlike in-process watermarking, which embeds the watermark during generation, post-hoc watermarking modifies text after it has been generated (Brassil et al., 1995; Por et al., 2012; Sato et al., 2023; Rizzo et al., 2016; Yang et al., 2023, 2022; Meral et al., 2009; Topkara et al., 2006; An et al., 2025, 2026; Chang et al., 2024; Zhang et al., 2024b; Qiang et al., 2023). Moreover, some works investigate the quality or alignment problem caused by existing watermarking techniques (Ajith et al., 2024; Molenda et al., 2024; Verma et al., 2025).

Proprietary LLM Extraction. For proprietary LLMs, the adversary typically has access only to the model's output (e.g., text) via its API. These outputs can then be used to label a substitute dataset, which enables the adversary to train a surrogate model. The most common strategy is LLM text watermarking, which involves manipulating the decoding process of proprietary LLMs (Kirchenbauer et al., 2023; Zhao et al., 2023a; Gu et al., 2023) without requiring additional training. The core idea is that the watermark signal embedded in the model-generated text can be learned by the surrogate model trained on those watermarked outputs, resulting in the surrogate's outputs also carrying detectable watermark information. Specifically, (Zhao et al., 2023b) injects a secret sinusoidal modulation into the token-generation logits, creating invisible 'spectral' signatures in the sequence of chosen tokens. (Sander et al., 2024) demonstrates the radioactivity of existing LLM text watermarks (Kirchenbauer et al., 2023), showing that when watermarked text is used as fine-tuning data, the watermark signal is transferred to the fine-tuned model.

Unauthorized Dataset Misuse. In addition to model IP infringement, protecting dataset IP is also crucial. Unauthorized users may incorporate proprietary datasets into their model training data or use them in Retrieval-Augmented Generation (RAG) systems (Karpukhin et al., 2020; Xiong et al., 2020; Lewis et al., 2020) without permission (Panaitescu-Liess et al., 2025; Liu et al., 2023b; Cui et al., 2025; Liu et al., 2025a; Li et al., 2024c; Anderson et al., 2024). The dataset owner usually embeds a backdoor (Chaudhari et al., 2024; Cheng et al., 2024; Chen et al., 2024) or watermark (Liu et al., 2023b; Jovanović et al., 2024a; Liu et al., 2025c; Wei et al., 2024) into the dataset for reliable detection. Specifically, (Zou et al., 2024) introduces a technique that involves inserting crafted malicious content into the dataset, causing retrieval-augmented LLMs to produce a specific, incorrect response to a targeted query. (Liu et al., 2025c) inserts carefully crafted watermarked canaries into the proprietary dataset to detect unauthorized use of the dataset in RAG systems.

B Experimental results of ICWs

In this section, we present brief experimental results on the detection and robustness performance of ICWs, given that the concept is relatively new. The performance is evaluated under the indirect prompt injection (IPI) setting, where we consider a scenario in which academic conference organizers embed watermarking instructions into submitted manuscripts and then detect if a review is generated by inputting the manuscript to an LLM.

The experiments use ICLR papers from 2020 to 2023 as a dataset. The ICWs are evaluated on `gpt-4o-mini` and `gpt-o3-mini`. The performance is evaluated using ROC-AUC, true posi-

Table 3: Comparative summary of watermarking settings across deployment actors, goals, incentive alignment, and potential failure modes.

Settings	Primary Deployment Actors	Primary Goal/Use Case	Incentive Alignment	Failure Modes
Model Watermarking	LLM developers or Platforms (e.g., Hugging Face)	IP protection	**Aligned:** Protect the provider's core asset without degrading user experience.	Transparency, disputes over ownership proofs
LLM Text Watermarking	LLM provider (e.g., OpenAI, Google)	Provenance of AI-misuse	**Misaligned:** Create competitive risk for model providers; users may switch to unwatermarked models.	Market rejection, key management
In-Context Watermarking	Trusted third parties (e.g., Conference organizers)	Provenance of AI-misuse	**Aligned:** Trusted parties gain detection tools; model providers remain neutral; users get fair services.	Dependence on model's instruction-following ability

tive rate at 1% false positive rate (TPR@1%FPR), and true positive rate at 10% false positive rate (TPR@10%FPR). The robustness is evaluated by paraphrasing the watermarked text using LLMs.

As shown in Table 4, ICW performance improves with increasing LLM capability, for example, from GPT-4o-mini to GPT-o3-mini. For advanced models such as GPT-o3-mini, ICWs achieve strong detection performance. Moreover, ICWs, such as Initials, Lexical, and Acrostics ICWs, remain certain robustness even when the watermarked text is completely paraphrased, demonstrating their potential for practical deployment. More extensive experimental results can be found in Liu et al. (2025b).

C Alternative Views and Discussion

Watermarking as A Mandatory Safety Baseline. Some people argue that waiting until all parties voluntarily adopt watermarks sets the bar too high. Instead, they treat provenance watermarking as a basic feature in the era of genAI, like seat belts or food allergy labels, that should be mandated through policy, not left to voluntary adoption. From this perspective, watermarking is not merely a market feature driven by aligned incentives but a necessary safeguard to protect society from the risks of AI-generated content.

