

Robust Because Coarse, or Robust Because Planned? A Research Agenda for Training-Data Provenance After the Aggregate-Disclosure Turn, in Which Granularity Buys Statistical Power, Prior Randomisation Buys Soundness, and Per-Example Value Has Neither

Pranay Mahendrakar

SONYTECH · ORCID 0009-0003-7224-029X

pranaymahendrakar@sonytech.in · mahendrakarpranay@gmail.com · +91 6361723454

Abstract. Regulators have chosen coarse disclosure of training data: the template for the EU AI Act's public summary of training content asks for lists of large datasets, a summary of the most relevant crawled domains and a description of how opt-outs were honoured, and does not require a list of individual works. Compensation schemes, opt-out enforcement and licensing markets, by contrast, need a debt from a model to a particular work, and the technical literature reports that per-example attribution is unstable across training runs and can be manipulated. The natural reading is that provenance is robust where it is coarse and gameable where it pays. In this analysis the author reviews and synthesises published results to show that granularity is not the variable that separates the robust claims from the gameable ones. The paper separates four quantities that travel under the word provenance - composition, inclusion, exposure and contribution - and three parties who may be adversarial - the developer, the contributor and the valuator. Aggregation raises statistical power: membership inference that is at chance for single sentences reaches an AUROC of 0.943 on 500-document collections, and influence estimates for groups track actual group effects with rank correlations of 0.8 to 1.0 in convex models. Power is not soundness. A post-hoc test cannot sample its own null hypothesis at any granularity, while a randomised mark inserted before release gives a bounded false-positive rate for a single rightsholder. Coarse units leak: filtering a corpus by current robots.txt opt-outs cut the share of documents matching New York Times text only from 0.354 to 0.267 percent, much of the remainder being copies republished elsewhere. Per-example contribution is the one quantity that is unstable, gameable by contributors, who raised their compensation share to between 185 and 644 percent of its unperturbed value, and gameable by valuers who choose the utility. The paper consolidates the measured values in a table, maps six obligations to the quantity each needs, gives a decision procedure for choosing an instrument, and names seven measurements that would settle what the published record leaves open.

Keywords: training data provenance; data attribution; data valuation; membership inference; EU AI Act; opt-out; copyright; compensation

SONYTECH ORCID: 0009-0003-7224-029X

1. Introduction

Between July 2025 and August 2026 the question of what a general-purpose model was trained on stopped being a matter of voluntary documentation in the European Union. The template that implements Article 53(1)(d) of the AI Act was published on 24 July 2025, and the AI Office began enforcement on 2 August 2026, with models already on the market given until 2 August 2027 (Blankvoort et al., 2026). The template is a six-page document in three sections: general information, a list of data sources, and data processing. Its data-source section asks about public and non-public datasets, crawled and scraped data, user data and synthetic data; the crawled-data subsection asks for the crawlers used, their behaviour and "the most relevant domain names crawled"; the processing section asks the provider to describe the measures it took to respect reservations of rights under the text and data mining exception. In the rubric Blankvoort et al. (2026) built for the template, the URLs of individual works figure only as an optional extra in a comments field. The regulator chose coarse.

Other actors want something finer. Proposals for compensating the people whose writing trains a model allocate payments by each work's contribution to model outputs, computed with influence functions or cooperative-game values (Wang et al., 2024b; Deng et al., 2023; Zhang et al., 2025b). Kandpal and Raffel (2025) estimate that paying for the text behind 64 models released between 2016 and 2024 at conservative wage rates would cost 10 to 1000 times what it cost to train them, and argue that the data should be the most expensive part of a model. Opt-out regimes need to know whether a particular rightsholder's work was used after it was withdrawn. Licensing markets need to price a dataset before it is used and audit after. Each of these needs a quantity that attaches to a document or a rightsholder, not to a domain list.

The technical literature has been unkind to that finer quantity. Influence estimates in deep networks are fragile (Basu et al., 2021); the influence of a single example is often smaller than the noise from initialisation and batch order (Nguyen et al., 2023); data

values can rank points inconsistently across runs of the same procedure (Wang and Jia, 2023); and small, defensible changes to the utility function reorder who is worth what (Diehl and Wilson, 2025). Contributors can inflate their attributed share with imperceptible perturbations to their data (Wang et al., 2025a), and in federated training with local updates that leave global accuracy intact (Gao et al., 2026). Read together, these results suggest a tidy division of labour: coarse disclosure is robust because it is coarse, and fine-grained attribution is gameable because it is where the money is.

In this analysis the author reviews and synthesises published results to show that this division is wrong in a specific and useful way. Granularity does change something real: aggregating evidence over many documents raises the statistical power of a test, and aggregating training examples into groups makes influence estimates more stable. But power and soundness are different properties. Whether a provenance claim can be relied on against a party who would rather it were false depends on two things the granularity framing leaves out. The first is whether the test's null hypothesis can be sampled, which in practice means whether the rightsholder randomised something before the data was released. The second is who chooses the quantity being measured, which decides whether the valuator can game it as easily as the contributor can.

Four observations carry the argument.

First, coarse aggregation helps detection, and the effect is large. Puerto et al. (2025) report membership-inference AUROC of 0.501 for single sentences of arXiv text against Pythia 6.9B, 0.697 for whole documents and 0.943 for collections of 500 documents (Figure 1a). Maini et al. (2024) distinguish training from validation subsets of the Pile with about 1000 sequences and no false positives in their checks. Koh et al. (2019) find that influence-function predictions of group effects correlate with actual effects at Spearman 0.8 to 1.0 in regularised logistic regression, even where absolute errors are large.

Second, power is not soundness, at any granularity. Zhang et al. (2025a) argue that a membership test cannot prove training on a work, because the null distribution - the model's behaviour had the work been absent - cannot be sampled for a production model; they extend the objection to dataset inference, since publishing a work changes other documents that may have been trained on. Duan et al. (2024), Das et al. (2024) and Meeus et al. (2025) show that the published successes of sentence-level attacks largely measure distribution shift between the member and non-member sets. Aggregating a test whose null is not identified multiplies its power and its bias alike. What does provide a bounded false-positive rate is planned in advance: a random watermark or canary placed in the work before release (Wei et al., 2024; Zhang et al., 2025a). Meeus et al. (2024) use a randomised design of this kind but

report a detection AUC, not a false-positive bound. Verbatim extraction of long passages (Carlini et al., 2021; Cooper et al., 2025) is argued to have a negligible false-positive rate under heuristic conditions that Zhang et al. (2025a) only partially formalise.

Third, coarse units are neither self-verifying nor leak-proof. Fan et al. (2025b) filtered FineWeb-Edu by current robots.txt opt-outs and found that the share of documents with a 50-gram match to New York Times content fell only from 0.354 to 0.267 percent, and that many of the remaining matches were copies republished on domains that had not opted out; for news they needed a further decontamination step that removed 14 percent of the remaining tokens (Figure 1b). Crawlers comply unevenly with robots.txt (Kim et al., 2025; Liu et al., 2025a). The five public summaries Blankvoort et al. (2026) located by January 2026 scored between 33.3 and 92.9 percent on their transparency rubric. A domain list is a developer's self-report, and nothing in the model lets a third party check it.

Fourth, per-example contribution is the one quantity that is unstable, gameable by contributors and gameable by the valuator. Wang et al. (2025a) raise an adversary's share of top-k compensation to between 185.2 and 643.9 percent of its unperturbed value; Diehl and Wilson (2025) show that a valuator choosing among reasonable utility functions can move the median individual's Shapley payout by about five average payouts in one dataset; and Fan et al. (2025a) show that paying data owners their leave-one-out or Shapley value gives them a reason to misreport their costs. Coarsening helps against noise. It helps less against a valuator, because Diehl and Wilson's valuator can favour a group in aggregate.

From these the paper draws its contribution. It separates four quantities and three adversaries (Section 3) and shows which published results bear on which pair. It sets out the evidence on contribution (Sections 4 and 5), inclusion (Section 6), composition (Section 7) and exposure (Section 8). Table 1 consolidates the measured values; Table 2 maps six obligations to the quantity each needs and the instrument that can deliver it; Algorithm 1 states the decision procedure; Section 11 names seven measurements that would settle what the record leaves open.

The scope is text pretraining data for large language models, with results from image, music and federated settings brought in where they are the only evidence on a mechanism. The paper does not argue for or against any reading of copyright law. Where it says a quantity is or is not "sound", it means in the statistical sense of a test whose false-positive rate can be bounded against a party who would rather the test fail. Where it says a result is not known, the author searched for it and did not find it.

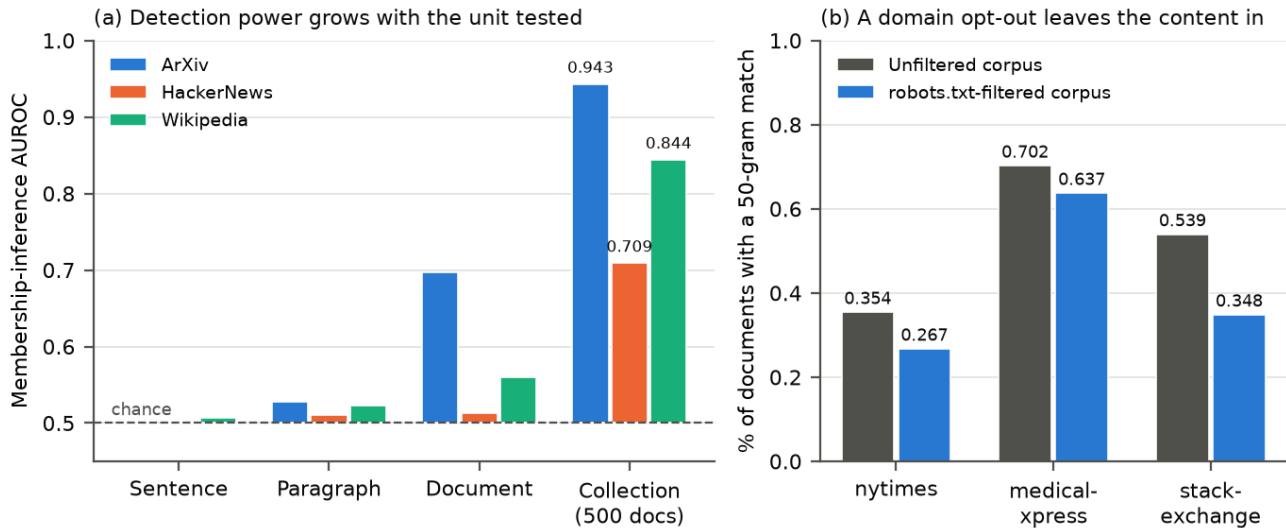


Figure 1. Two published measurements re-plotted. (a) Membership-inference AUROC for Pythia 6.9B at four text scales, from Puerto et al. (2025, Table 2); collections are sets of 500 documents. (b) Percentage of FineWeb-Edu documents with a 50-gram match to three domains, in the unfiltered corpus and in the corpus filtered by current robots.txt opt-outs, from Fan et al. (2025b, Table 1).

