Training-Data Attribution

training-data-attribution · safety · concept

Source: https://policywindow.org/wiki/training-data-attribution

Generated 2026-07-09T20:26:36 UTC

Summary

Technical methods that identify which training examples most influenced a specific AI model output, enabling provenance claims about generated content and supporting copyright / consent / accountability disputes downstream.

At a glance

Used by

3 instrument(s)

Related concepts

ai-supply-chain, model-card, data-poisoning

Primary source

Grosse, R., et al. (2023), 'Studying Large Language Model Generalization with Influence Functions' (Anthropic) — the canonical articulation of scalable influence-function-based attribution for foundation models.

Details

Training-data attribution (TDA) is the inverse of training: given an output, recover the training examples that caused it. The technical lineage runs from influence functions (Koh & Liang 2017, 'Understanding Black-box Predictions via Influence Functions,' ICML) through gradient-based methods (Pruthi et al. 2020, TracIn) to recent scalable approximations for foundation models (Grosse et al. 2023, Anthropic, 'Studying Large Language Model Generalization with Influence Functions'; Park et al. 2023 TRAK). Adjacent methods include training-data extraction (Carlini et al. 2021, 'Extracting Training Data from Large Language Models') which surfaces verbatim memorisation rather than influence.

Governance relevance is now legally acute. The NYT v. OpenAI complaint (Dec 2023) used training-data extraction to show verbatim NYT articles in GPT-4 outputs; ongoing US copyright suits (Authors Guild v. OpenAI, Getty v. Stability AI, Tremblay v. OpenAI) turn partly on whether attribution methods can demonstrate substantial similarity at training-corpus scale. EU AI Act Art. 53(1)(c) requires GPAI providers to publish a 'sufficiently detailed summary' of training-data content — a disclosure obligation that is the regulatory analogue of attribution. China's GenAI Measures Art. 7 requires legal sourcing of training data. Brazil's PL 2338/2023 includes an explicit author-compensation provision. India's DPDPA does not yet address training-data rights directly, but the 2024 MEITY advisories signal forthcoming guidance.

Methodologically, TDA at frontier-model scale remains contested: influence-function approximations require restrictive assumptions (locally-linear loss surface) that don't hold for over-parameterised LLMs, and verbatim-extraction methods undercount the (likely larger) population of paraphrased or compositionally-derived outputs.

How to cite this article

APA

Policy Window. (n.d.). Training-Data Attribution [Wiki article — Concept]. https://policywindow.org/wiki/training-data-attribution

CHICAGO

Policy Window. n.d.. "Training-Data Attribution." Wiki article (Concept). https://policywindow.org/wiki/training-data-attribution.

HARVARD

Policy Window (n.d.) 'Training-Data Attribution', Wiki article — Concept, available at: https://policywindow.org/wiki/training-data-attribution.

OSCOLA

Policy Window, 'Training-Data Attribution' (Wiki article — Concept, n.d.) <https://policywindow.org/wiki/training-data-attribution> accessed [date].

BIBTEX

```
@misc{policywindow-training-data-attribution,
title  = {Training-Data Attribution},
author = {Policy Window},
year   = {n.d.},
howpublished = {training-data-attribution — safety},
url    = {https://policywindow.org/wiki/training-data-attribution},
note   = {Primary source: https://arxiv.org/abs/2308.03296}
}
```