ALTRU.DEV

From Model Provenance to Human Attribution

The Compositional Privacy Risk of AI Text Watermarking

DOCUMENT TYPE	Technical Research Note / Privacy Threat Analysis
AUTHOR	Valentyn Rukhaylo · Altru.dev
DATE	August 18, 2026
VERSION	1.0

Research classification: Privacy engineering · AI provenance · Generative AI governance · Attribution systems
Repository:github.com/altrudev/ai-watermarking-compositional-privacy

Central proposition

“A provenance signal may contain no identifying information and still participate in an identifying system.”

Abstract

AI text watermarking is increasingly being deployed as infrastructure for determining whether a generative model participated in producing a piece of text. Anthropic's recently announced Claude text watermark illustrates the intended model: the watermark is imperceptible to readers, provides evidence that Claude was involved in producing or processing text, and - according to Anthropic - carries no information identifying a particular user, organization, or conversation. Anthropic has also announced a forthcoming watermark detection API. [1]

This note argues that the privacy question cannot end there. A provenance signal does not need to contain personal identity to participate in a system capable of establishing personal identity. If a watermarked public artifact can be associated with a provider, and the provider separately retains generation outputs, request records, timestamps, sessions, account relationships, or other operational metadata, the artifact may in principle become linkable to a particular generation event and, from there, to an account or person.

The resulting privacy risk is therefore not limited to what information is encoded inside a watermark. It arises from composition. This note calls the resulting threat class **provenance-mediated identity linkage**: a sequence in which model-provenance information becomes one component in a broader correlation process capable of moving from artifact detection toward generation-event, account, organization, or human attribution.

CENTRAL PROPOSITION: A provenance signal may contain no identifying information and still participate in an identifying system.

This is not an allegation that Anthropic currently performs user attribution through Claude's watermark. Anthropic explicitly states that its present watermark does not identify individual users, and no evidence reviewed for this note establishes otherwise. Rather, this paper identifies a system-level privacy property that should be evaluated before AI provenance infrastructure becomes ubiquitous.

The issue is particularly timely because AI-generated-content marking is moving from experimental technology toward regulatory infrastructure. Article 50 of the EU AI Act applies from August 2, 2026, and the European Commission's Code of Practice addresses marking and detection of AI-generated content. [2][3] The policy problem is therefore no longer simply whether AI output should be detectable. It is also: **How far should detection be allowed to resolve identity?**

1. Research origin and scope

The central research question formalized in this note was proposed by **Valentyn Rukhaylo of Altru.dev on August 18, 2026**: if AI-generated text becomes systematically fingerprinted or watermarked, does the long-term privacy problem stop at identifying the model - or can the same provenance infrastructure ultimately contribute to identifying the human who used it?

There are already research systems explicitly designed to embed different watermarks for different users. Jiang et al. studied watermark-based user-level attribution by assigning a unique watermark to each user. PersonaMark similarly proposed personalized LLM watermarks for individual-user attribution. [4][5] More recent work argues that watermarking should be treated as a monitoring primitive and shows that entity-level inference can emerge from aggregated watermark signals in multi-key settings. [6]

Accordingly, this note does **not** claim to discover the general possibility of personalized or user-attributable watermarking. Its narrower concern is different: **Can user attribution become possible even when the public watermark itself is intentionally non-identifying and shared at the provider/model level?**

2. What Anthropic currently says Claude's watermark does

Anthropic announced Claude text watermarking on August 14, 2026. Future Claude models will generate text containing a watermark based on a version of the SynthID-Text approach. Anthropic describes the technique as

changing the source of randomness used when choosing among acceptable next-token options; it does not add hidden characters to the text. [1]

Anthropic says the watermark can estimate the likelihood that Claude was involved in producing or processing the text. It does not establish that Claude wrote the entire text, whether a person later edited it, or whether another model also participated. [1]

DOCUMENTED STATEMENT: Anthropic states that its text watermark “carries no identifying information” and cannot be traced through the watermark or its key to a specific person, organization, or chat. This note accepts that as the documented description of the present system. [1]

The privacy issue identified here does not require that statement to be false.