2. The aggregate-disclosure turn

2.1 What the template asks for

Blankvoort et al. (2026) reproduce the template's fields in the course of building a 242-metric quality rubric for it, and their reproduction is the most detailed public account of what the template asks. Three features matter here. The crawled-data section asks for crawler names, their purposes, a general description of their behaviour - including whether they respect robots.txt, paywalls and captchas - and a summary of the most relevant domain names crawled. The dataset section asks for a list of large publicly available datasets, with links. The processing section asks whether the provider signed the Code of Practice and asks it to describe the measures taken before training to respect reservations of rights, including opt-out protocols honoured by the provider or by third parties from whom datasets were obtained. Disclosure of the URLs of individual works is assessed only as an optional extra in the comments field.

Every one of those fields is either a list of sources at a coarse unit (dataset, domain, crawler) or a description of a process. None is an attestation that a particular work was or was not used, and none asks for a quantity that a third party could recompute from the model. That is a design choice, and Section 7 argues it is defensible for some obligations and empty for others.

2.2 What developers disclose when they are not required to

The Foundation Model Transparency Index scores developers against 100 indicators. Its first edition scored ten developers (Bommasani et al., 2023), at an average of 37 out of 100 as the second edition reports; the second, in which developers submitted reports, averaged 58 and found sustained, systemic opacity on copyright

status, data access and data labour (Bommasani et al., 2024); the third found the average had fallen to 40, with companies most opaque about training data and training compute (Wan et al., 2025). Where developers choose what to say about data, they mostly say little.

Below the level of model developers, dataset documentation is unreliable. The Data Provenance Initiative audited more than 1800 text datasets and found licences omitted in over 70 percent of cases on popular hosting sites and miscategorised in over half (Longpre et al., 2023). Its multimodal extension found that while fewer than a third of datasets carry restrictive licences, over 80 percent of the source content in widely used text, speech and video datasets carries non-commercial restrictions (Longpre et al., 2025). The licence a dataset declares and the terms attached to its contents diverge. Datasheets were proposed as a fix for exactly this gap (Gebru et al., 2018), and the audits show how far practice remains from it.

2.3 What the consent landscape looks like

Longpre et al. (2024a) audited the robots.txt files and terms of service of 14,000 web domains behind C4, RefinedWeb and Dolma. Within a single year, 2023 to 2024, restrictions rose until about 5 percent of C4's tokens, and over 28 percent of tokens from its most actively maintained and critical sources, were fully restricted by robots.txt, and terms of service restricted crawling for 45 percent of C4. They also found that robots.txt and terms of service frequently disagree with each other. Longpre et al. (2024b) argue that authenticity, consent and provenance tooling for training data is broken at every layer and call for universal provenance standards.

These three facts frame the problem. A regulator has asked for coarse disclosure. Developers volunteer little. And the signals that would tell a developer what it may use are inconsistent at the source. Into this landscape come proposals that require per-work quantities. The rest of the paper asks which of them the evidence can support.

3. Four quantities and three adversaries

3.1 Four quantities under one word

"Provenance" names at least four different quantities, and the published results do not transfer between them.

Composition is a property of the corpus: what fraction came from which datasets, domains, languages or licence classes. It is computed from the training data, not the model, and its natural unit is coarse. The template asks for composition.

Inclusion is a binary property of a work relative to a model: was it in the training set. Opt-out enforcement and most litigation turn on inclusion. Its unit can be a sentence, a document, a collection or a rightsholder's whole output.

Exposure is a property of model behaviour: does the model reproduce, or produce something substantially similar to, a given work. Output-side copyright questions turn on exposure (Henderson et al., 2023). Exposure can be tested without knowing the training set.

Contribution is a counterfactual: how much would some output, loss or capability change if a work were removed or reweighted. Compensation proposals turn on contribution. It is the quantity estimated by influence functions (Koh and Liang, 2017), TracIn (Pruthi et al., 2020), datamodels (Ilyas et al., 2022), TRAK (Park et al., 2023) and data valuation by Shapley, Banzhaf or Beta-Shapley values (Ghorbani and Zou, 2019; Jia et al., 2019; Wang and Jia, 2023; Kwon and Zou, 2022). Surveys organise these methods by the counterfactual each approximates (Hammoudeh and Lowd, 2024; Zhu and Cangelosi, 2025).

The four come apart in both directions. A work can be included and contribute nothing measurable; it can contribute without being reproduced, as Wang et al. (2025b) show when a Wikipedia passage ranks, on average, about 146th of roughly 320,000 training examples in value for a synthetic story on a similar topic that shares no tokens with it; and it can be exposed without being included, if a near copy was trained on instead. Chang et al. (2024) make the split between attribution and influence explicit at 8B scale: classical retrieval finds the passages that state a fact better than gradient methods do, but the gradient methods find the passages whose further training moves the model more.

3.2 Three adversaries

A claim is robust or gameable relative to someone. Three parties matter.

The **developer** controls the training set and the disclosure, and may prefer that inclusion not be provable or that composition look different from what it was. The developer is the adversary for opt-out enforcement and for disclosure accuracy.

The **contributor** controls the data offered and is paid by some function of it. The contributor is the adversary for compensation: it gains by inflating its measured contribution, by replicating its data, or by misreporting what the data cost (Wang et al., 2025a; Han et al., 2022; Fan et al., 2025a).

The **valuator** chooses the measurement: the utility function, the test set, the attribution method and its hyperparameters. When the valuator is also the payer, it gains by choosing the specification that favours whomever it prefers (Diehl and Wilson, 2025).

Robustness against one does not imply robustness against another. A provenance quantity that the developer cannot fake may still be one the contributor can inflate, and a quantity robust against contributors may still be at the valuator's discretion. The granularity framing treats "gameable" as one property. It is three.

3.3 What would settle the question

The tidy hypothesis predicts that robustness against each adversary rises as the unit of measurement coarsens. The alternative argued here predicts something else: coarsening raises power against noise for every quantity; robustness against the developer depends on whether the test's null can be sampled, which coarsening does not change; robustness against the contributor depends on whether the payment rule is replication-robust and incentive-compatible, which coarsening only partly changes; and robustness against the valuator depends on whether the utility is fixed before the data arrive, which coarsening does not change at all. Sections 4 to 8 take each quantity in turn.

4. Contribution is unstable at the unit where payments would attach

4.1 Training noise swamps single-example effects

Contribution is a counterfactual about training, and training is random. Wang and Jia (2023) computed leave-one-out and Shapley values for 20 CIFAR-10 images across five runs that differed only in the training seed. The run-to-run variance of each image's value exceeded its magnitude, and the average Spearman correlation between runs was about 0.001 for leave-one-out and about 0.038 for Shapley. Five mislabelled images could not be separated from the clean ones by their value distributions. Their proposed fix, the Banzhaf value, is the semivalue with the largest "safety margin"

against perturbations of the utility, and is more stable in their experiments; it is a fix to the weighting over coalitions, not to the noise in each coalition's utility.

Nguyen et al. (2023) treat the trained model as a draw from a posterior and test whether each train-test influence is distinguishable from zero. Averaged across pairs, even ground-truth leave-one-out influence had p-values of 0.331 on a three-class MNIST task and 0.692 on CIFAR-10 under the scheme that varies initialisation (SWA+DE-Init), and approximate methods were about as unstable, the influence function giving 0.352 and 0.575 on the same tasks. Under the scheme that fixes initialisation and varies only batch order (SWA+DE-Batch), the MNIST3 results are significant - leave-one-out at $p = 0.025$ and the influence function at $p < 0.001$ - while CIFAR-10 leave-one-out stays at 0.487. Under that batch-order scheme, moving from a small CNN to a large ViT fine-tuned with LoRA raised the leave-one-out p-value on MNIST3 from 0.025 to 0.786. Their recommendation is to trust attribution only for the rare pairs whose influence survives the noise.

Karthikeyan and Sogaard (2021) find that influence functions, TracIn, representer points and gradient-similarity methods all change their rankings with initialisation, data order and batch size, and argue that the instability follows from treating an example's influence as independent of model state and of other examples. Basu et al. (2021) find influence estimates fairly accurate for shallow networks and often wrong for deep ones, with accuracy varying by test point and depending on weight decay.

Each of these results has a counter-reading, and the counter-readings matter for an agenda. Epifano et al. (2023) argue that part of the reported fragility comes from the validation procedures rather than the estimator. Bae et al. (2022) show that practical influence estimates in nonlinear networks approximate a different object, the proximal Bregman response function, which still answers many of the questions influence functions are used for. Mlodozieniec et al. (2025) argue that influence functions are better read as predicting how the distribution of outputs over training runs depends on the data, and that this distributional reading explains why they work at all in deep learning. Deng et al. (2024) report that ensembling over several trained models improves attribution substantially, and that dropout or LoRA ensembles can recover most of the gain at a fraction of the cost. The common thread is that attribution can be made to measure something stable, but the stable thing is an expectation over training runs, not a fact about the one model a court or a market would be asked to pay on.

