

# UnPII: Unlearning Personally Identifiable Information with Quantifiable Exposure Risk

*ICSE -SEIP' 26*

Intae Jeon<sup>1</sup>, Yujeong Kwon<sup>2</sup>, Hyungjoon Koo<sup>2</sup>

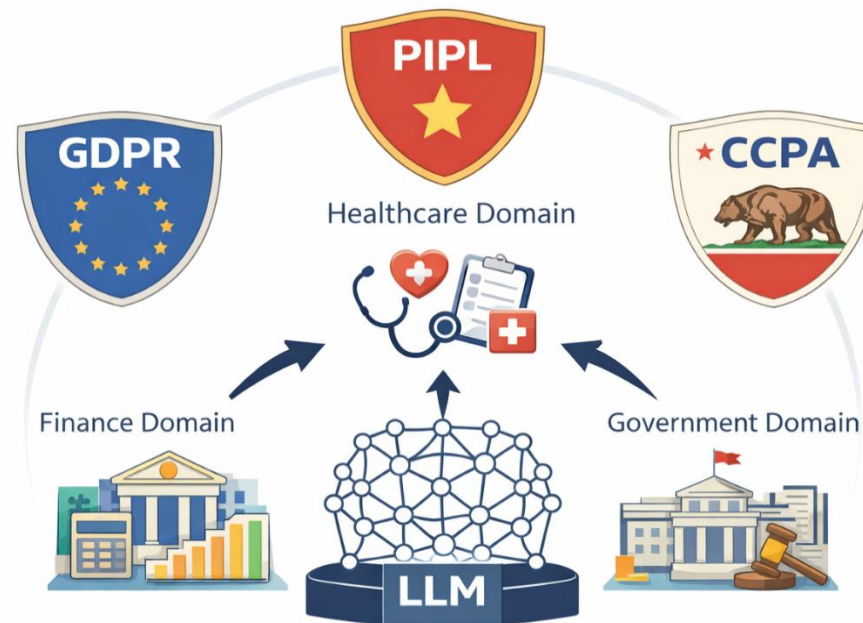
<sup>1</sup>Samsung Research, Seoul, South Korea

<sup>2</sup>Sungkyunkwan University, Suwon, South Korea



# Background

- LLM deployment has been growing in regulated domains
- Training data may contain PII, which can be memorized and exposed at inference time
- Regulations (e.g., GDPR, CCPA, PIPL) require effective data erasure, and full retraining is impractical



# Motivation

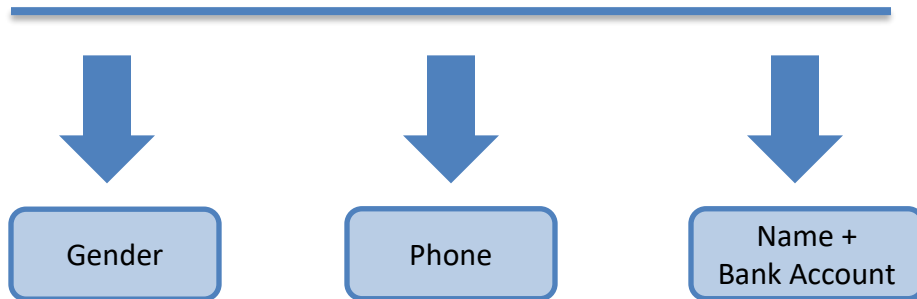
- Existing machine unlearning is not well-aligned with practical PII erasure requirements:
  - PII attributes carry different privacy risks
  - Combining PII attributes amplifies exposure risk
  - Practical LLM unlearning requires calibrated forgetting

