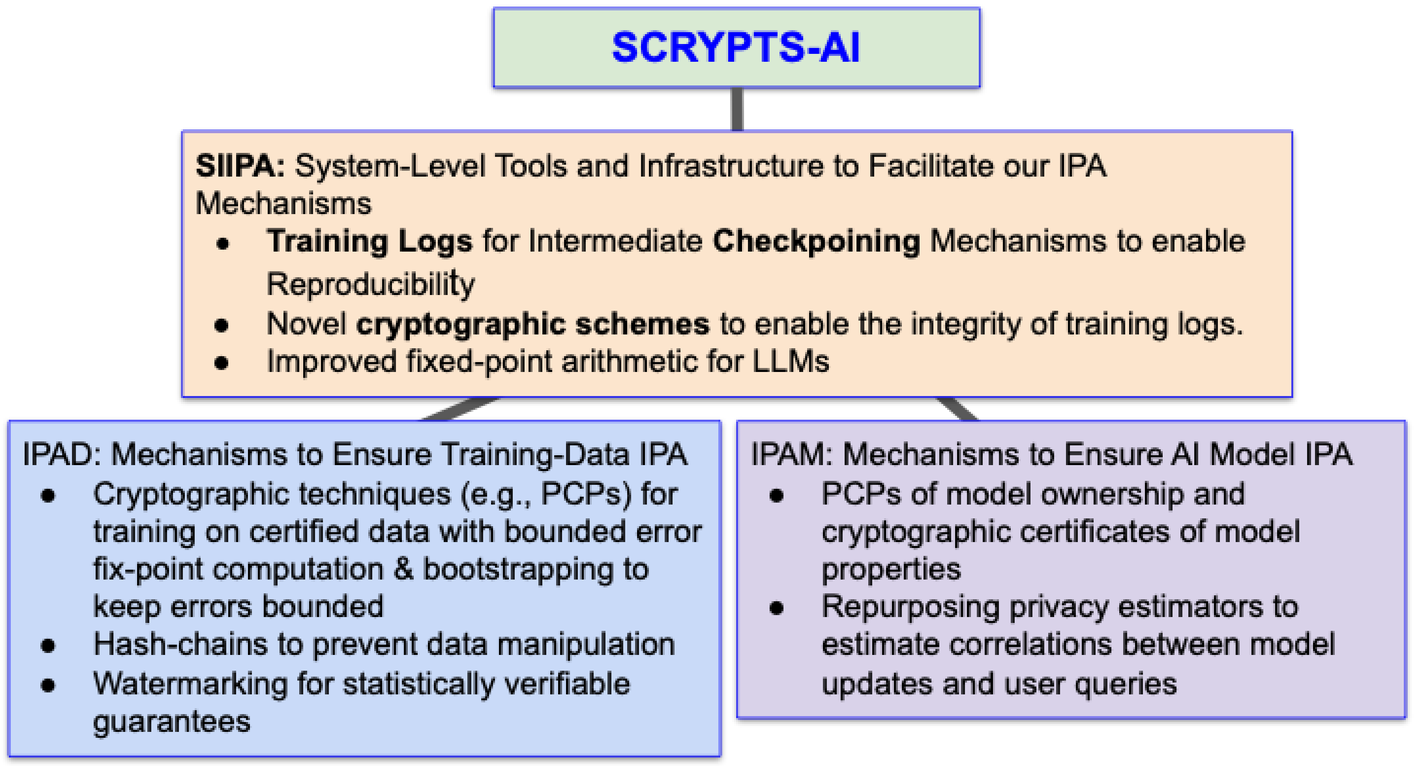
SCRYPTS-AI: Systems and CRYPtographic Tools for Scientific AI

PI: Teodora Baluta (Georgia Tech) | Co-PIs: Vassilis Zikas (Georgia Tech), Rafail Ostrovsky (UCLA)

NSF CICI Award #2531010 • 3-year project, started January 2026

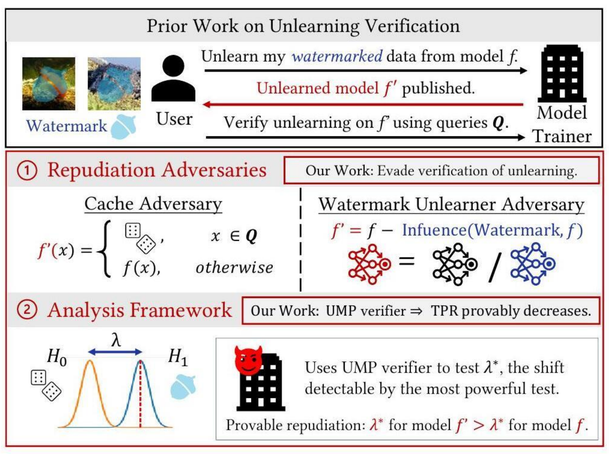
Need for Integrity, Provenance, and Authenticity (IPA) in Scientific AI

