Recent Progress on PhraseDP+ Evaluation and Improvements

Meeting Presentation

Tech4HSE Team

2025-01-XX

Recent Progress on PhraseDP+ Evaluation and Improvements

Meeting Presentation

Updated PPI Evaluation • Ablation Study • Medical Improvements • Few-Shot Analysis

1. Updated PPI Evaluation (Fine-Grained)

Problem with Previous Approach

- Binary exact-match detection was **too strict**
- Flat protection curves that didn't reflect privacy-utility trade-off
- Protection rate stayed constant across epsilon values

New Granular Evaluation

Protection Score (PS)

```
PS = 1 - semantic_similarity
```

- Continuous 0-1 scale (not binary)

Semantic Similarity

- Uses **SBERT embeddings** to measure closeness
- More sensitive to subtle perturbations

Updated Plots

- **Line plots:** Protection Score vs Epsilon for each PII type (email, phone, address, name)
- **Radar plots:** Multi-dimensional comparison across epsilon values (1.0, 2.0, 3.0)
- **All 5 mechanisms:** PhraseDP, InferDPT, SANTEXT+, CusText+, CluSanT on same scale
- **Token-level experiments:** N=1000 examples, epsilon coverage: 1.0, 1.5, 2.0, 2.5, 3.0

Figure 1: PII Protection Scores Across Privacy Budgets

PII protection scores across privacy budgets for individual PII types (higher is better). For token-wise mechanisms (InferDPT, SANTEXT+, CusText+, CluSanT), the score is PS $= 1-$ semantic similarity between the original PII string and its perturbed value. PhraseDP is evaluated with normalized exact-match binary protection and plotted on the same 0–1 scale.

Name protection Name	Email protection Email	Phone protection Phone
Address protection Address		

Figure 2: Multi-dimensional PII Protection Radar Charts

Multi-dimensional PII protection radar charts at different privacy budgets ($\varepsilon \in \{1.0, 2.0, 3.0\}$). PhraseDP (yellow) and InferDPT (blue) consistently occupy larger protection areas across all dimensions (email, phone, address, name), demonstrating superior comprehensive PII protection capabilities.

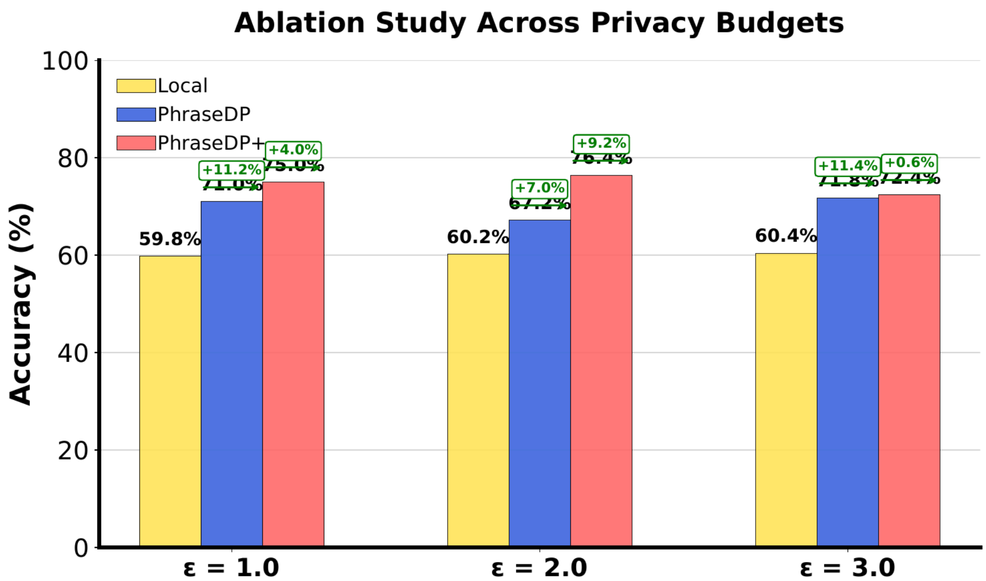
ε = 1.0	**ε = 2.0**	**ε = 3.0**
Radar chart epsilon 1.0	Radar chart epsilon 2.0	Radar chart epsilon 3.0

