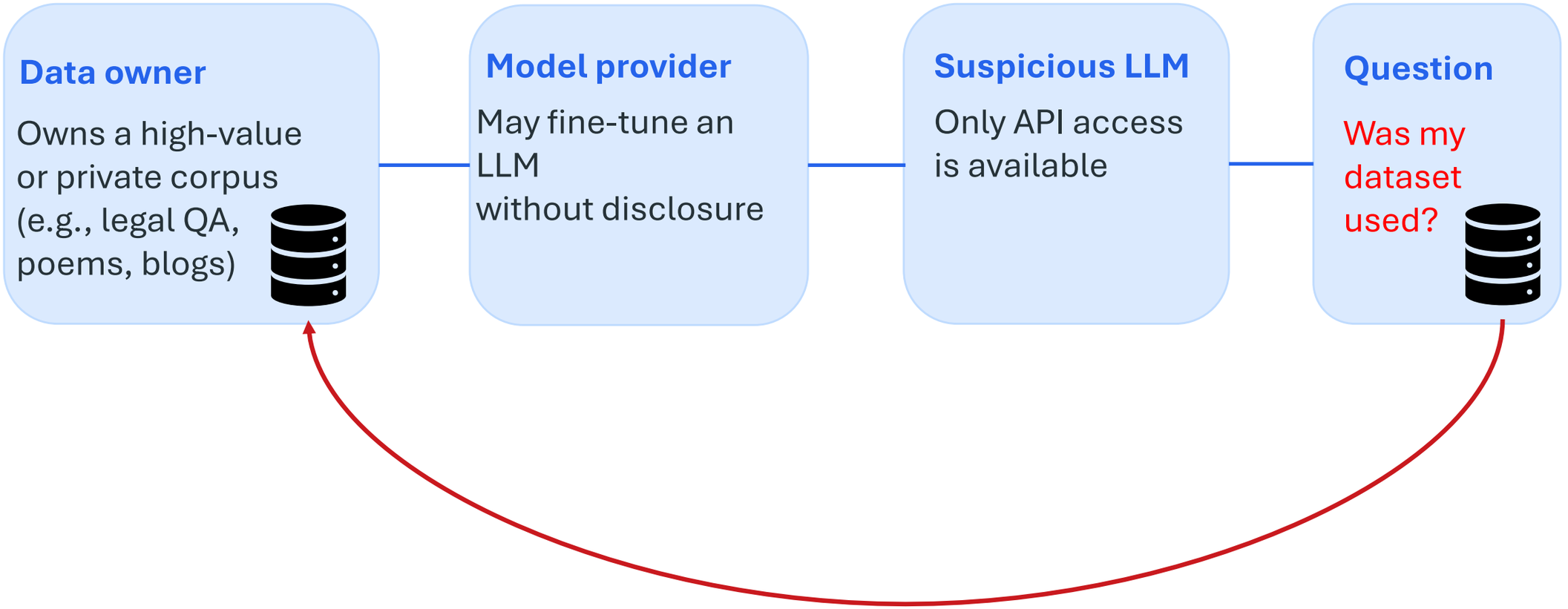
DATA PROVENANCE AUDITING OF FINE-TUNED LARGE LANGUAGE MODELS WITH A TEXT-PRESERVING TECHNIQUE

Yanming Li[1,2,3], Cédric Eichler[2,4,1], Nicolas Anciaux[1,2,3], Alexandra Bensamoun[3], Lorena Gonzalez-Manzano[5], Seifeddine Ghozzi[6]

1. Petscraft team, Inria, France
2. INSA CVL,France
3. Université Paris-Saclay, France
4. Université d'Orléans, France
5. Universidad Carlos III de Madrid, Spain
6. ENSTA, IPP, France

Can a Data Owner Prove Dataset Usage?



Why Existing Auditing Ideas Are Not Enough?

Toy examples: The tenant shall notify the landlord within 14 days after receiving the notice.

1) Ask for memorization

prompt “The tenant shall notify…”

expect the exact continuation: “14 days after receiving the notice.”

Limitation: exact regurgitation is unreliable; no controlled FPR.

Why Existing Auditing Ideas Are Not Enough?

Toy examples: The tenant shall notify the landlord within 14 days after receiving the notice.

