Profile-then-Simulate:
Can LLMs Faithfully Generate DP Synthetic Data?

Nassima Bouzid

USENIX PEPR '26 · June 2026

PEPR '26

Agenda

01 **The challenge: DP synthesis at scale**

02 **Our approach: Profile-then-Simulate**

03 **Results: What worked and what broke**

04 **Lessons for privacy engineering practice**

01
The Challenge

The high-dimensional DP synthesis problem

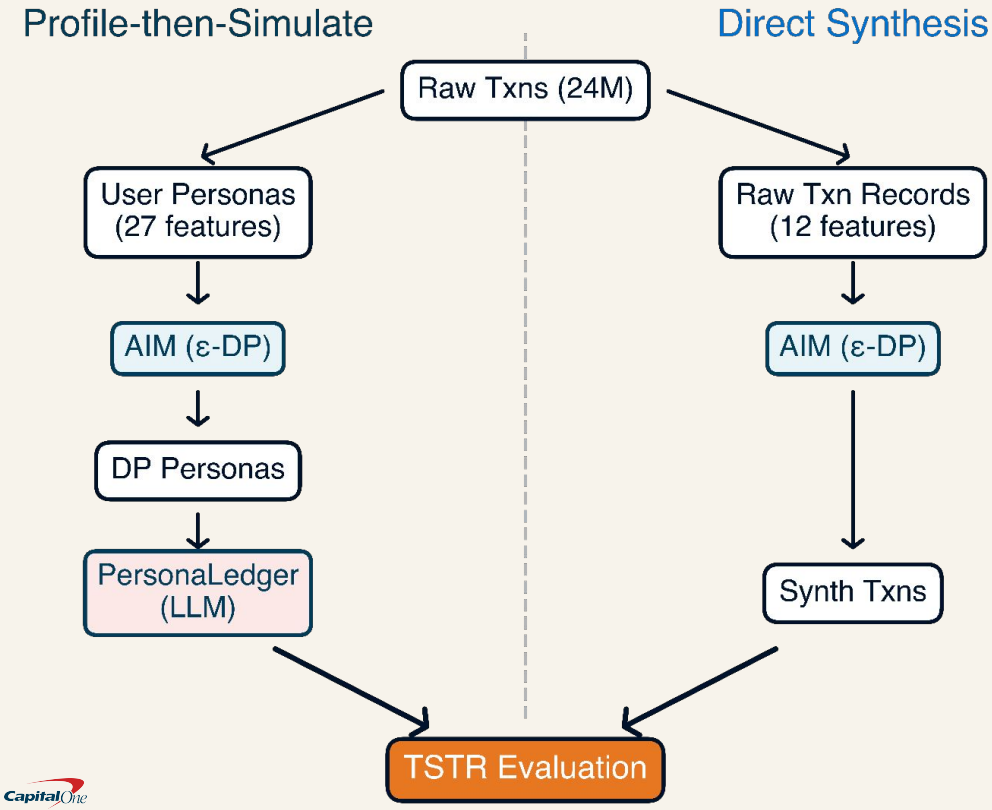
- Fraud detection requires realistic transaction data
- Privacy regulations prevent sharing real customer records
- DP provides formal guarantees but traditional methods struggle as domain size grows
- Utility degrades exponentially with dimensionality

RESEARCH QUESTION

Can LLMs faithfully reproduce statistical distributions from DP-protected user profiles?

02
Our Approach: Profile-then-Simulate

Two-phase architecture: privacy then generation



Privacy budget spent entirely on user-level statistics.

LLM is post-processing → no additional privacy cost.

