Waffle: Watermarking in Federated Learning

Can we *watermark ML models during federated learning*?

The setting:
- De-centralized training in client-server federated learning
- Main goal: Ownership demonstration for the model owner

The problem:
- Clients are potentially malicious
- Model owner has no access to training data

Our solution: Waffle
- Does not require access to training data
- Negligible overhead: performance degradation (-0.17%), computational overhead (+3.2%)
- Robust against watermark removal methods