4.2 The utility function is underdetermined

A contribution score needs a utility: a model-quality measure evaluated on some test set. Diehl and Wilson (2025) take this seriously. They define three families of reasonable utilities - different handling of very small

coalitions, monotone transformations of a loss, and a cost-ratio threshold in a band from 0.5 to 0.6 - and compute Shapley and Banzhaf values under each on five medical and social datasets. In the credit dataset, switching between the worst and best small-coalition rule raised the median individual's Shapley payout by about five average payouts, and some by up to fifteen. Between 2 and 5 percent of observations changed fate under a filter that drops the bottom tenth, depending only on which utility was chosen. Banzhaf was more rank-stable under the small-coalition perturbation, as Wang and Jia's analysis predicts, but about as unstable as Shapley under the other two.

Wang et al. (2024a) reach a related negative from a different direction: without constraints on the utility, Data Shapley can select data no better than random, and it is optimal only for a narrow class of utilities. Hu et al. (2024) show that influence-based greedy selection of the most influential subset can fail even in linear regression because collective influence is not additive, and question whether additive evaluation metrics such as the linear datamodeling score are the right test. If the sum of individual contributions is not the contribution of the set, a compensation scheme that pays each work its individual score is paying for a quantity that does not add up to the value it is dividing.

4.3 At language-model scale, attribution and influence diverge

The results above are mostly on small models. At language-model scale the picture is mixed in a way that matters for compensation.

Akyurek et al. (2022) built a fact-tracing benchmark and found that a BM25 retriever with no access to the model retrieved the training examples that taught a fact better than gradient-based or embedding-based attribution: BM25 reached a mean reciprocal rank of 77.55 against 67.66 for the best neural combination. Chang et al. (2024) scaled gradient attribution to an 8B model and a 160B-token corpus. Retrieving from all of roughly 20 million T-REx sentences, BM25 reached an MRR of 0.592 and their method, TrackStar, 0.365. But when the retrieved passages were used for further training, TrackStar's proponents raised the target fact's probability by 0.90 percent against 0.41 percent for BM25's. The passages that state a fact and the passages that move the model are different passages. A TRAK replication that scored an MRR of 0.401 in the older 300-distractor setting scored 0.001 when retrieving from all 19.6 million sentences. What changed is the size of the candidate pool, not the model, and the authors trace the collapse largely to correctable design choices: margin-function gradients corrected "per example rather than per token" and "the lack of unit normalization". The number shows that a method tuned on a small candidate set can fail on a large one; it is not evidence about model scale.

DATE-LM complicates this from the other side. Jiao et al. (2025) argue that earlier fact-tracing benchmarks reward lexical overlap, and introduce a counterfactual variant that corrupts entities to break it; in that setting, attribution methods beat BM25 and representation similarity. On data selection, however, they find simple non-attribution baselines beating expensive gradient-similarity methods, no method dominating across tasks, and method rankings sensitive to evaluation design. Li et al. (2024) report that influence functions perform poorly on most of the language-model tasks they test, and attribute this to inverse-Hessian approximation error, uncertain convergence and the more basic problem that parameter change does not track behaviour change.

Influence analysis at scale does produce findings about what models learn from. Grosse et al. (2023) used EK-FAC influence functions on models up to 52B parameters and found influence patterns that grow more abstract with scale, alongside the limitation that influence drops to near zero when the order of key phrases is reversed. Ruis et al. (2025) used influence over 2.5B pretraining tokens for 7B and 35B models and found that reasoning answers draw on documents carrying procedural knowledge shared across questions, while factual answers draw on distinct documents that often contain the answer. Choe et al. (2024) scaled influence-based valuation to Llama3-8B-Instruct with up to a 6,500-fold throughput gain. Wang et al. (2025b) compute Shapley values within a single training run at negligible overhead.

What none of these establishes is that the per-document number is stable across retraining, robust to the choice of target output, or robust against a party with an interest in it. They establish that influence is a useful scientific instrument for studying a model's dependence on its data. A payment rule needs more.

4.4 Coarsening helps against noise

If single-example contribution is drowned by noise, aggregating examples should help, and the evidence says it does. Koh et al. (2019) removed groups of training points - points sharing feature values, feature or gradient clusters, random points within or across classes, and points of high positive or negative influence, in 1,700 subsets per dataset covering 0.25 to 25 percent of the data - and compared influence-function predictions with the effects of actually retraining. In L2-regularised logistic regression, predicted and actual group effects correlated at Spearman 0.8 to 1.0, despite large absolute errors. In two case studies with naturally occurring groups, the correlation was 1 for labelling functions on a chemical-disease task and 0.77 to 0.86 for crowdworkers on MultiNLI. Their theory shows the correlation need not hold in general, and their models are convex. Ilyas et al. (2022) show that linear datamodels predict the outcome of training on subsets well enough to support dataset counterfactuals. Na et al. (2024) find that the perplexity

of a model trained on a candidate data mixture is strongly correlated with that of a parameter average of models trained on the mixture's partitions, which makes partition-level ablations cheap. Wang et al. (2025b) report a domain-level value composition for a math validation set in which arXiv carries a share far above its share of the Pile.

So the tidy hypothesis is half right about contribution. Coarse contribution is more stable than fine contribution. Whether it is less gameable is a separate question, taken next.

5. Contribution is gameable from two sides

5.1 Contributors can inflate their share

Wang et al. (2025a) give the first systematic threat model. A data provider who contributes periodically, and who can either sample from the training distribution or query the deployed model, perturbs its next contribution so that more of it lands in the top-100 influential examples for validation points. Against influence functions, TRAK and Data Shapley on MNIST, CIFAR-10 and digits, the Shadow Attack raised the adversary's compensation share to between 206.3 and 456.1 percent of its unperturbed value; the black-box Outlier Attack raised it to between 185.2 and 643.9 percent. In the one text setting, NanoGPT on Shakespeare with TRAK, an adversary supplying 0.13 percent of the data raised its share from 0.0035 to 0.0092, a ratio of 262.9 percent. The perturbations are small enough that the authors describe them as imperceptible, and they exploit an inductive bias the influence literature has known about since Cook: outliers are influential.

Gao et al. (2026) move the attack into federated training with ten clients evaluated by a federated Shapley value. A client that would otherwise be ranked last mixes a small synthetic batch, generated by optimising a decoder's latent vector each round, into its local training. Its attribution rank rises from last (10th of 10) to between 2nd and 7th, in the upper half in 3 of 9 dataset-architecture settings, while global accuracy is preserved, and the credit of benign clients is reshuffled as a side effect. Label flipping, random noise and free-riding did not achieve the same effect. Zheng et al. (2025) describe an analogous overvaluation attack in collaborative learning and propose a Shapley variant under which truthful behaviour is optimal for clients under stated conditions.

These attacks are on small models. Nobody has published a contributor attack against attribution for a pretrained language model. The absence should not be read as safety: the inductive bias they exploit is generic, and the evidence in Section 4 suggests that the noise floor is higher at scale, which would make manipulation harder to detect as well as harder to measure.

5.2 Contributors can replicate

A second contributor strategy needs no optimisation. Data can be copied at zero cost, and many value notions reward a player who splits into several identities or submits duplicates. Agarwal et al. (2019) list free replicability as the first property that makes data unlike other goods in a market and design a fairness notion around it. Ohrimenko et al. (2019) build a collaborative market with replication-robust payments, and Han et al. (2022) characterise when semivalues, including Shapley and Banzhaf, are robust to replication in submodular games, finding that naive application can leave them open to it.

Duplication interacts with language models in a specific way. Kandpal et al. (2022) found that a sequence present ten times in training data is generated about a thousand times more often than one present once, and that detectors of memorisation are near chance on non-duplicated sequences. Carlini et al. (2022) describe memorisation growing log-linearly with the number of duplicates. Lee et al. (2022) found one 61-word sentence repeated over 60,000 times in C4. Elazar et al. (2024) found about half the documents in RedPajama duplicated. A compensation rule that pays by measured influence or by detectable memorisation rewards a rightsholder for seeding copies of their work across the web, and the corpora already contain the copies. Deduplication by the developer partly defends against this, and is in the developer's interest for model quality (Lee et al., 2022), but it is also a valuator choice: whether two near-copies count as one work or two is a utility specification.

5.3 The valuator can choose the answer

Diehl and Wilson's (2025) gameability results are about the valuator, not the contributor. A valuator who enumerates a small set of defensible utility specifications and computes values under each can pick the one that favours a chosen individual or group, and their Propositions 1 to 3 show that this search is feasible relative to the cost of computing the values once, for finite candidate sets, for small-coalition rules and for cost-ratio thresholds. Their appendix shows that randomly chosen groups containing 10 percent of a dataset can be favoured in aggregate, with no offsetting losses within the group. Coarsening the unit of payment from individuals to groups therefore does not remove valuator gaming; it changes the target from a person to a group.

Valuator gaming is the form of the problem most often missed in compensation proposals. Wang et al. (2024b) propose compensating copyright owners in proportion to Shapley-based contributions to each generation, and Deng et al. (2023) propose a royalty scheme for music generation that attributes value to training pieces by approximate causal influence, reporting reasonably high correlation with retraining-based attribution and finding

that perceived similarity identifies the most influential pieces but misses the broad contribution. Zhang et al. (2025b) propose fairshare pricing grounded in data valuation. In each, the developer or platform that pays also chooses the utility, the target outputs and the attribution method. Diehl and Wilson's result says that this discretion is enough to redirect the payments.

5.4 Values are not prices

Fan et al. (2025a) take the market-design view. Data owners incur private costs; a payment rule should make reporting those costs truthfully the owner's best strategy. They show that paying owners their leave-one-out value or Data Shapley value does not ensure truthful reporting and leads to inefficient outcomes, and that Myerson payments are the minimal truthful mechanism. This is a third way contribution fails as a basis for compensation: even with no noise and no manipulation of the data, an accurate value is not an incentive-compatible price.

The upshot for contribution is that coarsening addresses one of four problems. It reduces noise. It does not remove valuator discretion, it does not make a rule replication-robust, and it does not make a value a truthful price. Each of those has a known partial fix in the economics and mechanism-design literature (fix the utility in advance; use replication-robust rules; use truthful payment rules). None of the fixes is granularity.

