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
- **Radar plots:** Multi-dimensional comparison across epsilon values
- **All 5 mechanisms:** PhraseDP, InferDPT, SANTEXT+, CusText+, CluSanT
- **Token-level experiments:** N=1000 examples

FIGURE 1: PII PROTECTION SCORES

Name | Email | Phone | Address

PS = 1− semantic similarity (0-1 scale)

FIGURE 2: MULTI-DIMENSIONAL PII PROTECTION

ε = 1.0 **ε = 2.0** **ε = 3.0**

PhraseDP (yellow) and InferDPT (blue) show superior protection

IMPACT

Now captures expected downward trend in protection as epsilon increases, *better reflecting the privacy-utility trade-off.*

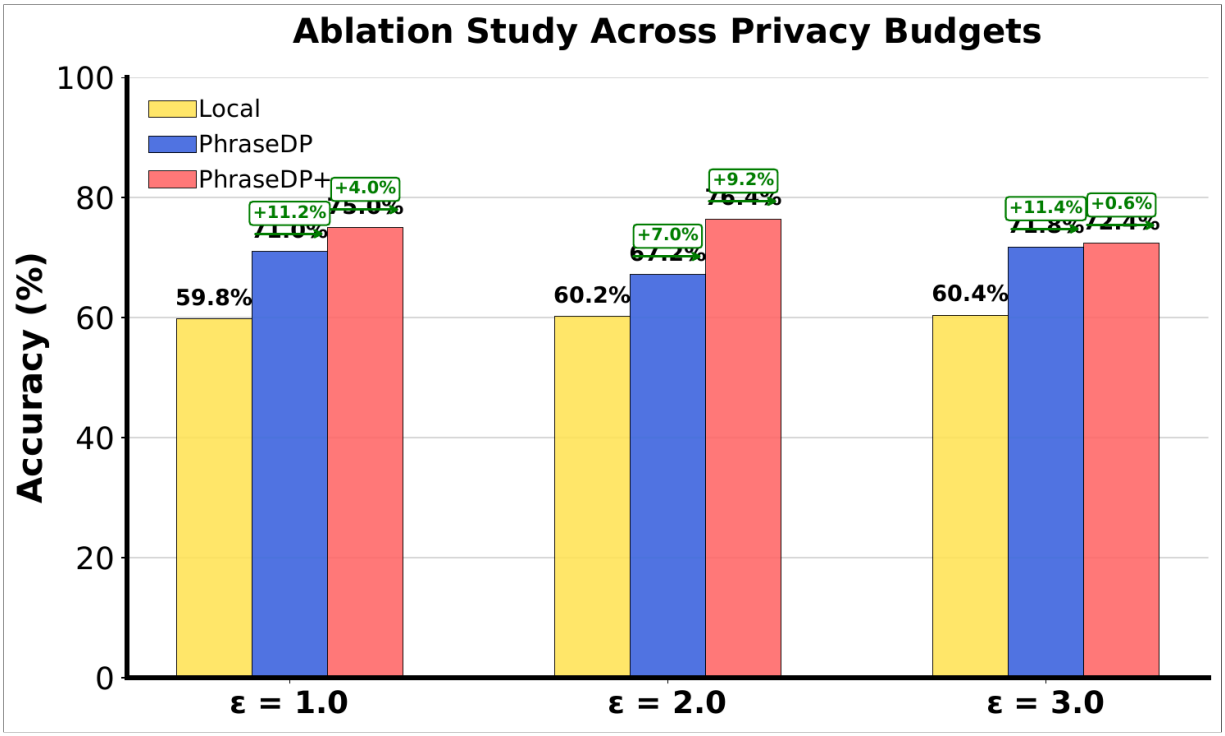
Plots already updated in Overleaf.

2. ABLATION STUDY

COMPONENT CONTRIBUTIONS

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

ABLATION STUDY RESULTS



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)

3. MEDICAL MODE IMPROVEMENT RESULTS

OVERVIEW

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

MEDICAL IMPROVEMENT: EPSILON VALUES

```
<strong>ε = 1.0</strong><br>
<img
src="medqa_medical_improvement_epsilon_1
1.png" style="width: 90%;"
alt="Medical 1.0" />
```

```
<strong>ε = 2.0</strong><br
<img
src="medqa_medical_improvem
1.png" style="width: 90%;"
alt="Medical 2.0" />
```

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*

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 (1/2)

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

ROOT CAUSES (2/2)

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

- **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

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

SUMMARY (CONTINUED)

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.

THANK YOU

Questions?