3. Embedded identity is not the only form of identity

The phrase “contains no identifying information” describes a property of one data object. It does not necessarily describe the privacy properties of the system surrounding that object.

```
ARTIFACT SIGNAL:  4f92a187

By itself:        non-identifying
External mapping: 4f92a187 -> Account 1837 -> Person A
Result:           identifying through association
```

Privacy engineering has dealt with this problem for decades. NIST SP 800-188 treats re-identification and linkage risk as system-level concerns and emphasizes that de-identification must consider whether data can be linked with other information. [7]

The same principle applies to AI provenance. The relevant privacy question is not only “Does the watermark encode user identity?” It is also “Can the artifact be linked to information that can eventually resolve user identity?” These are different properties.

4. Provenance-mediated identity linkage

This note proposes **provenance-mediated identity linkage** as a threat-model category for analyzing the transition from model provenance toward personal attribution.

```
PUBLIC ARTIFACT
      |
      v
MODEL-PROVENANCE DETECTION
      |
      v
PROVIDER / MODEL FAMILY
      |
      v
GENERATION-EVENT CORRELATION
      |
      v
REQUEST / SESSION
      |
      v
ACCOUNT
      |
      v
ORGANIZATION OR HUMAN
```

Level 1 - Model provenance

A detector determines that a particular model or provider likely participated in producing the artifact. This is the capability Anthropic presently describes. [1]

Level 2 - Provider attribution

The provenance result narrows the infrastructure family responsible for the generation. A provider-specific watermark therefore has informational value even when it carries no user identifier.

Level 3 - Generation-event attribution

The artifact is correlated with a particular generation event. Possible correlation material could include complete generated output, an output digest, an approximate textual fingerprint, semantic representations, timestamps, model/version data, or combinations of these. Whether a particular provider maintains any such index is an empirical question and should not be assumed.

Levels 4-7 - Session, account, organization, human

If a generation event corresponds to a service request or session, that record may be associated with an account. Enterprise or organizational environments may add another mapping layer. Registration, billing, employment, identity-verification, platform, or lawful-access records may ultimately map an account to a natural person.

KEY OBSERVATION: The watermark does not need to carry the information traversed in Levels 3-7. It only needs to participate in the chain.

5. A concrete system-level scenario

Suppose a person generates a passage through an AI provider and later publishes it pseudonymously. The public text contains only a model-level watermark. There is no user identifier in the watermark.

```
PUBLIC TEXT
  -> watermark detection -> Provider A likely involved

PROVIDER RECORDS
  Account -> Session -> Request -> Output

POSSIBLE COMPOSITION
  Public artifact -> output correlation -> generation event -> account
```

The first process provides provenance. The second provides linkage. Together they can potentially provide attribution. This note does **not** claim that Anthropic currently performs this operation. It demonstrates that user identity need not be encoded into the watermark for such an architecture to be technically possible.

6. Why data retention changes the threat model

Linkage becomes meaningful whenever generation-related records persist. Anthropic's current consumer-data documentation states that user conversations can remain in history until deletion; deleted conversations are generally removed from backend systems within 30 days. If users allow chats or coding sessions to be used to improve Claude, some data may be retained in de-identified form for up to five years. [8][9]

These policies are not evidence of watermark-to-account attribution. They establish a narrower but relevant fact: **generation-related data and account-associated service records may coexist for non-zero periods of time.** Once datasets coexist, privacy analysis should evaluate not only what each contains independently but what can be inferred when they are joined.

7. Component privacy is not compositional privacy

A recurring privacy failure occurs when systems evaluate every component independently. Three records can each appear non-identifying while their combination materially narrows identity.

```
Dataset A: Claude likely generated this text.
Dataset B: This text appeared publicly at 14:03:17.
Dataset C: Account 381 generated a near-match at 14:03:12.

Each alone: insufficient.
Combined: potentially identifying.
```

DISTINCTION: Component Privacy != Compositional Privacy.

This is not unique to AI. It is a standard problem in de-identification, database linkage, quasi-identifiers, and re-identification analysis. AI watermarking introduces a new context in which the same problem can operate over public text at large scale. [7]

8. Watermark detection is already evolving toward attribution

The research literature makes clear that detection and attribution are neighboring technical capabilities. Watermark-based user attribution has been studied explicitly, including unique-user watermarks and personalized LLM watermarking. [4][5]

Aremu, Lukas, and Zhang argue that watermarking should be treated as a monitoring primitive and show that even zero-bit watermarking can enable entity-level inference in multi-key settings when signals are aggregated. [6] Multi-bit watermark research likewise treats recoverable messages as a provenance capability rather than simple binary detection. [10]

GOVERNANCE LESSON: Detection, attribution, and identity resolution should not be governed as though movement between them were technically impossible.