6. Inclusion: power grows with the unit, soundness does not

6.1 Single-sequence membership inference does not work on pretrained models

Duan et al. (2024) evaluated membership attacks against Pythia models from 160M to 12B parameters on the Pile and found AUC below 0.6 in most domains, with the main exception of GitHub at about 0.70, which they suspect reflects non-members that overlap heavily with members. Earlier reports of AUC above 0.7 came from benchmarks whose non-members were drawn from a later period. Shi et al. (2024) introduced Min-K% Prob and the WikiMIA benchmark built on exactly that temporal split. Maini et al. (2024) found that Min-K% Prob reached an AUC near 0.7 on WikiMIA but near 0.5 on same-period Wikipedia splits of the Pile, and near 0.4 on arXiv and OpenWebText2, which they read as false positives. Das et al. (2024) built "blind" classifiers that look only at the samples, not the model, and found that they beat published attacks on eight membership benchmarks. Meeus et al. (2025) quantified distribution shift in six post-hoc benchmarks with a bag-of-words classifier and found strong shift in all of them. At the unit of a sentence or paragraph, the published evidence that inclusion can be detected in a pretrained model is mostly evidence that two sets of text differ.

6.2 Aggregation restores power

Maini et al. (2024) reframe the task as dataset inference: a rightsholder with a suspect set and a private validation set from the same distribution combines several weak membership features, learns which carry signal, and runs a t-test across many sequences. On Pile subsets with Pythia, this separated training from validation splits with p-values below 0.1 and produced p-values above 0.5 when two validation subsets were compared, using about 1000 sequences. Puerto et al. (2025) measure the same idea on a continuous scale. For Pythia 6.9B, AUROC on arXiv rose from 0.501 for sentences to 0.528 for paragraphs, 0.697 for documents and 0.943 for collections of 500 documents; on Wikipedia from 0.507 to 0.844; on HackerNews from 0.500 to 0.709, but with a reported spread of ± 0.340 (Figure 1a). At 500 documents only arXiv (0.943 ± 0.025) and Wikipedia (0.844 ± 0.132) are clearly above chance. Collections of only dozens of documents stayed near chance. The aggregation gain depends on the paragraph-level signal being above chance to begin with.

This is the strongest published support for the tidy hypothesis. A rightsholder with a large body of work is in a much better position than one with a single article. It also matches how disputes arise: Maini et al. note that copyright claims typically concern many documents by one author, not one paragraph.

6.3 But the null is not identified

Zhang et al. (2025a) argue that none of this yields a proof. A training-data proof must bound the probability that the test fires when the model was not trained on the work. That requires sampling from the distribution of models trained without the work, which cannot be done for a production model whose training set is unknown and whose training cannot be repeated. Dataset inference substitutes a held-out validation set for the missing counterfactual. Zhang et al.'s objection is that the substitution is unsound because publishing a work changes other documents: quotations, summaries, discussion and republished copies enter the web, the model may train on those, and the published work then looks more familiar to the model than the unpublished drafts even if the work itself was never trained on. Fan et al.'s (2025b) finding that much opted-out New York Times text remained in a compliant corpus as republished copies is a measured instance of exactly this channel.

Aggregation does not repair this. If the null is mis-specified, averaging over 500 documents makes the test more confident in the wrong comparison. Puerto et al.'s splits are clean train and held-out splits of the Pile, so their result establishes power under a correctly specified null; it does not establish what the false-positive rate would be for a published work that was excluded but quoted. Nobody has measured that rate. Section 11 proposes how.

6.4 Ex-ante randomisation restores soundness

The same authors who argue that post-hoc membership inference cannot prove inclusion identify what can. If the rightsholder inserts a randomly drawn mark before release, the null distribution is the distribution of the test statistic over marks that were not drawn, and it can be sampled exactly (Zhang et al., 2025a). Wei et al. (2024) implement this as data watermarks - random character sequences or random Unicode look-alike substitutions applied across a rightsholder's documents - and frame detection as a hypothesis test with a guaranteed false-detection rate. Detection strengthens with the number of watermarked documents and the watermark's length, weakens as the training set grows, and holds up if the model grows with it. Treating SHA hashes in StackExchange as natural watermarks, they detected hashes in BLOOM-176B robustly once a hash occurred more than about 90 times, with some occurring only 12 times detected at ten sigma. Meeus et al. (2024) insert fictitious "trap" sequences into books in a randomised design and train a 1.3B model from scratch. Medium-length traps repeated 100 times were not detectable; 100-token traps repeated 1000 times reached an AUC of 0.748. That is a measure of power, not a false-positive bound: Zhang et al. (2025a) note that Meeus et al. "do not consider a rank-based test and do not show how to bound the FPR of this approach". The randomised design would admit such a test, but the published result does not supply one.

Two features of this evidence matter for the agenda. The first is that it is fine-grained and sound at once: a single rightsholder with a single collection gets a bounded false-positive rate. That directly contradicts the claim that robustness lives at coarse units. The second is that it costs power. Marks must be repeated or spread over many documents to be learned, which is aggregation again, and they must be present before the training crawl. A mark cannot be added to work already trained on. Ex-ante design and aggregation are complements: aggregation supplies power, randomisation supplies the null.

Ex-ante marks have their own adversary. A developer who rewrites training text before use - rephrasing, translating or distilling it through another model - may remove a character-level mark and lower memorisation signals on the originals. Li et al. (2026) show that post-hoc detection with original texts becomes brittle when a model is trained on style-transformed surrogates, and propose a method that searches for the transformation and, in their evaluation, restores the detection signal. Wang et al. (2023a) propose watermarking the outputs of a model so that synthetic text carries information about its data providers, which moves the mark to a layer the developer controls. How well watermarks survive deliberate laundering at pretraining scale is not established.

6.5 Architecture can make inclusion exact

A different route avoids inference entirely. Min et al. (2024) train a parametric model, SILO, only on 228B tokens of public-domain and permissively licensed text, and place higher-risk text in a nonparametric datastore queried at inference. The datastore supports attribution at the level of the retrieved sentence and opt-out by deletion, and access to it closed 90 percent of the performance gap to a model trained on the Pile. For the datastore portion, inclusion is a lookup, not an inference, and it is exact at the finest grain. This is robustness by construction at fine granularity, again the opposite of the tidy hypothesis. Its limit is that it covers only the data the developer chose to keep out of the weights.

7. Composition: the coarse quantity is not self-verifying

7.1 A domain-level opt-out does not remove the content

Composition at the level of domains is what the template asks for and what robots.txt controls. Fan et al. (2025b) measure what filtering by it achieves. Applying current robots.txt restrictions retroactively to FineWeb-Edu removed 8 percent of tokens. In the filtered corpus, 0.267 percent of documents still had a 50-gram match to nytimes.com content, against 0.354 percent before filtering; for medicalxpress.com the figures were 0.637 and 0.702 percent, and for stackexchange.com 0.348 and 0.539 percent (Figure 1b). Manual inspection showed that many of the remaining New York Times matches were republished copies on other domains, while the matched medicalxpress content was itself republished from other sources. To build a news-free corpus they excluded news domains, a 4 percent reduction, and then removed segments replicated from news articles on other domains, a further 14 percent. They note that New York Times articles reposted on personal blogs that do not block crawlers remain accessible, and conclude that document-level compliance is hard to establish.

This is a measured failure of the coarse unit, not a failure of the developer. The rightsholder's unit is the work. The opt-out signal's unit is the domain. Works cross domains. A domain-level opt-out honoured perfectly still lets most of the matching text through, and a domain list disclosed perfectly still says little about whether a given work was trained on.

7.2 The signal is honoured unevenly

Kim et al. (2025) observed 130 self-declared bots over 40 days on their institution's sites while varying robots.txt directives. Bots were less likely to comply with stricter directives, and 15 of 47 bots in the endpoint-access and disallow-all experiments did not fetch robots.txt during the experiments, though some of these still complied in part. Liu et al. (2025a) surveyed 203 professional artists and measured crawler behaviour,

finding strong demand for crawler blocking, low awareness and agency in deploying it, and limited efficacy against crawlers that do not respond to robots.txt; network-level blocking by reverse proxies protected better. Longpre et al. (2024a) found robots.txt and terms of service in frequent disagreement. The coarse signal is a request, and its observance is itself something a disclosure would have to report.

7.3 The disclosure is a self-report

The first public summaries under the template varied widely. Blankvoort et al. (2026) found five as of 12 January 2026 and scored them for transparency and usefulness: Apertus 92.9 and 97.1 percent, SmolLM 82.5 and 86.0, Bria 87.0 and 94.7, Bielik 88.0 and 71.1, and a Phi-4 document they could not confirm was intended as a public summary 33.3 and 24.5. Their rubric includes correctness metrics - is the domain list accurate, were all listed crawlers used - but the assessors could judge correctness only against what the provider published. A third party has no instrument for checking a domain list against the weights.

That is the central weakness of composition as a provenance quantity. It is coarse, and against the contributor and the valuator it is robust - neither can change what a corpus contained. Against the developer, who writes it, it is exactly as robust as the developer's honesty, plus whatever an audit of the training pipeline can establish. The AI Office's supervisory powers are such an audit. Whether model-side tests could corroborate a domain list - for example, collection-level membership inference over samples from the listed and unlisted domains - has not been studied. Section 11 proposes the test.

7.4 Synthetic data moves the question

The template asks separately about synthetic data and the models that produced it (Blankvoort et al., 2026). A corpus that is partly generated by a teacher model inherits whatever the teacher was trained on, and a disclosure that names the teacher does not disclose the teacher's sources. Li et al. (2026) show that training on transformed surrogates makes original-text detection brittle. Coarse disclosure of synthetic data is therefore robust in form - the developer reports a model name - and silent in substance about the provenance of the content. No published work measures how much of a teacher's training text can be recovered, or detected, through a student trained on its outputs.

8. Exposure: the fine-grained quantity with a sound test

8.1 Extraction bypasses the counterfactual

Exposure has the cleanest test in this literature. If a model can be made to emit a long passage verbatim, the probability that it would do so without having trained on the passage, or on a copy, is taken to be negligible, and

no null distribution needs to be sampled. Zhang et al. (2025a) point out that this rests on heuristics - the emitted string should be "long", of "high entropy", and produced from "short" or "random looking" prompts - and offer a way to "(partially) formalize these intuitions" by casting extraction as a ranking test. Carlini et al. (2021) extracted hundreds of verbatim sequences from GPT-2, some of which appeared in only one training document. Nasr et al. (2023) extracted gigabytes from open models and found that a divergence attack made ChatGPT emit training data at about 150 times its normal rate. Cooper et al. (2025) measured memorisation of 200 books in 14 open-weight models: most models did not memorise most books, but Llama 3.1 70B memorised some books, including the first Harry Potter novel, so completely that the whole text can be extracted nearly verbatim from its opening words.