> **Impact:** Now captures expected downward trend in protection as epsilon increases, better reflecting the privacy-utility trade-off. **Plots already updated in Overleaf.**

2. Ablation Study - Component Contributions

Study Design

Local (baseline) → PhraseDP (CoT effect) → PhraseDP+ (medical mode effect)



Ablation Study: Incremental Improvements Across Privacy Budgets

Key Findings

- Demonstrates **incremental contribution** of each component
- Shows value of:
  1. **PhraseDP-induced CoT:** Improves over Local baseline
  2. **Medical mode:** Additional improvement on top of PhraseDP
- Results across epsilon values (1.0, 2.0, 3.0)
- Publication-ready PDF output for Overleaf integration

3. Medical Mode Improvement Results

Overview

Medical mode preserves medical terminology while removing PII, showing consistent improvements across epsilon values.

Key Statistics

Epsilon	Questions Tested	Questions Improved	Improvement Rate
1.0	127	53	**41.7%**
2.0	149	61	**40.9%**
3.0	149	61	**40.9%**

> **Consistency:** Medical mode shows consistent effectiveness (~41%) across all epsilon values, indicating robust performance regardless of privacy level.

4. PhraseDP++ Few-Shot Prompting - The Backfire

Performance Degradation

Method	Accuracy	Notes
PhraseDP+ (no few-shot)	**82.4%**	Baseline
PhraseDP++ (with few-shot)	**75.6%**	Current experiment
Degradation	**-6.8 pp**	34 questions lost

Root Causes

1. Shorter CoT Responses

- 10.6% too short (<100 chars)
- Correct answers: 559.5 chars avg
- Incorrect answers: 429.3 chars avg

2. Few-Shot Interference

- 10.2% mention example keywords inappropriately
- DIC, endotoxin, hydronephrosis, ACS, PCI
- Examples bias reasoning toward specific topics

3. Reduced Reasoning Depth

- 11.2% lack reasoning keywords
- 10.8% direct answer only
- 11.2% error messages

4. Overfitting

- 15.2% show inappropriate overfitting
- Mimics example structure/terminology
- Pattern-matches rather than reasons

> **Key Finding:** Few-shot examples are too specific and cause overfitting. Dialog-style format may encourage brevity over detailed reasoning.

5. Current Directions - What We're Trying

Approach 1: Better Few-Shot Prompting Techniques

- **Redesign few-shot examples:**
  - Make examples more general and less topic-specific
  - Test different few-shot styles (system_block vs dialog)
  - Avoid examples that bias toward specific medical scenarios
- **Goal:** Maintain reasoning structure benefits without content mimicry

Approach 2: Better COT-Inducing Prompts

- **Investigate CoT quality differences:**
  - Compare average CoT length between few-shot and non-few-shot
  - Analyze reasoning structure differences
  - Check if prompts can encourage deeper reasoning without few-shot examples
- **Focus on:**
  - Prompts that explicitly request step-by-step reasoning
  - Medical terminology preservation in prompts
  - Structured reasoning format requirements

Next Steps

Immediate Actions

- Remove few-shot from PhraseDP++ (restore to PhraseDP+ baseline)
- Test PhraseDP+ without few-shot to confirm baseline (82.4%)
- Continue experiments with alternative prompting strategies

Future Work

- Evaluate model-specific behavior (GPT-5) affects
- Document findings for future improvements
- Develop improved COT-inducing prompts

Summary & Next Steps

Completed ✅

- ✅ Fine-grained PPI evaluation with semantic similarity
- ✅ Updated plots (line plots + radar plots) - **Already in Overleaf**

- ✅ Ablation study showing component contributions
- ✅ Medical improvement analysis across epsilon values

In Progress 🔄

- 🔄 Investigating why few-shot prompting backfired
- 🔄 Testing better few-shot prompting techniques
- 🔄 Developing improved COT-inducing prompts

Immediate Actions 📋

- Remove few-shot from PhraseDP++ (restore to PhraseDP+ baseline)
- Continue experiments with alternative prompting strategies
- Document findings for future improvements

> **Key Takeaway:** Medical mode (PhraseDP+) shows consistent ~41% improvement rate across all epsilon values. Few-shot prompting (PhraseDP++) degrades performance by 6.8 percentage points, indicating that current few-shot examples are counterproductive.