# Challenges and Our Solutions

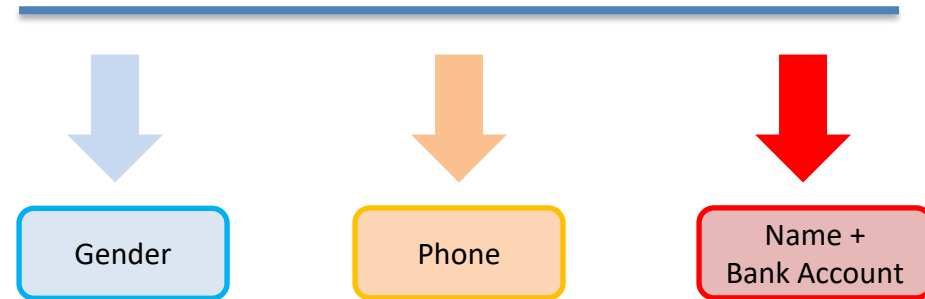
- Dataset: Legal barriers to real-world PII data accessibility → **Synthetic PII**
- Measurement: Quantifying diverse and combined PII exposure risks → **PII Risk Index**
- Method: PII-risk-aware unlearning → **UnPII**  
(Practical design applied on top of existing approaches)
- Evaluation metric: Lack of standardized unlearning effectiveness → **H-AUG**  
(Trade-off between unlearning accuracy, utility, and generalizability)

# Main Idea

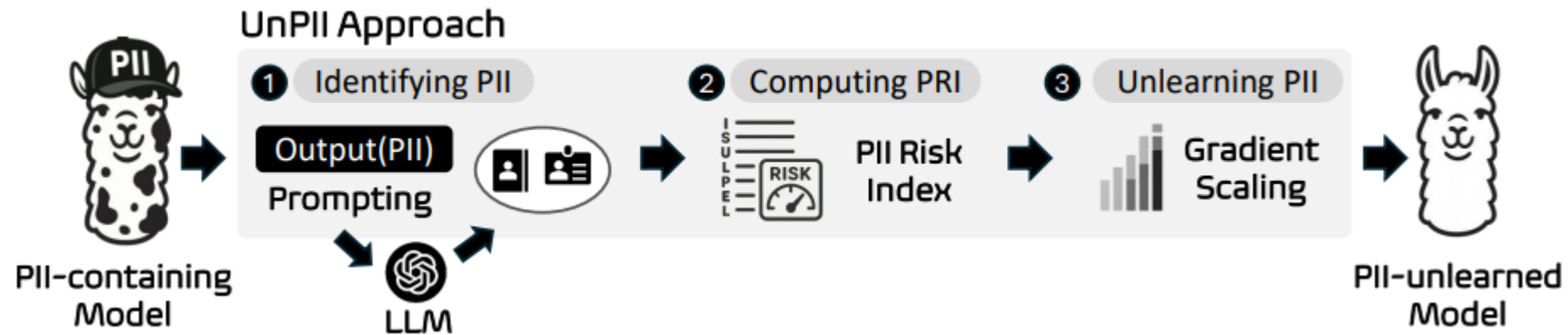
## Existing Approaches Uniform Unlearning Signal



## Our Approach (UnPII) Risk-aware Unlearning Signal



# Proposed Method: UnPII Overview



**Step 1: Identifying PII** using an LLM prompt to detect sensitive tokens

**Step 2: Quantifying PII** by computing the *PII Risk Index (PRI)*

**Step 3: Unlearning PII** with gradient scaling based on the calculated risk

# PII Dataset & Model

- Dataset: 1,700 synthetic PII QA samples (single-PII and compositional-PII categories)
- Model: LLaMA2-7B with LoRA-based fine-tuning
- Examples
  - Single-type PII: e.g., Full Name, Gender, Postal Code, Address
  - Combined-type PII: e.g., Full Name + Address1, Full Name + Address2, Full Name + Medical Record

# Quantifiable Metric for PII Leakage Risk

- **PII Risk Index (PRI):** quantifying potential PII leakage risk
  - Define seven dimensions of PII-related risk factors
    - Identifiability, Sensitivity, Usability, Linkability, Permanency, Exposability, Compliancy
  - Risk increases when multiple PII attributes are jointly exposed, as identifiability rises through composition
    - Example: combining ZIP code, date of birth, and gender can uniquely identify 87% of U.S. residents<sup>[1]</sup>
  - Represent the overall risk as a normalized score in (0, 1)

$$r = \lambda kl + \sum_{i=1}^l \prod_{j=1}^k w_{ij} a_{ij}$$

$$R = \tanh(r) = \frac{e^r - e^{-r}}{e^r + e^{-r}} \in (0, 1) \quad \text{where } r > 0$$

[1] Latanya Sweeney. 2000. Simple demographics often identify people uniquely. Health (San Francisco) (2000).