Exposure is fine-grained - it attaches to a passage - and sound against the developer. It is also limited in exactly the way the duplication results predict: verbatim reproduction concentrates on heavily duplicated text (Kandpal et al., 2022; Carlini et al., 2022), so exposure tests favour rightsholders whose work was copied widely, which is the same distortion a memorisation-based payment rule would have.

8.2 Exposure is not contribution, and not infringement

Two cautions keep exposure in its place. First, exposure says nothing about the much larger share of training data that is not reproduced. Wang et al. (2025b) find that a training passage can rank within the top 0.1 percent of contributors to a validation text on a similar topic that shares no tokens with it, and Deng et al. (2023) find that similarity-based proxies miss the broader contribution that drives a music model's utility. A provenance regime built on exposure would pay for regurgitation and nothing else. Second, exposure is not a legal conclusion. Henderson et al. (2023) argue that fair use for foundation models is not guaranteed and depends on outputs and market effects. Vyas et al. (2023) define near access-freeness, a bound on how much a model's output distribution can diverge from that of a model that never saw a protected work, and Elkin-Koren et al. (2023) argue that such stability guarantees are privacy notions that do not map onto what copyright protects. Cooper et al. (2024) argue that unlearning, often invoked as the remedy for exposure, mismatches both the removal of information and the suppression of outputs it is asked to achieve.

8.3 Exact-match tooling makes corpus-side exposure cheap

When the training corpus is available, exposure can be checked from the corpus side exactly. Infini-gram indexes trillions of tokens for arbitrary-length n-gram lookups with millisecond latency (Liu et al., 2024), and OLMoTrace uses it to show verbatim matches between a model's outputs and its multi-trillion-token training data

in seconds (Liu et al., 2025b). For open models with open data, the question "did this passage appear in training, and does the model reproduce it" is a lookup. For closed models it is the extraction problem.

9. The consolidated evidence

Table 1 collects the measured values the argument rests on, one row per cited result, grouped by the quantity each bears on. Figure 1 re-plots two of them. Every value is the source's own; the last column states the reading this paper draws, which is the author's.

Read down the table and three patterns appear. In the contribution rows, the stable results are the coarse one (Koh et al.) and the scale results are the ones that come apart (Chang et al.); every gaming result is at the unit of an example or a client. In the inclusion rows, power arrives with aggregation (Puerto, Maini) and soundness arrives with planning (Wei; Meeus et al. supply a planned design but report power only), and those are different rows. In the composition rows, the coarse quantity leaks at the boundary between the rightsholder's unit and the signal's unit, and its verification depends on the party that produces it.

10. An obligation map and a decision procedure

10.1 Six obligations, four quantities

Table 2 assigns each obligation that has been proposed for training data to the quantity it needs, the coarsest unit at which that quantity still serves the obligation, whether a test with a bounded false-positive rate against the relevant adversary exists, and who can game it. The assignments are the author's; the evidence column points to the sections above.

Three things follow from Table 2. First, only one obligation - per-use compensation by contribution - needs the quantity that is both unstable and gameable from two sides. The others need composition, inclusion or exposure, each of which has a route to robustness that does not run through granularity. Second, the regulator's coarse disclosure serves the first row only, and it serves it as a self-report. It does not by itself serve opt-out enforcement, because opt-out is about a rightsholder's works and the domain is not the rightsholder's unit (Section 7.1). Third, a compensation scheme can avoid contribution altogether by allocating on composition - pro rata by licensed volume, say - at the cost of rewarding volume and inviting replication, a cost the replication-robust payment literature addresses (Ohrimenko et al., 2019; Han et al., 2022).

10.2 The decision procedure

Algorithm 1 states the decision structure the paper analyses: given an obligation, which quantity it needs, which adversary matters, and which instrument can deliver a claim that survives that adversary. It is a

Table 1. Measured values on the four provenance quantities, as reported by each source. "Ratio" is manipulated over unperturbed compensation share.

Quantity	Source	Setting	Unit	Reported value	Reading
Contribution	Wang and Jia (2023)	CIFAR-10, 5 seeds	Example	Spearman ~ 0.001 (LOO), ~ 0.038 (Shapley)	Noise dominates
Contribution	Nguyen et al. (2023)	CNN, MNIST3 / CIFAR-10, SWA + DE-Init	Pair	LOO p = 0.331 / 0.692	Noise dominates
Contribution	Nguyen et al. (2023)	ViT + LoRA vs CNN, MNIST3, SWA + DE-Batch	Pair	LOO p 0.025 to 0.786	Worse with scale
Contribution	Diehl and Wilson (2025)	Credit data, Shapley	Example	Median payout + ~ 5 avg. units	Valuator discretion
Contribution	Diehl and Wilson (2025)	5 datasets	Example	2-5% flip a bottom-10% filter	Valuator discretion
Contribution	Akyurek et al. (2022)	Fact tracing	Example	MRR 77.55 BM25 vs 67.66 TDA	Retrieval beats TDA
Contribution	Chang et al. (2024)	8B, 20M sentences	Example	MRR 0.592 BM25 vs 0.365	Attribution
Contribution	Chang et al. (2024)	8B, tail-patch	Example	+0.90% vs +0.41% (BM25)	Influence
Contribution	Chang et al. (2024)	TRAK, 300 vs 19.6M candidates	Example	MRR 0.401 to 0.001	Fails on full pool
Contribution	Koh et al. (2019)	Logistic regression	Group	Spearman 0.8-1.0	Coarse is stabler
Contribution	Wang et al. (2025a)	MNIST, Digits, CIFAR-10	Example	Ratio 185.2-643.9%	Contributor gaming
Contribution	Wang et al. (2025a)	NanoGPT, 0.13% of data	Example	Ratio 262.9%	Contributor gaming
Contribution	Gao et al. (2026)	FL, 10 clients, FedSV	Client	Rank 10 to 2-7 (of 10)	Contributor gaming
Inclusion	Duan et al. (2024)	Pythia 160M-12B	Sequence	AUC < 0.6 most domains	At chance
Inclusion	Puerto et al. (2025)	Pythia 6.9B, arXiv	Sentence	AUROC 0.501	At chance
Inclusion	Puerto et al. (2025)	Pythia 6.9B, arXiv	500 docs	AUROC 0.943	Power from scale
Inclusion	Maini et al. (2024)	Pythia, Pile subsets	~ 1000 seqs	p < 0.1; no false positives	Power from scale
Inclusion	Meeus et al. (2024)	1.3B, trap sequences	100 tok x 1000	AUC 0.748	Planned mark, power only (AUC)
Inclusion	Wei et al. (2024)	BLOOM-176B hashes	Hash	Robust past ~ 90 copies	Planned mark
Composition	Fan et al. (2025b)	FineWeb-Edu, NYT	Document	0.354% to 0.267%	Opt-out leaks
Composition	Fan et al. (2025b)	News removal	Tokens	4% domains + 14% copies	Opt-out leaks
Composition	Kim et al. (2025)	47 bots, disallow-all	Bot	15 of 47 did not fetch robots.txt	Signal often unread
Composition	Blankvoort et al. (2026)	5 public summaries	Summary	Transparency 33.3-92.9%	Self-report varies
Composition	Longpre et al. (2024a)	C4	Tokens	$\sim 5\%$ robots; 45% ToS	Consent shifting
Exposure	Kandpal et al. (2022)	LM regeneration	Sequence	10 copies $\sim 1000x$ output	Duplication bias
Exposure	Cooper et al. (2025)	200 books, 14 models	Book	Most not memorised	Rare but extreme

Table 2. Obligations mapped to provenance quantities. "Sound" means a test whose false-positive rate can be bounded against the party in the last column.

Obligation	Quantity	Coarsest unit that serves	Sound test?	Who can game it	Evidence
Public summary (Art. 53(1)(d))	Composition	Dataset, domain	No; audit only	Developer	Sec. 2, 7.3
Honour an opt-out	Inclusion	Rightsholder's corpus	Only with prior mark	Developer	Sec. 6, 7.1
Prove use in a dispute	Inclusion	Collection	Mark or extraction	Developer	Sec. 6.3-6.4, 8.1
Output non-reproduction	Exposure	Passage	Yes, extraction	Developer (weakly)	Sec. 8
Licence-fee allocation	Composition	Licensed source	N/A (contract)	Contributor (volume)	Sec. 5.2
Per-use compensation	Contribution	Work or source	No	Contributor and valuator	Sec. 4, 5

procedure for choosing an instrument, not a guarantee; each branch names the evidence it rests on.

Line 13 is the line most likely to be resisted. Dataset inference and collection-level membership inference are powerful, and Puerto et al.'s numbers are impressive. The procedure keeps them, but labels their output as association rather than proof, because Zhang et al.'s

objection has not been answered by a measurement. Line 21 is the line that answers the topic's question directly: when a per-work debt is demanded and the four mitigations above it do not make it stable and ungameable, the honest output is that the evidence does not support one.

Algorithm 1. Choosing a provenance instrument for an obligation

```

Input: obligation O; adversary set A (developer, contributor, valuator)
1  q <- quantity O requires (composition | inclusion | exposure | contribution)
2  if q = composition:
3    report at dataset/domain unit; record crawler compliance and dedup policy
4    if developer in A: require pipeline audit; optionally corroborate with
5      collection-level inference on listed vs unlisted sources (Sec. 7.3, M4)
6    if O allocates money: fix the unit and rule in the contract; use a
7      replication-robust rule (Sec. 5.2)
8  if q = inclusion:
9    if rightsholder marked the work before release:
10     run the mark's hypothesis test; report its false-positive bound (Sec. 6.4)
11   else if long verbatim extraction succeeds: report extraction (Sec. 8.1)
12   else if collection is large: report dataset inference as evidence
13     of association only; do not report it as proof (Sec. 6.2-6.3)
14   else: report "not determinable"
15 if q = exposure: run extraction or corpus lookup; report rate per passage (Sec. 8)
16 if q = contribution:
17   if valuator in A: fix utility, target set and method before data arrive (Sec. 5.3)
18   aggregate to the coarsest unit the obligation allows (Sec. 4.4)
19   report variance across retraining or ensembles, not a point value (Sec. 4.1)
20   if contributor in A: pay with a truthful, replication-robust rule (Sec. 5.2, 5.4)
21   if O needs a per-work debt that survives 17-20: report "not supported"
Output: instrument, unit, and the adversaries against which the claim is sound

```

11. Measurements that would settle the open part

Seven measurements, each feasible with published methods and models, would move the questions above from argument to evidence.