9. The detector itself is a privacy boundary

Anthropic says it plans to offer a watermark detection API and is still working out implementation details. [1] That interface should itself be treated as an information-disclosure boundary.

```
MINIMAL DETECTOR
  watermark detected: yes / no / confidence

RICH DETECTOR (hypothetical)
  provider / model / version / key epoch / deployment / cohort / generation ID
```

There is a large privacy difference between these designs. Even metadata that does not name a person can narrow an attribution search. Detector design is therefore part of the privacy architecture, not merely a cryptographic verification concern.

10. The risk is larger than deliberate user-ID watermarks

A straightforward privacy debate asks whether AI companies should directly encode user IDs into generated content. That question matters, but it is too narrow. There are at least three routes to user attribution:

1. Direct identity encoding: watermark -> explicit user identifier.
2. Entity-specific watermarking: watermark pattern/key -> user or customer.

3. Compositional linkage: shared model watermark + generation records + auxiliary metadata -> user.

The third path is the focus of this paper. It allows both of the following statements to be simultaneously true: **The watermark contains no user identity. The surrounding system can associate the artifact with a user.**

11. Why this matters for anonymous and pseudonymous speech

Attribution infrastructure has legitimate uses: investigations of malicious automation, fraud, impersonation, coordinated manipulation, copyright disputes, abuse, academic misconduct, and other harms. But the same capability can affect legitimate anonymity and pseudonymity.

- whistleblowers using AI to improve the readability of a disclosure;
- journalists or confidential sources using AI-assisted editing;
- employees discussing misconduct under pseudonyms;
- dissidents or activists for whom identification creates physical or legal risk;
- people participating anonymously in health, abuse-support, sexuality, addiction, or other sensitive communities;
- researchers or ordinary users maintaining legitimate pseudonymous identities.

The question is not whether every anonymous user should be immune from lawful investigation. It is whether **model-provenance infrastructure should silently become identity-resolution infrastructure without a separately defined governance boundary.**

12. A proposed attribution taxonomy

Level	Capability	Question answered
1	**AI detection**	Was generative AI likely involved?
2	**Provider/model attribution**	Which provider or model was involved?
3	**Generation attribution**	Which generation event produced it?
4	**Session attribution**	Which interaction/session produced it?
5	**Account attribution**	Which service account produced it?
6	**Organization attribution**	Which organization controlled the account?
7	**Human attribution**	Which natural person was responsible?

A system's privacy disclosure should state which levels it supports. It should not describe Level 2 and leave users to infer that Levels 3-7 are therefore impossible.

13. Privacy properties that should be required

Model Attribution != User Attribution

Evidence that a model participated in producing content does not constitute authorization to identify its user.

Artifact Identifier != Human Identifier

An artifact-level signal should not silently become a personal identifier through hidden mappings.

Detection != Identity Resolution

The ability to verify provenance should not automatically grant access to account or identity records.

Retention != Attribution Authority

Possessing historical generation records does not by itself establish permission to use them for identity resolution.

Provider Knowledge != Public Disclosure

Information known internally to a service provider should not automatically become available through a public provenance detector.

Correlation != Authorization

Technical ability to correlate records does not determine whether the correlation is justified.

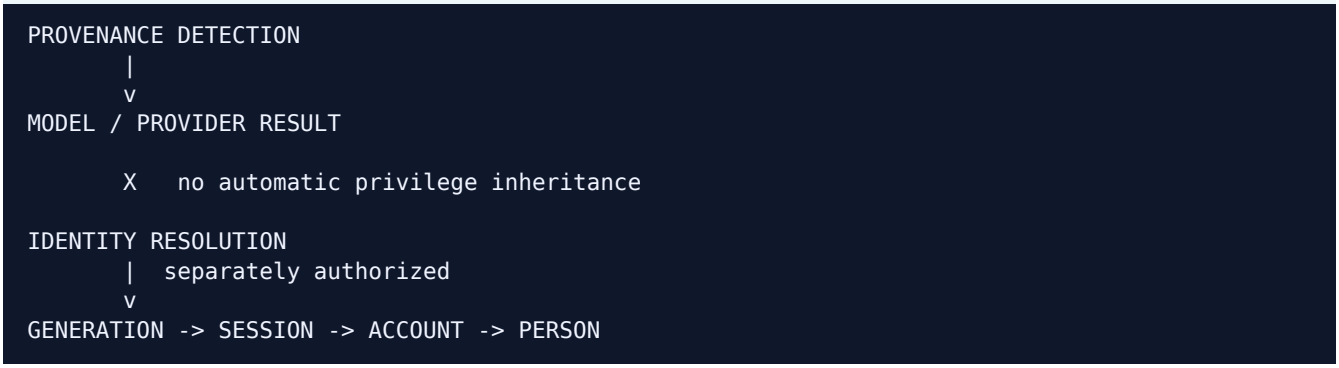
Provenance != Authorship != Ownership != Responsibility