# Unlearning PII atop Existing Techniques

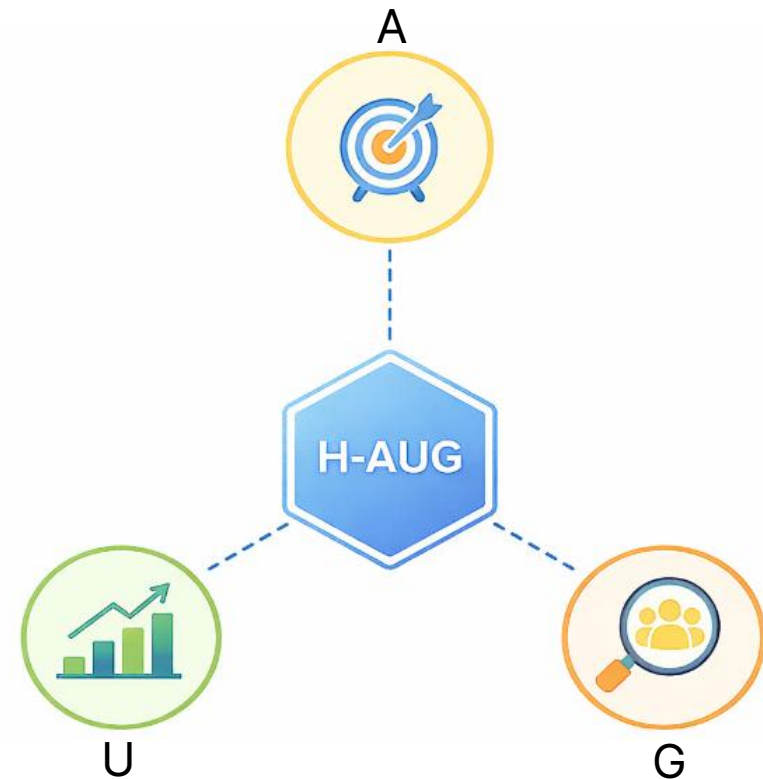
- PRI-weighted loss
  - Easy adoption for gradient-based algorithms (GA/NPO/DPO)
- Seamless integration

$$\mathcal{L}_{\text{UnPII}} = \mathcal{L}_{\text{base}}(1 + \text{PRI}) \quad \text{where } \mathcal{L}_{\text{base}} \in \{\mathcal{L}_{\text{GA}}, \mathcal{L}_{\text{NPO}}, \mathcal{L}_{\text{DPO}}\}$$

Base Loss	UnPII Loss
$L_{\text{GA}} = -\log \pi_{\theta}(y_f   x_f)$	$L_{\text{GA}+\text{UnPII}} = -\log \pi_{\theta}(y_f   x_f) \times (1 + \text{PRI})$
$L_{\text{NPO}} = -\frac{2}{\beta} \log \sigma(-\beta \log \frac{\pi_{\theta}(y_f x_f)}{\pi(y_f x_f)})$	$L_{\text{NPO}+\text{UnPII}} = L_{\text{NPO}} \times (1 + \text{PRI})$
$L_{\text{DPO}} = -\frac{2}{\beta} \log \sigma(\beta \log \frac{\pi_{\theta}(y_p x_f)}{\pi(y_p x_f)} - \beta \log \frac{\pi_{\theta}(y_f x_f)}{\pi(y_f x_f)})$	$L_{\text{DPO}+\text{UnPII}} = L_{\text{DPO}} \times (1 + \text{PRI})$