M1. The false-positive rate of collection-level inference for published-but-excluded works. Train a model from scratch on a corpus from which a set of works has been excluded while quotations, summaries and republished fragments of those works remain, as they do in practice (Fan et al., 2025b). Run dataset inference with unpublished drafts as the validation set, as Maini et al. (2024) propose. The rate at which the test fires on the excluded works is the quantity Zhang et al. (2025a) argue cannot be bounded by the validation-set design, and nobody has measured it. Meeus et al.'s (2024) randomised design shows the experiment can be piggybacked on a 1.3B from-scratch pretraining run.

M2. Stability of source-level contribution across retraining at language-model scale. Koh et al. (2019) establish group-level stability in convex models; Nguyen et al. (2023) establish example-level instability in small deep ones. Compute contribution for sources of increasing size - document, author, domain - across several pretraining seeds, using ensembles (Deng et al., 2024) or partition merging (Na et al., 2024), and report the rank correlation across seeds as a function of source size. The size at which correlation crosses a usable threshold is the coarsest unit at which contribution-based payment is even possible.

M3. Valuator gaming at the source level for a pretraining corpus. Repeat Diehl and Wilson's (2025) range analysis with domain-level or source-level values for a language model, over a candidate set of utilities a reasonable valuator could defend (different target benchmarks, loss versus accuracy, different

deduplication rules). The spread of each source's share across that set is the discretion a payer would hold.

M4. Corroborating a domain list from the model. For an open model with open data, draw samples from domains listed in its composition and from domains known to be absent, and test whether collection-level inference separates them. Then repeat with a model whose list is deliberately perturbed. If the test can detect a misreported top domain, composition becomes partially verifiable against the developer; if not, the public summary remains an audit-only instrument.

M5. Republication leakage across rightsholder types. Fan et al. (2025b) measured leakage for three domains. Measure, across many opted-out domains, the fraction of each domain's text that reaches a robots.txt-compliant corpus through other domains. This is the quantity that decides whether domain-level opt-outs approximate work-level opt-outs for a given class of rightsholder, and it is likely to differ sharply between news, which is widely republished, and long-form books, which are not.

M6. A contributor attack on language-model attribution. Wang et al. (2025a) attacked NanoGPT on Shakespeare. Attempt the Outlier Attack against an influence-based valuation of a pretrained model at the scale of Choe et al. (2024) or Chang et al. (2024), with the adversary controlling a small fraction of fine-tuning or continued-pretraining data. Either the noise floor at scale hides the attack's effect - which would also mean the honest scores are noise - or the attack works and per-work compensation by influence is broken in the setting where it was proposed.

M7. Watermark and trap survival under laundering. Insert random watermarks (Wei et al., 2024) or traps (Meeus et al., 2024) into a collection, then train on rephrased, translated or distilled versions of it, and measure detection power against the transformation applied. Li et al. (2026) show original-text detection

degrades under such transformations; whether ex-ante marks degrade as badly is the question that decides whether planned provenance survives a developer who wants to evade it.

M1, M4 and M7 test robustness against the developer; M2 tests noise; M3 tests the valuator; M6 tests the contributor; M5 tests the coarse unit itself. Together they cover each cell of the quantity-by-adversary grid that Section 3 identifies as open.

12. Objections

"Granularity still matters; the paper concedes as much." It does, and the paper's claim is narrower than its title may suggest. Granularity buys power against noise for inclusion and stability for contribution, and Figure 1a and the Koh et al. result are large effects. The claim is that granularity does not buy robustness against the developer, who controls whether a null can be sampled, or against the valuator, who controls the utility, and buys only part of the robustness against the contributor. A provenance regime that coarsens and stops there will be powerful and gameable.

"Ex-ante marks are impractical for the existing web." Agreed, and Algorithm 1 handles it: for work already released unmarked, the procedure falls back to extraction or to association-only reporting. The point is that the regime that makes sound fine-grained claims possible exists, has been demonstrated at 176B scale with natural marks (Wei et al., 2024), and costs the rightsholder a decision before publication. Policy that wants enforceable opt-outs for future work could ask for it.

"Influence at scale is now good enough; the instability results are on toy models." Some of them are. But the scale results do not show stability either: Li et al. (2024) find influence functions poor on language models, and Nguyen et al. (2023) find instability rising sharply when moving to a large pretrained backbone. The TRAK collapse from 0.401 to 0.001 (Chang et al., 2024) is not among them: it comes from enlarging the candidate pool, not the model, and is traced largely to correctable design choices. DATE-LM's counterfactual setting, where attribution beats retrieval, is the strongest evidence for attribution at scale, and it is evidence about ranking supporting facts, not about the stability of a payment-grade number across retraining (Jiao et al., 2025).

"A regulator never claimed the template would support compensation." Correct, and the paper does not criticise the template for failing at something it did not attempt. The argument concerns the inference, common in commentary, from "coarse disclosure is what we have" to "coarse is what is robust". The template's coarse fields are robust against contributors and valuers by construction, and as robust against the developer as an audit makes them. That is a property of

who writes the disclosure, not of its grain.

"Contributor attacks need repeated contributions and model access." Wang et al.'s (2025a) threat model assumes periodic contribution and either distributional knowledge or black-box queries, which is the setting of a data marketplace with recurring payments - the setting compensation schemes propose. Replication needs neither.

13. What is not known

The false-positive rate of dataset inference or collection-level membership inference for a published work that was excluded from training but whose quotations and copies were not is not known (M1). Whether contribution at the level of a source becomes stable across pretraining seeds, and at what source size, is not known for any language model (M2). How much a valuator's choice of utility moves source-level shares of a pretraining corpus is not known (M3). Whether a public summary's domain list can be corroborated or contradicted from the model is not known (M4). How much opted-out text reaches compliant corpora through republication, beyond three domains, is not known (M5). Whether any contributor attack works against attribution for a pretrained language model is not known (M6). How well ex-ante data watermarks survive rephrasing or distillation at pretraining scale is not known (M7). How much of a teacher model's training text is detectable through a student trained on its outputs is not known.

14. Related work and how this paper differs

Attribution and valuation methods. Surveys organise influence and valuation methods by the counterfactual they estimate and by their assumptions (Hammoudeh and Lowd, 2024; Zhu and Cangelosi, 2025). Attribution for generative image models has its own evaluation literature: Wang et al. (2023b) construct exemplar-influenced images by customisation to test attribution in text-to-image models, and Zheng et al. (2024) find that design choices without theoretical justification outperform theoretically motivated ones for diffusion models, which echoes Bae et al.'s (2022) point that the estimators may be answering a different question from the one posed. This paper uses that literature as evidence about one quantity, contribution, and does not propose a method.

Robustness of attribution. Wang et al. (2025a), Gao et al. (2026) and Zheng et al. (2025) study contributors; Diehl and Wilson (2025) study valuers; Wang and Jia (2023), Nguyen et al. (2023) and Karthikeyan and Sogaard (2021) study noise. The present paper's addition is to place these as three different adversaries against one quantity and to show that coarsening addresses the noise and not the other two.

Membership and dataset inference. Meeus et al. (2025) and Zhang et al. (2025a) are the closest critiques;

Puerto et al. (2025) and Maini et al. (2024) the strongest positive results. This paper's addition is to read the two sides as answering different questions - power and soundness - and to tie that distinction to the granularity claim in the disclosure debate.

Provenance and consent. Longpre et al. (2023, 2024a, 2024b, 2025) document the state of dataset licensing, consent signals and provenance tooling; Kandpal and Raffel (2025) estimate the value of the labour behind training data; Fan et al. (2025b) measure the performance cost and the leakage of compliance; Blankvoort et al. (2026) assess the first public summaries. This paper takes their measurements as inputs and asks which obligation each supports.

Copyright and technical guarantees. Henderson et al. (2023), Vyas et al. (2023), Elkin-Koren et al. (2023), Cooper et al. (2024) and Cooper et al. (2025) address what technical properties bear on copyright. This paper confines itself to exposure as one quantity among four and draws no legal conclusion.

The paper differs from all of these in its unit of analysis. It treats the granularity claim - robust where coarse, gameable where it pays - as a hypothesis, decomposes "provenance" and "gameable" until the hypothesis makes distinct predictions, and checks each prediction against the published measurements.

15. Conclusion

The aggregate-disclosure turn is often read as a concession to what is measurable: per-work attribution is noisy and manipulable, so regulators settled for domain lists. The published evidence supports a different account. Coarsening helps against noise - for detecting inclusion, for stabilising contribution - and the effects are large. But the claims that hold up against a party who would rather they did not are the ones that were planned: a random mark placed before release, a utility fixed before the data arrive, a payment rule designed to be truthful and replication-robust, an architecture that keeps risky data in a datastore where inclusion is a lookup. Those can be fine-grained. The claims that fail are the ones whose null cannot be sampled or whose measurement is chosen by an interested party, and those fail at every grain.

Per-work compensation by measured contribution is the one proposal that needs a quantity which is unstable, contributor-gameable and valuator-gameable together. If it is to be built, it should be built on the mitigations in Algorithm 1, lines 17 to 20, and tested by M2, M3 and M6 before anyone is paid by it. Opt-out enforcement and dispute resolution do not need contribution, and have sound instruments if rightsholders mark their work in advance. Public summaries serve the purpose they were designed for, and are exactly as reliable as the audit behind them.