15,000:1 compression ratio (24M txns → 1.6K personas)

Evaluation setup

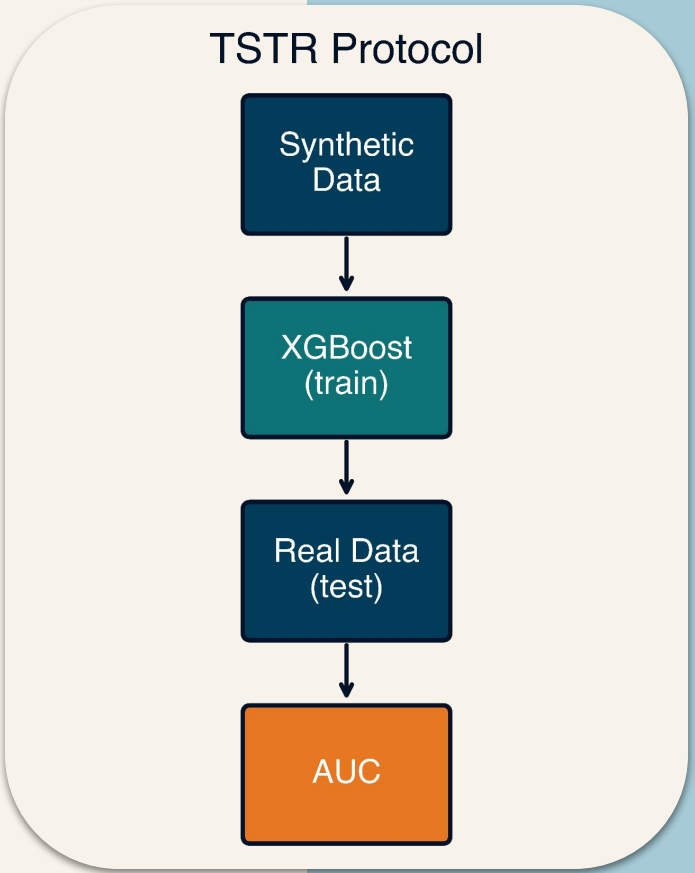
Evaluation protocol: TSTR

- Train XGBoost on synthetic data (reproducible, interpretable baseline)