Detecting model involvement does not establish authorship, ownership, or legal responsibility. Anthropic explicitly makes this distinction for its current watermark. [1]

14. Proposed Compositional Privacy Invariant

COMPOSITIONAL PRIVACY INVARIANT: Information classified as non-identifying in isolation must not be assumed non-identifying after correlation with provenance signals, generation records, content fingerprints, semantic representations, timestamps, account metadata, platform records, network observations, or other auxiliary datasets.

AI PROVENANCE RULE: A transition from model or artifact provenance to generation-event, session, account, organization, or human attribution should be treated as a separate identity-disclosure operation requiring an independent purpose, authority basis, access decision, evidence record, and audit trail.



15. Recommended safeguards

4. **Purpose separation.** Detection services and account-identification systems should have separate purposes, access controls, and audit domains.
5. **Minimum detector disclosure.** Public detection interfaces should return no more information than necessary for their stated purpose.
6. **No silent per-user keying.** If watermarks vary by account, user, organization, session, geography, or identifiable cohort, that property should be explicitly disclosed.

7. **Identity-resolution logging.** Any attempt to map an artifact to a generation event, account, organization, or person should itself create tamper-evident audit evidence.
8. **Independent authorization.** Access to a watermark detector should not imply access to generation or identity records.
9. **Retention minimization.** Output and request records should not be retained merely because they could later support attribution.
10. **Correlation analysis in privacy reviews.** Privacy assessments should examine datasets in combination rather than certifying each independently.
11. **Bulk-monitoring controls.** Detection APIs should consider abuse cases involving mass surveillance, longitudinal tracking, and automated scanning of public communications.
12. **False-attribution safeguards.** Any identity-resolution process must account for detector uncertainty, copied text, collaborative authorship, human editing, shared accounts, quotations, and ambiguous provenance.
13. **No retroactive purpose expansion without review.** Data collected to deliver an AI service should not silently acquire a new identity-tracing purpose merely because future watermarking makes correlation technically possible.

16. Regulatory significance

The European Commission states that Article 50 transparency obligations apply from August 2, 2026 and include machine-readable marking to enable detection of AI-generated or manipulated content. The Code of Practice provides rules for marking and detection. [2][3]

Those goals can be legitimate while still creating a second-order privacy problem. If marking becomes universal, provenance infrastructure can become part of the information architecture of the public internet.

POLICY PAIR: “How detectable should AI content be?” must be accompanied by “How linkable should AI users become?”

The two goals can conflict. A watermark made stronger, more persistent, more individualized, more interoperable, and more robust may improve provenance while simultaneously increasing monitoring capability. Recent watermarking research identifies this dual-use tension directly. [6]

17. Open research questions

Watermark architecture

- Does every user receive the same watermark key?
- Are keys different across model versions, deployment environments, regions, dates, organizations, or accounts?
- Can key families themselves become quasi-identifiers?

Detector architecture

- What information will public detection APIs return?
- Who may access higher-resolution internal detection?
- Are detector queries logged?
- Can detectors be used at bulk scale?

Generation correlation

- Are generated outputs retained in searchable form?
- Are exact output hashes, semantic embeddings, or approximate fingerprints retained?
- Could edited public text be associated probabilistically with retained generations?

Identity boundary

- Are watermark detection systems technically isolated from account systems?
- Who can perform a generation-event lookup?
- What authorization is required?
- Does such a lookup produce an auditable event?

Cross-platform correlation

- Could publication timestamps, platform accounts, IP records, browser identifiers, or other datasets materially narrow candidate generations?
- How does risk change when several organizations cooperate?

Governance

- What is the legitimate threshold for moving from content provenance to user identity?
- Should that transition require legal process, a defined abuse investigation, user consent, or another explicit authority basis?