References

- Agarwal, A., Dahleh, M., & Sarkar, T. (2019). A Marketplace for Data: An Algorithmic Solution. arXiv:1805.08125. doi:10.48550/arXiv.1805.08125.
- Akyurek, E., Bolukbasi, T., Liu, F., Xiong, B., Tenney, I., Andreas, J., & Guu, K. (2022). Towards Tracing Factual Knowledge in Language Models Back to the Training Data. Findings of EMNLP 2022. arXiv:2205.11482. doi:10.48550/arXiv.2205.11482.
- Bae, J., Ng, N., Lo, A., Ghassemi, M., & Grosse, R. (2022). If Influence Functions are the Answer, Then What is the Question? arXiv:2209.05364. doi:10.48550/arXiv.2209.05364.
- Basu, S., Pope, P., & Feizi, S. (2021). Influence Functions in Deep Learning Are Fragile. ICLR 2021. arXiv:2006.14651. doi:10.48550/arXiv.2006.14651.
- Blankvoort, D. A. H., Pandit, H. J., & Gahntz, M. (2026). Quality Assessment of Public Summary of Training Content for GPAI models required by AI Act Article 53(1)(d). arXiv:2603.13270. doi:10.48550/arXiv.2603.13270.
- Bommasani, R., Klyman, K., Longpre, S., Kapoor, S., Maslej, N., Xiong, B., Zhang, D., & Liang, P. (2023). The Foundation Model Transparency Index. arXiv:2310.12941. doi:10.48550/arXiv.2310.12941.
- Bommasani, R., Klyman, K., Kapoor, S., Longpre, S., Xiong, B., Maslej, N., & Liang, P. (2024). The 2024 Foundation Model Transparency Index. Transactions on Machine Learning Research. arXiv:2407.12929. doi:10.48550/arXiv.2407.12929.
- Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., Oprea, A., & Raffel, C. (2021). Extracting Training Data from Large Language Models. arXiv:2012.07805. doi:10.48550/arXiv.2012.07805.
- Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., & Zhang, C. (2022). Quantifying Memorization Across Neural Language Models. arXiv:2202.07646. doi:10.48550/arXiv.2202.07646.
- Chang, T. A., Rajagopal, D., Bolukbasi, T., Dixon, L., & Tenney, I. (2024). Scalable Influence and Fact Tracing for Large Language Model Pretraining. arXiv:2410.17413. doi:10.48550/arXiv.2410.17413.
- Choe, S. K., Ahn, H., Bae, J., Zhao, K., Kang, M., Chung, Y., Pratapa, A., Neiswanger, W., Strubell, E., Mitamura, T., Schneider, J., Hovy, E., Grosse, R., & Xing, E. (2024). What is Your Data Worth to GPT? LLM-Scale Data Valuation with Influence Functions. arXiv:2405.13954. doi:10.48550/arXiv.2405.13954.
- Cooper, A. F., Choquette-Choo, C. A., Bogen, M., Klyman, K., Jagielski, M., Filippova, K., Liu, K., et al. (2024). Machine Unlearning Doesn't Do What You Think: Lessons for Generative AI Policy and Research. arXiv:2412.06966. doi:10.48550/arXiv.2412.06966.
- Cooper, A. F., Lemley, M. A., Casasola, A., Ahmed, A., Gokaslan, A., Cyphert, A. B., De Sa, C., Ho, D. E., & Liang, P. (2025). Extracting memorized pieces of (copyrighted) books from open-weight language models. COLM 2026. arXiv:2505.12546. doi:10.48550/arXiv.2505.12546.
- Das, D., Zhang, J., & Tramer, F. (2024). Blind Baselines Beat Membership Inference Attacks for Foundation Models. arXiv:2406.16201. doi:10.48550/arXiv.2406.16201.
- Deng, J., Jiang, X., Zhang, S., Zhang, S., Lakkaraju, H., Gao, R., Donahue, C., & Ma, J. W. (2023). Computational Copyright: Towards A Royalty Model for Music Generative AI. arXiv:2312.06646. doi:10.48550/arXiv.2312.06646.
- Deng, J., Li, T.-W., Zhang, S., & Ma, J. (2024). Efficient Ensembles Improve Training Data Attribution. arXiv:2405.17293. doi:10.48550/arXiv.2405.17293.
- Diehl, H., & Wilson, A. C. (2025). Semivalue-based data valuation is arbitrary and gameable. arXiv:2506.12619. doi:10.48550/arXiv.2506.12619.
- Duan, M., Suri, A., Mireshghallah, N., Min, S., Shi, W., Zettlemoyer, L., Tsvetkov, Y., Choi, Y., Evans, D., & Hajishirzi, H. (2024). Do

- Membership Inference Attacks Work on Large Language Models? COLM 2024. arXiv:2402.07841. doi:10.48550/arXiv.2402.07841.
- Elazar, Y., Bhagia, A., Magnusson, I., Ravichander, A., Schwenk, D., Suhr, A., Walsh, P., Groeneveld, D., Soldaini, L., Singh, S., Hajishirzi, H., Smith, N. A., & Dodge, J. (2024). What's In My Big Data? ICLR 2024. arXiv:2310.20707. doi:10.48550/arXiv.2310.20707.
- Elkin-Koren, N., Hacohen, U., Livni, R., & Moran, S. (2023). Can Copyright be Reduced to Privacy? arXiv:2305.14822. doi:10.48550/arXiv.2305.14822.
- Epifano, J. R., Ramachandran, R. P., Masino, A. J., & Rasool, G. (2023). Revisiting the Fragility of Influence Functions. *Neural Networks* 162, 581-588. arXiv:2303.12922. doi:10.48550/arXiv.2303.12922.
- Fan, D., Rotello, T. J., & Karimireddy, S. P. (2025a). Do Data Valuations Make Good Data Prices? arXiv:2504.05563. doi:10.48550/arXiv.2504.05563.
- Fan, D., Sabolcec, V., Ansaripour, M., Tarun, A. K., Jaggi, M., Bosselut, A., & Schlag, I. (2025b). Can Performant LLMs Be Ethical? Quantifying the Impact of Web Crawling Opt-Outs. COLM 2025. arXiv:2504.06219. doi:10.48550/arXiv.2504.06219.
- Gao, X., Hui, B., Sun, M.-T., & Ku, W.-S. (2026). On the Fragility of Data Attribution When Learning Is Distributed. arXiv:2605.15520. doi:10.48550/arXiv.2605.15520.
- Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daume, H., & Crawford, K. (2018). Datasheets for Datasets. arXiv:1803.09010. doi:10.48550/arXiv.1803.09010.
- Ghorbani, A., & Zou, J. (2019). Data Shapley: Equitable Valuation of Data for Machine Learning. arXiv:1904.02868. doi:10.48550/arXiv.1904.02868.
- Grosse, R., Bae, J., Anil, C., Elhage, N., Tamkin, A., Tajdini, A., Steiner, B., Li, D., Durmus, E., Perez, E., Hubinger, E., Lukosiute, K., Nguyen, K., Joseph, N., McCandlish, S., Kaplan, J., & Bowman, S. R. (2023). Studying Large Language Model Generalization with Influence Functions. arXiv:2308.03296. doi:10.48550/arXiv.2308.03296.
- Hammoudeh, Z., & Lowd, D. (2024). Training Data Influence Analysis and Estimation: A Survey. *Machine Learning*. arXiv:2212.04612. doi:10.48550/arXiv.2212.04612.
- Han, D., Wooldridge, M., Rogers, A., Ohrimenko, O., & Tschitschek, S. (2022). Replication-Robust Payoff Allocation for Machine Learning Data Markets. *IEEE Transactions on Artificial Intelligence*. arXiv:2006.14583. doi:10.48550/arXiv.2006.14583.
- Henderson, P., Li, X., Jurafsky, D., Hashimoto, T., Lemley, M. A., & Liang, P. (2023). Foundation Models and Fair Use. arXiv:2303.15715. doi:10.48550/arXiv.2303.15715.
- Hu, Y., Hu, P., Zhao, H., & Ma, J. W. (2024). Most Influential Subset Selection: Challenges, Promises, and Beyond. *NeurIPS 2024*. arXiv:2409.18153. doi:10.48550/arXiv.2409.18153.
- Ilyas, A., Park, S. M., Engstrom, L., Leclerc, G., & Madry, A. (2022). Datamodels: Predicting Predictions from Training Data. arXiv:2202.00622. doi:10.48550/arXiv.2202.00622.
- Jia, R., Dao, D., Wang, B., Hubis, F. A., Hynes, N., Gurel, N. M., Li, B., Zhang, C., Song, D., & Spanos, C. (2019). Towards Efficient Data Valuation Based on the Shapley Value. arXiv:1902.10275. doi:10.48550/arXiv.1902.10275.
- Jiao, C., Pan, Y., Xiao, E., Sheng, D., Jain, N., Zhao, H., Dasgupta, I., Ma, J. W., & Xiong, C. (2025). DATE-LM: Benchmarking Data Attribution Evaluation for Large Language Models. *NeurIPS 2025 Datasets and Benchmarks Track*. arXiv:2507.09424. doi:10.48550/arXiv.2507.09424.
- Kandpal, N., & Raffel, C. (2025). Position: The Most Expensive Part of an LLM should be its Training Data. arXiv:2504.12427. doi:10.48550/arXiv.2504.12427.
- Kandpal, N., Wallace, E., & Raffel, C. (2022). Deduplicating Training Data Mitigates Privacy Risks in Language Models. arXiv:2202.06539. doi:10.48550/arXiv.2202.06539.
- Karthikeyan, K., & Sogaard, A. (2021). Revisiting Methods for Finding Influential Examples. arXiv:2111.04683. doi:10.48550/arXiv.2111.04683.
- Kim, T., Bock, K., Luo, C., Liswood, A., Poroslay, C., & Wenger, E. (2025). Scrapers selectively respect robots.txt directives: evidence from a large-scale empirical study. arXiv:2505.21733. doi:10.48550/arXiv.2505.21733.
- Koh, P. W., & Liang, P. (2017). Understanding Black-box Predictions via Influence Functions. arXiv:1703.04730. doi:10.48550/arXiv.1703.04730.
- Koh, P. W., Ang, K.-S., Teo, H. H. K., & Liang, P. (2019). On the Accuracy of Influence Functions for Measuring Group Effects. arXiv:1905.13289. doi:10.48550/arXiv.1905.13289.
- Kwon, Y., & Zou, J. (2022). Beta Shapley: a Unified and Noise-reduced Data Valuation Framework for Machine Learning. arXiv:2110.14049. doi:10.48550/arXiv.2110.14049.
- Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., & Carlini, N. (2022). Deduplicating Training Data Makes Language Models Better. arXiv:2107.06499. doi:10.48550/arXiv.2107.06499.
- Li, M., Ye, Z., Li, S., & Liu, F. (2026). Combating Data Laundering in LLM Training. arXiv:2604.01904. doi:10.48550/arXiv.2604.01904.
- Li, Z., Zhao, W., Li, Y., & Sun, J. (2024). Do Influence Functions Work on Large Language Models? arXiv:2409.19998. doi:10.48550/arXiv.2409.19998.
- Liu, E., Luo, E., Shan, S., Voelker, G. M., Zhao, B. Y., & Savage, S. (2025a). Somesite I Used To Crawl: Awareness, Agency and Efficacy in Protecting Content Creators From AI Crawlers. *IMC 2025*. arXiv:2411.15091. doi:10.48550/arXiv.2411.15091.
- Liu, J., Blanton, T., Elazar, Y., Min, S., Chen, Y., Chheda-Kothary, A., et al. (2025b). OLMoTrace: Tracing Language Model Outputs Back to Trillions of Training Tokens. *ACL 2025 demo track*. arXiv:2504.07096. doi:10.48550/arXiv.2504.07096.
- Liu, J., Min, S., Zettlemoyer, L., Choi, Y., & Hajishirzi, H. (2024). Infini-gram: Scaling Unbounded n-gram Language Models to a Trillion Tokens. arXiv:2401.17377. doi:10.48550/arXiv.2401.17377.
- Longpre, S., Mahari, R., Chen, A., Obeng-Marnu, N., Sileo, D., Brannon, W., Muennighoff, N., et al. (2023). The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI. arXiv:2310.16787. doi:10.48550/arXiv.2310.16787.
- Longpre, S., Mahari, R., Lee, A., Lund, C., Oderinwale, H., Brannon, W., Saxena, N., et al. (2024a). Consent in Crisis: The Rapid Decline of the AI Data Commons. arXiv:2407.14933. doi:10.48550/arXiv.2407.14933.
- Longpre, S., Mahari, R., Obeng-Marnu, N., Brannon, W., South, T., Gero, K., Pentland, S., & Kabbara, J. (2024b). Data Authenticity, Consent, & Provenance for AI are all broken: what will it take to fix them? *ICML 2024*. arXiv:2404.12691. doi:10.48550/arXiv.2404.12691.
- Longpre, S., Singh, N., Cherep, M., Tiwary, K., Materzynska, J., Brannon, W., Mahari, R., et al. (2025). Bridging the Data Provenance Gap Across Text, Speech and Video. *ICLR 2025*. arXiv:2412.17847. doi:10.48550/arXiv.2412.17847.
- Maini, P., Jia, H., Papernot, N., & Dziedzic, A. (2024). LLM Dataset Inference: Did you train on my dataset? arXiv:2406.06443. doi:10.48550/arXiv.2406.06443.
- Meeus, M., Shilov, I., Faysse, M., & de Montjoye, Y.-A. (2024). Copyright Traps for Large Language Models. *ICML 2024*. arXiv:2402.09363. doi:10.48550/arXiv.2402.09363.
- Meeus, M., Shilov, I., Jain, S., Faysse, M., Rei, M., & de Montjoye, Y.-A. (2025). SoK: Membership Inference Attacks on LLMs are Rushing Nowhere (and How to Fix It). *IEEE SaTML 2025*. arXiv:2406.17975. doi:10.48550/arXiv.2406.17975.
- Min, S., Gururangan, S., Wallace, E., Shi, W., Hajishirzi, H., Smith, N. A., & Zettlemoyer, L. (2024). SILO Language Models: Isolating