- Sensitive scientific data is increasingly used to train AI models, while regulations (GDPR, CCPA) mandate verifiable data removal and provenance.
- Adversarial trainers can misrepresent how data was used or where models originated. Existing tools provide weak, ad-hoc guarantees and rarely consider an adversary that actively evades audits.
- Black-box-only deployments (closed APIs, third-party fine-tunes) make auditing harder still: ground truth is unavailable, training data is secret, and customization recipes are proprietary.
- **Goal: provable, deployable mechanisms that let scientific collaborators trust AI artifacts — from training logs to released models — even when other parties may be adversarial.**

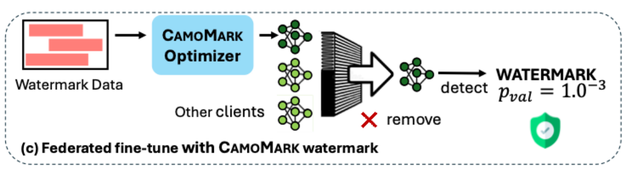


On-going Projects

Statistically Verifiable Watermarks for Data Provenance

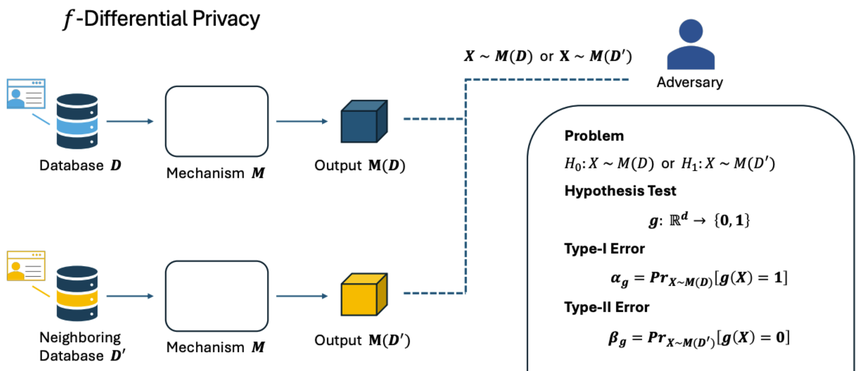


- We study statistical verification under adversarial settings. We propose repudiation attacks on watermarks, lower bounding the adversarial advantage.

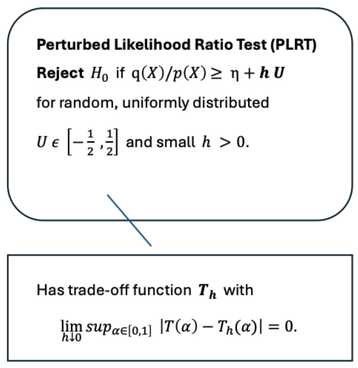


- We propose a data marking algorithm that enables highly detectable provenance under strong robust aggregators in federated learning

Data Auditing via DP Black-box Estimation

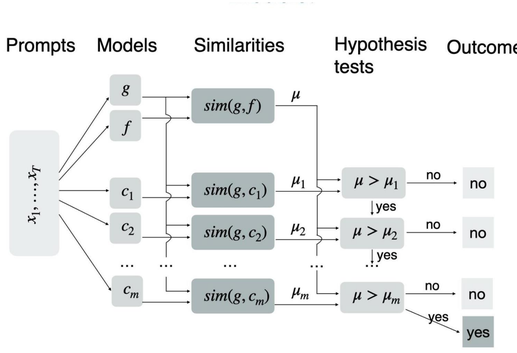


[Usenix'25] *"General-Purpose f-DP Estimation and Auditing in a Black-Box Setting"*

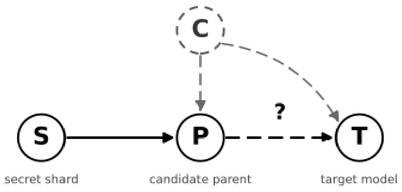


- We propose a general black-box framework that audits the full privacy curve via only black-box access to the mechanisms
- We further propose a sequential, sample-efficient auditing approaches that gives anytime-valid auditing verdict that adaptively stop once sufficient evidence is collected

Model Provenance Testing



[NeurIPS'25] *"Model Provenance Testing for Large Language Models"*



- We propose causal model provenance, together with reducing model provenance to data membership auditing

Benefits to Scientific Cyberinfrastructure

- Design solutions for **collaboration across disciplines;** develop new insights and approaches for distributed, sensitive, private or noisy datasets using AI.
- **Accountability and transparency** in scientific discovery enables trusted collaboration when aggregating data from multiple parties and when training models on such data.
- **Enhance confidence** by (1) checkpointing and reproducibility of AI training logs, (2) data and model provenance tracking and (3) post-training property certification.

Evaluating and Demonstrating IPA

- Public datasets on image classification, biology, math and language (such as Pile-CC)
- Mechanisms, implementations, and benchmarks released on GitHub under an open-source license.

Risks vs. Potential for Advances

- **Challenges:** Scaling cryptographic guarantees to large generative models. Floating-point determinism for reproducible training. Cryptographic overhead at modern AI throughput. Building tools that scientists can actually use.
- **Potential:** Securely sharing distributed, sensitive datasets unlocks high-impact discovery in biomedicine and the natural sciences. A success story for supply-chain trust across the broader AI ecosystem.