18. Claim maturity

Documented

- Anthropic announced Claude text watermarking based on SynthID-Text principles and states that the watermark carries no user, organization, or chat-identifying information. [1]
- Anthropic says a watermark detection API is forthcoming. [1]
- EU Article 50 transparency rules apply from August 2, 2026 and include marking/detection obligations for covered AI-generated content. [2][3]
- Academic research has demonstrated explicit user-level attribution, personalized watermarking, and monitoring/linkability under certain watermark architectures. [4][5][6][10]

Derived

- A model-level provenance signal can become one input into a larger correlation architecture.
- A system does not need to encode a person’s identity inside the public watermark in order to associate an artifact with that person’s account through independent records.
- Privacy analysis therefore needs to evaluate relationships among watermark, detector, retained generation data, request/session data, and identity systems.

Hypothesized

- A provider could potentially correlate a public watermarked artifact against retained generation outputs or derived fingerprints and then resolve a matching event to an account.
- Cross-platform and provider-side metadata could increase attribution confidence.
- Future provenance standards could unintentionally lower the cost of large-scale identity linkage if detector granularity, output retention, and identity systems are insufficiently separated.

Explicitly not claimed

- This paper does not claim that Anthropic currently embeds user identities in Claude text watermarks.
- It does not claim that Anthropic currently maintains an output-fingerprint-to-user attribution database.
- It does not claim that Anthropic’s planned detection API will expose user identity.
- It does not claim that every watermarked artifact can be re-identified.
- It does not claim that watermarking is inherently undesirable.

19. Conclusion

AI watermarking is usually discussed as a question of content: **Can we determine whether an AI produced this?** The next question is about identity: **If we can determine which AI produced it, what else can that fact eventually allow us to determine?**

Those are not equivalent questions. Anthropic's statement that Claude's current watermark contains no identifying information may be completely accurate at the watermark layer. [1] But privacy is not a property of one layer. It is a property of the composed system.

A model-level watermark, retained generation records, request metadata, account relationships, timestamps, semantic matching, and external platform information may each be individually insufficient to identify someone. Their combination can have substantially different properties.

CENTRAL DISTINCTION: Embedded Identity != Linkability != Attribution

As AI provenance becomes more standardized, interoperable, and ubiquitous, this distinction should become a first-class privacy requirement. The objective should not be to eliminate provenance. It should be to prevent an unexamined transition from "this came from AI" to "this came from this person." Each arrow changes the privacy state. Each arrow should therefore require its own justification.

Suggested citation

Rukhaylo, Valentyn. *From Model Provenance to Human Attribution: The Compositional Privacy Risk of AI Text Watermarking.* Altru.dev Technical Research Note, Version 1.0, August 18, 2026.

Research provenance and assistance

Author and originating threat-model framing: Valentyn Rukhaylo · Altru.dev

The central system-level research question - whether non-identifying model provenance can become human attribution through correlation with generation and identity records - was formulated by Valentyn Rukhaylo on August 18, 2026.

Research assistance was provided using ChatGPT for literature discovery, source comparison, technical synthesis, and drafting. The note distinguishes documented facts from derived architectural implications and hypotheses.

References

[1] Anthropic. "How Claude's text watermark works." Aug. 14, 2026. Source

[2] European Commission. "Code of Practice on Transparency of AI-Generated Content." 2026. Source

[3] European Commission. "Transparency obligations for AI providers and deployers." 2026. Source

[4] Jiang, Zhengyuan, et al. "Watermark-based Attribution of AI-Generated Content." 2024. Source

[5] Zhang, Yuehan, et al. "PersonaMark: Personalized LLM Watermarking for Model Protection and User Attribution." 2024. Source

[6] Aremu, Toluwani, Nils Lukas, and Jie Zhang. "Watermarking Should Be Treated as a Monitoring Primitive." 2026. Source

[7] Garfinkel, Simson. "De-Identifying Government Datasets: Techniques and Governance." NIST SP 800-188, 2023. Source

[8] Anthropic Privacy Center. "How long do you store my data?" 2026. Source

[9] Anthropic Privacy Center. "Can you delete data sent via Claude?" 2026. Source

[10] Zhu, Junlin, Baizhou Huang, and Xiaojun Wan. "QuantileMark: A Message-Symmetric Multi-bit Watermark for LLMs." ACL 2026. Source

Repository:github.com/altrudev/ai-watermarking-compositional-privacy