- Legal Risk In a Nonparametric Datastore. ICLR 2024. arXiv:2308.04430. doi:10.48550/arXiv.2308.04430.
- Mlodozienec, B., Reid, I., Power, S., Krueger, D., Erdogdu, M., Turner, R. E., & Grosse, R. (2025). Distributional Training Data Attribution: What do Influence Functions Sample? arXiv:2506.12965. doi:10.48550/arXiv.2506.12965.
- Na, C., Magnusson, I., Jha, A. H., Sherborne, T., Strubell, E., Dodge, J., & Dasigi, P. (2024). Scalable Data Ablation Approximations for Language Models through Modular Training and Merging. arXiv:2410.15661. doi:10.48550/arXiv.2410.15661.
- Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., Choquette-Choo, C. A., Wallace, E., Tramer, F., & Lee, K. (2023). Scalable Extraction of Training Data from (Production) Language Models. arXiv:2311.17035. doi:10.48550/arXiv.2311.17035.
- Nguyen, E., Seo, M., & Oh, S. J. (2023). A Bayesian Approach To Analysing Training Data Attribution In Deep Learning. arXiv:2305.19765. doi:10.48550/arXiv.2305.19765.
- Ohrimenko, O., Tople, S., & Tschitschek, S. (2019). Collaborative Machine Learning Markets with Data-Replication-Robust Payments. arXiv:1911.09052. doi:10.48550/arXiv.1911.09052.
- Park, S. M., Georgiev, K., Ilyas, A., Leclerc, G., & Madry, A. (2023). TRAK: Attributing Model Behavior at Scale. arXiv:2303.14186. doi:10.48550/arXiv.2303.14186.
- Pruthi, G., Liu, F., Sundararajan, M., & Kale, S. (2020). Estimating Training Data Influence by Tracing Gradient Descent. arXiv:2002.08484. doi:10.48550/arXiv.2002.08484.
- Puerto, H., Gubri, M., Yun, S., & Oh, S. J. (2025). Scaling Up Membership Inference: When and How Attacks Succeed on Large Language Models. Findings of NAACL 2025. arXiv:2411.00154. doi:10.48550/arXiv.2411.00154.
- Ruis, L., Mozes, M., Bae, J., Kamalakara, S. R., Talupuru, D., Locatelli, A., Kirk, R., Rocktaschel, T., Grefenstette, E., & Bartolo, M. (2025). Procedural Knowledge in Pretraining Drives Reasoning in Large Language Models. ICLR 2025. arXiv:2411.12580. doi:10.48550/arXiv.2411.12580.
- Shi, W., Ajith, A., Xia, M., Huang, Y., Liu, D., Blevins, T., Chen, D., & Zettlemoyer, L. (2024). Detecting Pretraining Data from Large Language Models. arXiv:2310.16789. doi:10.48550/arXiv.2310.16789.
- Vyas, N., Kakade, S., & Barak, B. (2023). On Provable Copyright Protection for Generative Models. ICML 2023. arXiv:2302.10870. doi:10.48550/arXiv.2302.10870.
- Wan, A., Klyman, K., Kapoor, S., Maslej, N., Longpre, S., Xiong, B., Liang, P., & Bommasani, R. (2025). The 2025 Foundation Model Transparency Index. arXiv:2512.10169. doi:10.48550/arXiv.2512.10169.
- Wang, J., Lu, X., Zhao, Z., Dai, Z., Foo, C.-S., Ng, S.-K., & Low, B. K. H. (2023a). Source Attribution for Large Language Model-Generated Data. arXiv:2310.00646. doi:10.48550/arXiv.2310.00646.
- Wang, J. T., & Jia, R. (2023). Data Banzhaf: A Robust Data Valuation Framework for Machine Learning. AISTATS 2023. arXiv:2205.15466. doi:10.48550/arXiv.2205.15466.
- Wang, J. T., Yang, T., Zou, J., Kwon, Y., & Jia, R. (2024a). Rethinking Data Shapley for Data Selection Tasks: Misleads and Merits. ICML 2024. arXiv:2405.03875. doi:10.48550/arXiv.2405.03875.
- Wang, J. T., Deng, Z., Chiba-Okabe, H., Barak, B., & Su, W. J. (2024b). An Economic Solution to Copyright Challenges of Generative AI. arXiv:2404.13964. doi:10.48550/arXiv.2404.13964.
- Wang, J. T., Mittal, P., Song, D., & Jia, R. (2025b). Data Shapley in One Training Run. ICLR 2025. arXiv:2406.11011. doi:10.48550/arXiv.2406.11011.
- Wang, S.-Y., Efros, A. A., Zhu, J.-Y., & Zhang, R. (2023b). Evaluating Data Attribution for Text-to-Image Models. ICCV 2023. arXiv:2306.09345. doi:10.48550/arXiv.2306.09345.
- Wang, X., Hu, P., Deng, J., & Ma, J. W. (2025a). Adversarial Attacks on Data Attribution. ICLR 2025. arXiv:2409.05657. doi:10.48550/arXiv.2409.05657.
- Wei, J. T.-Z., Wang, R. Y., & Jia, R. (2024). Proving membership in LLM pretraining data via data watermarks. Findings of ACL 2024. arXiv:2402.10892. doi:10.48550/arXiv.2402.10892.
- Zhang, J., Das, D., Kamath, G., & Tramer, F. (2025a). Membership Inference Attacks Cannot Prove that a Model Was Trained On Your Data. IEEE SaTML 2025. arXiv:2409.19798. doi:10.48550/arXiv.2409.19798.
- Zhang, L., Jiao, C., Li, B., & Xiong, C. (2025b). Fairshare Data Pricing via Data Valuation for Large Language Models. NeurIPS 2025. arXiv:2502.00198. doi:10.48550/arXiv.2502.00198.
- Zheng, S., Cai, S., Xiao, C., Cao, Y., Qin, J., Yoshikawa, M., & Onizuka, M. (2025). Data Overvaluation Attack and Truthful Data Valuation in Federated Learning. arXiv:2502.00494. doi:10.48550/arXiv.2502.00494.
- Zheng, X., Pang, T., Du, C., Jiang, J., & Lin, M. (2024). Intriguing Properties of Data Attribution on Diffusion Models. ICLR 2024. arXiv:2311.00500. doi:10.48550/arXiv.2311.00500.
- Zhu, H., & Cangelosi, A. (2025). Revisiting Data Attribution for Influence Functions. arXiv:2508.07297. doi:10.48550/arXiv.2508.07297.