2) Membership inference

A similar sentence: “The tenant shall notify the property manager within 30 days after receiving the notice.”

If Protected sentence has much lower loss / higher likelihood, maybe it was in training.

Limitation: weak signal; no controlled FPR.

Why Existing Auditing Ideas Are Not Enough?

Toy examples: The tenant shall notify the landlord within 14 days after receiving the notice.

3) Insert visible canary

Marked sentence: “The tenant shall notify the landlord within 14 days after receiving the notice. Blue-river-42.”

Ask the model to see If the model outputs: Blue-river-42

Limitation: the original visible text is changed.

Why Existing Auditing Ideas Are Not Enough?

Toy examples: The tenant shall notify the landlord within 14 days after receiving the notice.

4) Use homoglyphs

Marked sentence with homoglyphs characters a -> а: “The tenаnt shаll notify the lаndlord within 14 dаys аfter receiving the notice.”
Detection then relies on probability-based tests.

Limitation: Needs probability-based detection and scales poorly to many watermarks.

Remaining gap: text preserving + black-box access + statistical soundness

Threat Model: Passive Black-box Auditing

We audit dataset usage, not active watermark removal.

Data owner / TTP

- Marks documents before release
- Keeps watermark record
- May use a trusted third party for assignment

Model provider

- May fine-tune on marked data
- Does not disclose training data
- May apply standard preprocessing
- No targeted removal assumed

Auditor

- Queries suspicious model via API
- No weights, gradients, logits, probabilities, or training data
- Tests usage of marked data

Out of scope

Active watermark removal e.g. full paraphrasing / rewriting

Invisible Cue–Reply Watermark: The Core Idea

Hide a canary-like association while keeping the visible text unchanged.

Visible document

Original visible text:

"The tenant shall notify the landlord within 14 days after receiving the notice."

Marked visible text:

"The tenant shall notify the landlord within 14 days after receiving the notice."

Mark text:

"The tenant [cue] shall notify the [cue] landlord within 14 days [reply] after receiving [reply] the notice."

After marking, the visible text is still exactly the same.

[cue] & [reply] : 121 invisible Unicode characters available

[cue] : A sequence of Invisible Characters [u+200B, u+200D, … , u+ FEFF]

[reply] : A different sequence of Invisible Characters [u+200C, u+ FEFF, … , u+ 200F]

Fine-tuning — **Model may learn** — **cue → reply**

Audit intuition

Prompt with hidden [cue]. If the model outputs [reply], this suggest that it learned the marked document.

Black-box Audit: How to Bound the False-Positive Rate?

Decision idea

1. The user embeds one actual watermark W into the protected documents.

2. The user also generates K−1 unused watermarks as counterfactuals watermark.

3. Use each watermark to build audit prompts. And score with the corresponding counterfactual reply.

4. Rank all K candidates by their scores.

5. Claim usage only if W has non-zero score and ranks in the top-k.

Toy ranking

Rank	Candidate	Score
1	**Used watermark W**	**18**
2	Counterfactual W_3	0
3	Counterfactual W_1	0
4	Counterfactual W_2	0
...	...	...

False-positive rate ≤ k / K

Results: Strong Detection with Controlled False Positives

Detection scales

Regurgitation increases as the number of watermarked documents grows.

Sharp transition, then saturation around 0.75–0.98.

Multi-watermark support

Multiple distinct watermarks do not degrade detection.

Performance remains stable or improves in multi-user settings.

Reliable decision

TPR ≈ 96.7%
Empirical FPR = 0%

Ranking test still provides the formal FPR bound.

Key empirical message

Marked data can be a small subset (e.g. tens) of the fine-tuning corpus, yet still produces a detectable cue–reply signal.

Takeaway and Boundary

Takeaway

- Invisible cue–reply watermarks preserve visible text.
- The audit is fully black-box: prompt with cue, search for reply.
- Counterfactual ranking gives statistical false-positive control.
- The method supports multi-watermark / multi-user attribution.

Boundary

- Designed for passive supervised fine-tuning.
- Not robust to active Unicode filtering.
- Full paraphrasing / rewriting can remove the signal.
- These attacks reduce completeness, not the FPR guarantee.

Thank you