- Test on 100K real transactions
- Natural fraud rate: 0.13%

Metrics:

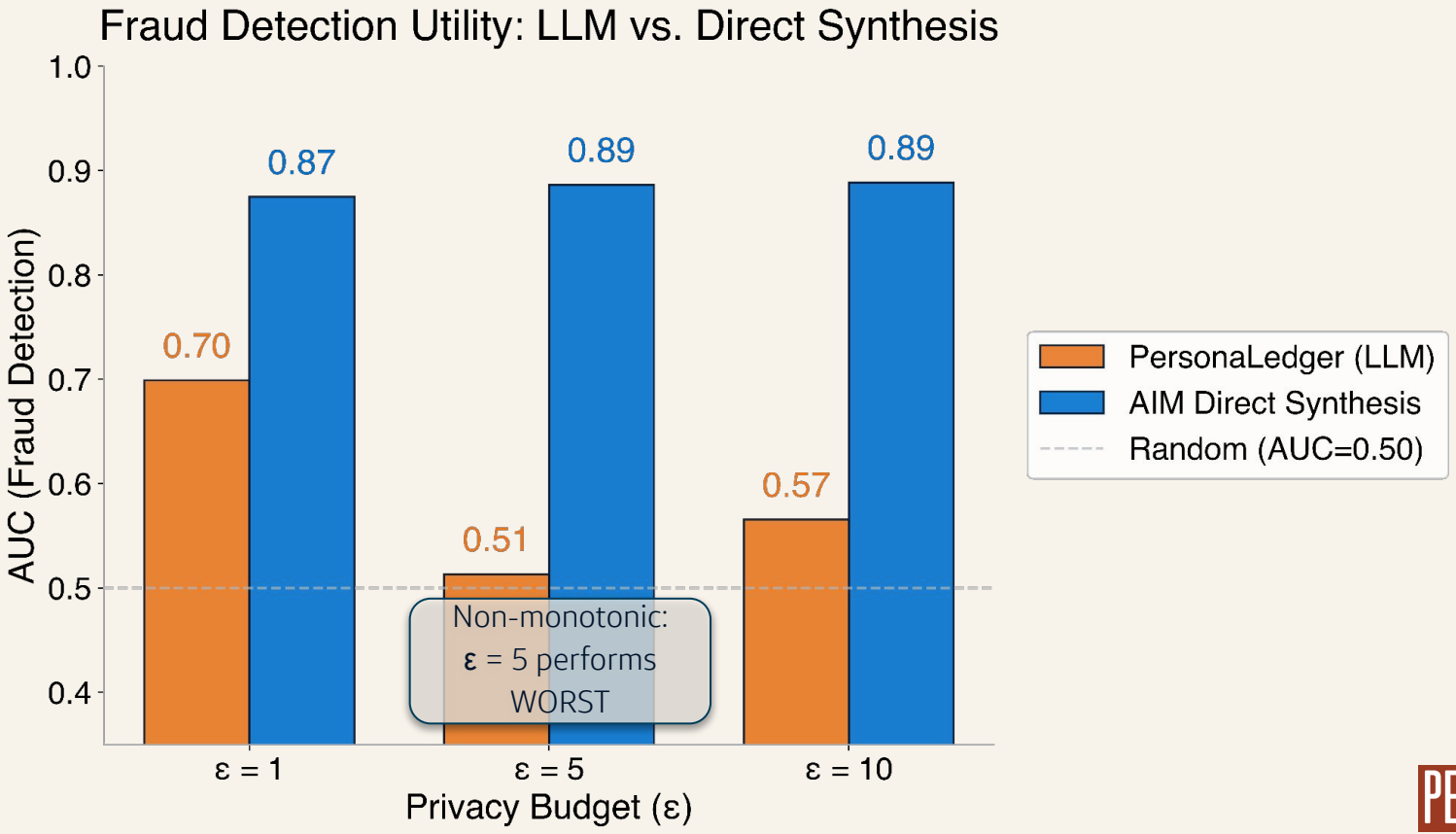
- Utility: AUC (fraud detection)
- Fidelity: Total Variation Distance