Regulatory momentum supports this view:

- The EU AI Act (European Union, 2024) explicitly requires providers to embed machine-readable watermarks in any system that generates or manipulates content, with enforcement set to take effect in 2025.
- China's Cyberspace Administration (Cyberspace Administration of China, 2023) has gone further, mandating both visible and invisible watermarks for generative content and requiring platforms to detect and flag unmarked media.
- In the U.S., NIST's 2024 report on synthetic content (Chandra et al., 2024) frames watermarking as a foundational content authentication tool, recommended even in the absence of strong commercial incentives.

Supporters of this approach argue that because the harms of misuse, such as deepfake-driven misinformation, copyright infringement, are broadly distributed and hard to monetize (Ma, 2022), no individual stakeholder has a strong financial reason to act alone. They contend that without regulatory pressure, the market will reward providers who skip provenance controls in favor of speed, cost, or user satisfaction. As a result, they advocate for watermarking mandates backed by penalties and procurement rules, arguing that these top-down mechanisms are more likely to ensure timely and universal adoption than waiting for stakeholder incentives to naturally align.

The Existence of Anti-detection Markets. The emergence of anti-detection markets may challenge incentive alignment and hinder efforts to detect AI misuse, as LLM providers could also have incentives to weaken or bypass watermarking.

Discussion. *In this paper, we advocate for the consideration of incentive alignment when designing the LLM watermarking algorithm for broader deployment in the real world.* However, there is no single method that is a panacea. A well-designed regulation may also play an important role in the ecosystem. However, a well-designed regulation usually requires substantial effort, moves slowly, varies across jurisdictions, and is hard to keep pace with the rapid deployment of LLMs. In the meantime, we believe it is important to consider mechanisms such as ICW that can be deployed immedi-

Table 4: Empirical performance of ICWs: Comparing detection effectiveness and robustness against paraphrasing on indirect prompt injection (IPI) setting.

Language Models	Methods	Detection (IPI Setting)			Robustness (Paraphrase)		
		ROC-AUC	TPR@1%FPR	TPR@10%FPR	ROC-AUC	TPR@1%FPR	TPR@10%FPR
GPT-4o-mini	Unicode ICW	0.857	0.714	0.735	0.500	0.010	0.100
	Initials ICW	0.620	0.006	0.076	0.616	0.000	0.070
	Lexical ICW	0.889	0.054	0.564	0.887	0.048	0.556
	Acrostics ICW	0.592	0.002	0.448	0.591	0.000	0.378
GPT-o3-mini	Unicode ICW	1.000	1.000	1.000	0.500	0.010	0.100
	Initials ICW	0.997	0.910	0.998	0.893	0.106	0.628
	Lexical ICW	0.997	0.974	0.989	0.940	0.558	0.872
	Acrostics ICW	0.997	0.982	0.998	0.874	0.448	0.724

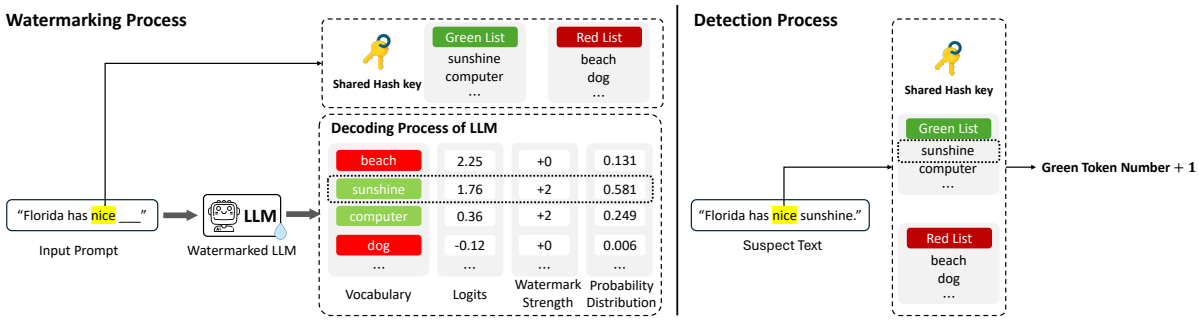

Figure 7: Illustration of Green/Red list LLM text Watermarking (Kirchenbauer et al., 2023).

ately without requiring international consensus or regulatory enforcement. Moreover, a simple, one-size-fits-all mandate risks overlooking the diverse incentives of different stakeholders. From a policymaking perspective, we believe it is important to account for these diverse incentives when designing watermarking requirements. In particular, effective regulation should aim to align incentives so that watermarking offers clear value not only to providers but also to trusted third parties and end users. By grounding regulation in incentive alignment, policymakers can reduce resistance and increase the likelihood of successful adoption.

Moreover, the existence of anti-detection markets may challenge incentive alignment, as LLM providers could have motives to weaken watermarking. However, this problem also arises under regulatory approaches. Moreover, incentive-aligned methods could better mitigate such risks: universal watermark mandates create a single target for evasion, while incentive-aligned methods apply mainly in high-stakes contexts with limited incentive to bypass detection. Incentive-aligned methods like ICW also align with providers' goals to improve instruction-following, requiring no universal cooperation, and opposing it would conflict with their broader interests. Because enterprises and educators all benefit from trustworthy watermarking, deliberately undermining it would bring reputational risks, making compatibility with incentive-aligned methods the rational choice.