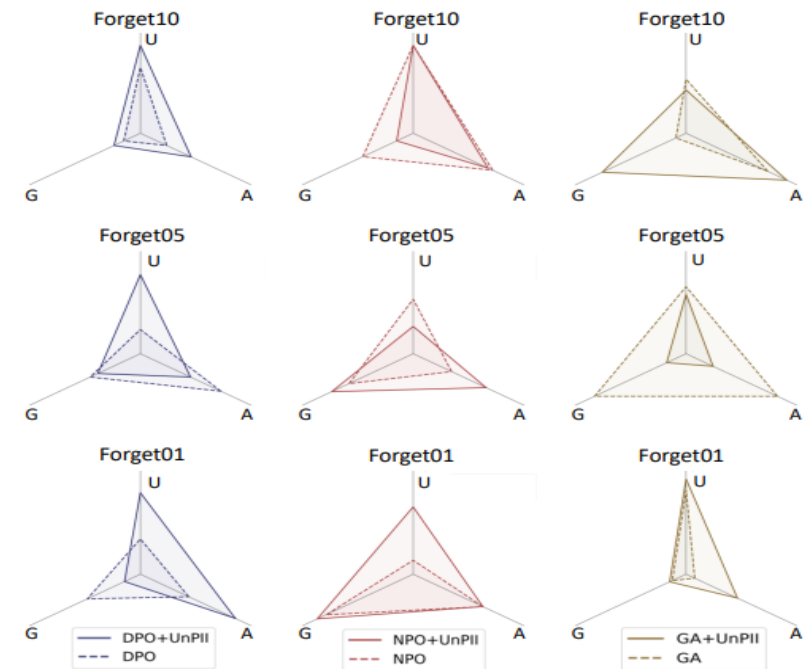
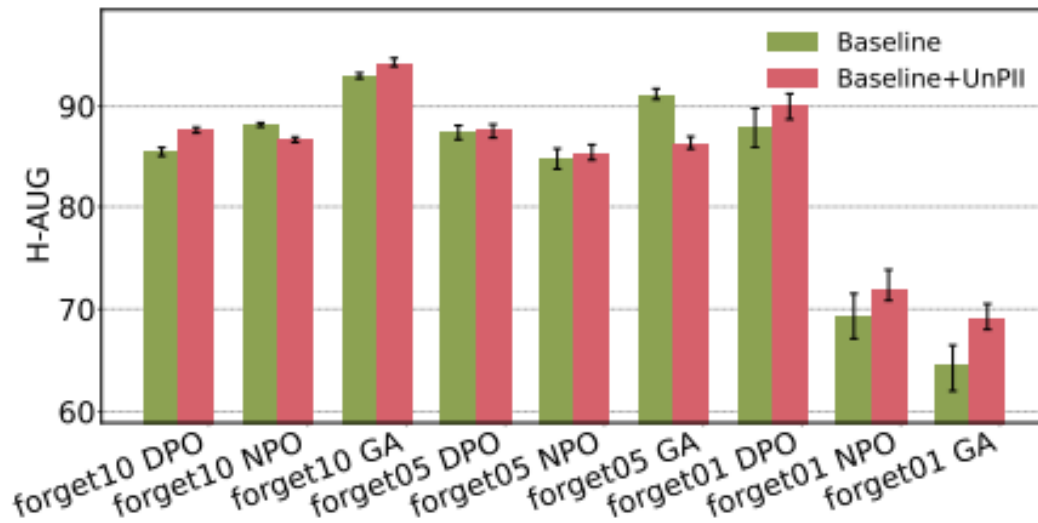
# Unlearning Metric: H-AUG

- Accuracy (A)
  - Effective forgetting of the target PII
- Utility (U)
  - Preserved performance on non-PII data
- Generalizability (G)
  - Forgetting about unseen PII cases
- Harmonic score (H-AUG)
  - Balanced overall unlearning quality



# Evaluation

- Higher H-AUG indicates better trade-off: effective forgetting + preserved utility
- Consistently competitive across forgetting ratios (1%, 5%, 10%)
- Practical implications with a focus on high-risk PII rather than uniform unlearning



# Conclusion

- Risk-aware PII unlearning
  - Risk-proportional forgetting via policy-configurable PRI (PII Risk Index)
- Seamless integration
  - Handy integration with existing (gradient-based) unlearning approaches
- Empirical evidence
  - Superior risk-targeted forgetting compared to uniform baselines

# Q&A

# Quantifiable Metric for PII Leakage Risk

- Simulation assuming 1–10 PII attributes are exposed
  - Based on 1,000 iterations to ensure statistical reliability
  - Overall risk increases as the number of exposed PII attributes grows
  - When 5 or more attributes are exposed, the risk saturates near 1

