



GenAI Systems Should Implement Contribution-Aware Revenue Sharing for Data Providers

The position (P1): Pay data providers in proportion to their **measurable impact on inference-time outputs verifiably and at scale**

Unfortunate current realities:

- (1) Mostly, **no one is paid**; (2) When pay happens, it is **coarse**; (3) No **scalable, transparent** mechanism

Alternative views:

- (1) **Fair use** permits uncompensated use of training data
- (2) **One-time licensing** is sufficient compensation

Call for Actions: Technical level

Call 1: Efficient and scalable **influence estimation**

Estimate top- k influential training samples at a fraction of TracIn [1], DataInfl [2], balancing **attribution accuracy** and **performance**.

Call 2: Influence estimation across full **multi-stage** GenAI **training pipelines**

Future estimation methods should support multi-stage training pipelines, including pretraining, fine-tuning, distillation, and RLHF.

Call 3: Efficient, Portable, and auditable **contracting**

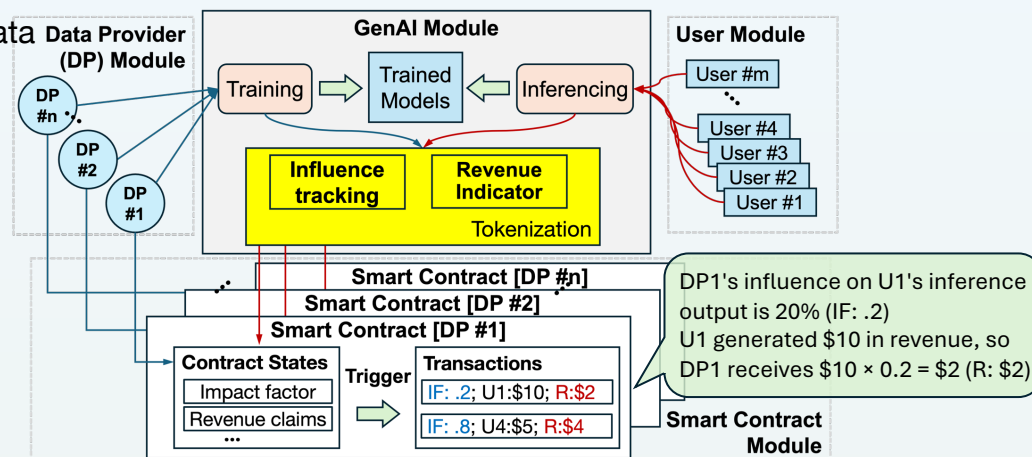
Settle payments on scalable and fault-tolerant contracts, from centralized to fully decentralized:

- Centralized:** GenAI providers centrally manage all contracts
- Semi-decentralized:** GenAI and data providers jointly maintain databases with auditing functions
- Fully decentralized:** all parties coordinate and audit self-sustaining contracts over peer-to-peer systems

P2: Compensation should scale with GenAI-generated revenue, including subscriptions and ads income

P3: Compensation should be auditable and verifiable

P4: Foster community and legal support for adoption



AI Community & Regulatory level

Call 4: Foster **ecosystems** and **community support**

ML community should promote contribution-aware data use by developing shared attribution benchmarks, publishing richer provenance datasets, and establishing standard verification.

Call 5: Build **regulatory** & **institutional frameworks**

Policy and regulation should define rights and obligations for GenAI and data providers in revenue sharing. Public institutions should audit deployments to ensure compliance, deter malicious behavior and preserve fair competition.

[1] Pruthi, G., Liu, F., and Sundararajan, M. *Estimating training data influence by tracing gradient descent*. Advances in Neural Information Processing Systems, 2020.

[2] Kwon, Y., Wu, E., Wu, K., and Zou, J. *DataInfl: Efficiently estimating data influence in lora-tuned LLMs and diffusion models*. International Conference on Learning Representations, 2024