Privacy regimes: $\varepsilon \in \{1, 5, 10\}$

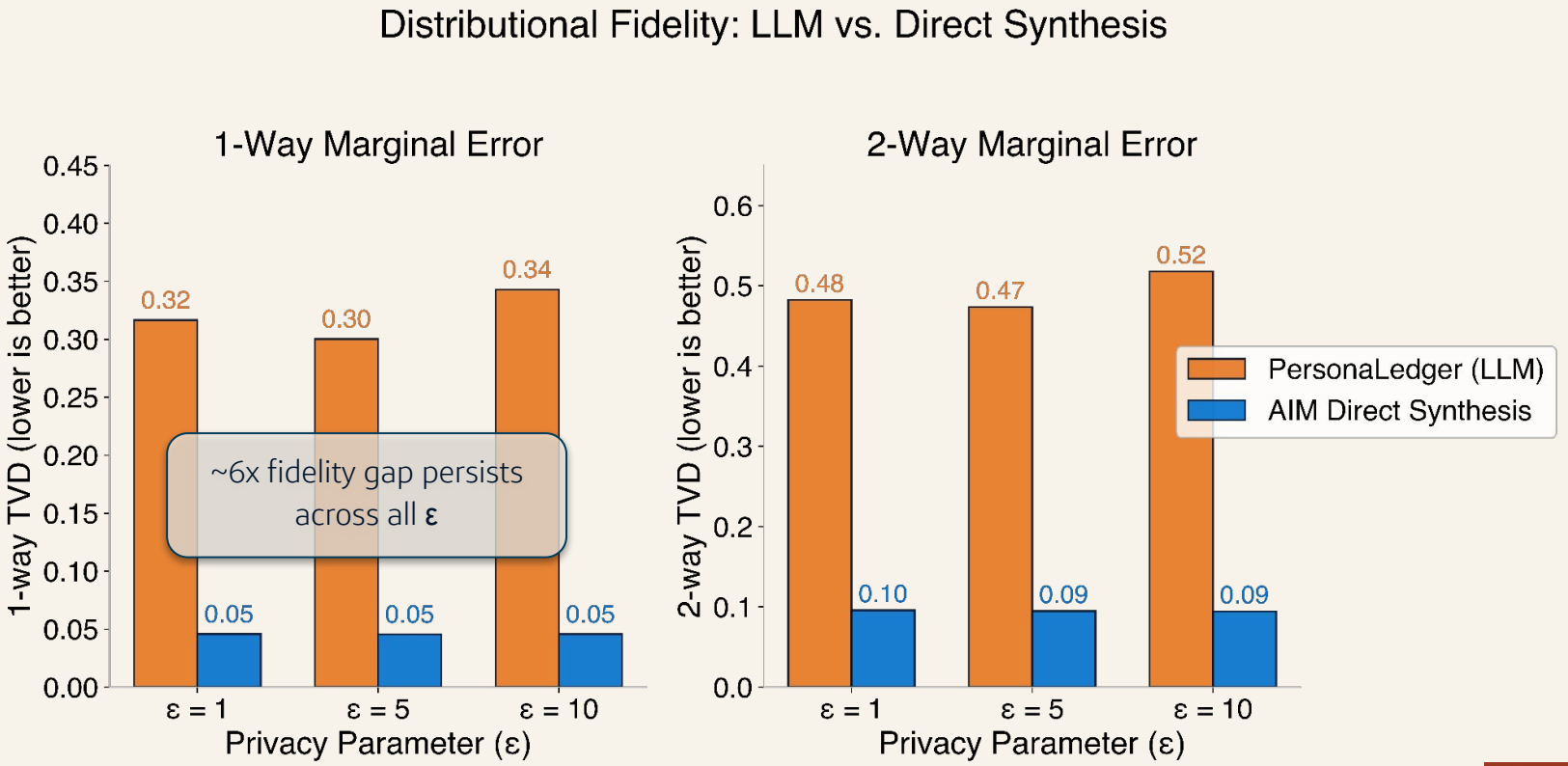


03
Results

Utility gap: AUC comparison across privacy regimes



Fidelity gap: 6x higher TVD in LLM-generated data



Root cause: LLM priors override input statistics

- DEMOGRAPHICS: ~80% retired in generated data vs. ~50% in real data
- TIME-OF-DAY: Transactions cluster 9am-2pm, night/evening patterns missing
- WHAT WORKS: Binary and low-cardinality features (gender, day-of-week, home zip) transfer accurately

Simple distributions transfer.
Complex or non-stereotypical patterns drift.

04
Lessons for Privacy Engineering

The bottleneck isn't privacy,
it's LLM faithfulness.

More epsilon doesn't help when the LLM ignores what you give it.

When to use direct synthesis vs. LLM approach

USE DIRECT SYNTHESIS:

- Schema < ~20 features
- Marginal methods feasible
- High fidelity required

USE LLM APPROACH:

- Profiles 50+ dimensions
- Direct synthesis collapses
- Behavioral realism needed

Measure fidelity, not just utility

- Decent AUC can hide distributional biases
- Feature-level TVD analysis reveals hidden drift
- Evaluate before deploying to production
- TVD on marginals is a minimum bar

ICLR 2026: validated solutions emerging

- **NEMOTRON DISTILLATION (NVIDIA)** — (Ghita, NVIDIA 2026)
  Smaller models have weaker priors, easier to override with input statistics
- **NATWEST SYNTHETIC PERSONAS** — (Iglesias de Oliveira *et al.*, 2026)
  Constrained generation at scale with explicit fidelity constraints on financial data
- **HOT PATE (Private Aggregation)** — (Cohen *et al.*, 2026)
  Aggregate multiple LLM outputs to wash out individual model biases. No extra privacy cost.
- **SecPE (Targeted Noise Allocation)** — (Wang *et al.*, 2026)
  Concentrate DP budget on sensitive attributes, give LLM more signal where it matters

Capital One

Profile-then-Simulate:
LLM faithfulness is the open problem.

Nassima Bouzid

Coauthors: Dehao Yuan, Nam H. Nguyen, Mayana Pereira

PEPR '26

Questions?