RUHR UNIVERSITÄT BOCHUM RUB

LLM Privacy in 2025: What Are Researchers Actually Working On?

Ahmad Mohammadi

December 30, 2025

www.trusthlt.org
Trustworthy Human Language Technologies Group (TrustHLT)
Ruhr University Bochum & Research Center Trustworthy Data Science and Security

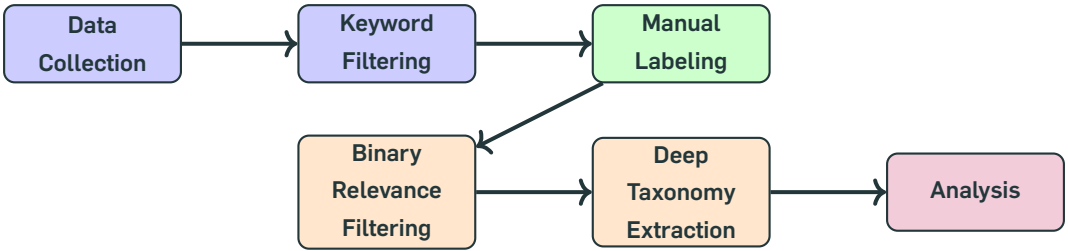
TrustHLT

CENTER FOR TRUSTWORTHY DATA SCIENCE AND SECURITY

Methodology

Experimental Setup

Research Methodology Pipeline



Target Conferences

EMNLP 2025
Suzhou, China | 中国苏州

ACL 2025
VIENNA

ICLR
International Conference On Learning Representations

ICML
International Conference On Machine Learning
2025

NAACL 2025

34TH USENIX SECURITY SYMPOSIUM
AUGUST 13–15, 2025
SEATTLE, WA, USA

NDSS
SYMPOSIUM/2025

Data Collection

Dataset Overview:

- **Total Papers:** 32,855
- **Time Period:** 2025
- **Conferences:** 10 venues

Conference Categories:

- **ML (3):** ICLR, NeurIPS, ICML
- **NLP (3):** ACL, EMNLP, NAACL
- **Security (4):** USENIX, S&P, CCS, NDSS

Keyword Filtering

Filtering Criteria: Papers must contain at least one privacy keyword AND one LLM keyword

Privacy Keywords:

- Core: privacy, private, PII, sensitive data
- Attacks: membership inference, memorization, data extraction, model inversion
- LLM-specific: chat leakage, prompt leakage, RAG privacy
- Defenses: differential privacy, DP-SGD, federated learning, unlearning
- Frameworks: contextual integrity, anonymization, k-anonymity
- Security: adversarial, backdoor, poisoning, side-channel
- Regulations: GDPR, CCPA, HIPAA

LLM Keywords:

- Models: language model, LLM, GPT, BERT, transformer, LLaMA
- Architectures: neural network, attention, encoder-decoder, autoregressive
- Capabilities: in-context learning, few-shot, RAG, agent, multimodal

Result: 3,203 papers matched filtering criteria

Manual Labeling

Sample Selection:

- **50 papers** randomly selected from filtered dataset
- Balanced across conferences and years
- Used as training examples for automated pipeline

Labeling Process:

1. **Binary Classification:** Yes/No for LLM privacy relevance
2. **Taxonomy Extraction:** 7 fields for relevant papers with Label Yes

Taxonomy Fields

1. Contribution Type
2. Privacy Framework
3. Threat Model
4. Methodology
5. Privacy Incident
6. Lifecycle Stage
7. Keywords (5 per paper)

Taxonomy Framework: 6-Dimensional Analysis

1. Contribution Type

Attack, Defense, Analysis, Theory, Survey, Audit

2. Privacy Framework

Differential Privacy (DP), Local DP (LDP), DP-SGD, f-Differential Privacy (f-DP), Contextual Integrity, None

3. Threat Model

Black-box, White-box, Gray-box

4. Methodology

Empirical, Theoretical, Both, Survey

5. Privacy Incident

Training Data Leakage, Direct Chat Leakage, Indirect Chat Leakage, Indirect Attribute Inference, Direct Attribute Aggregation

6. Lifecycle Stage

Pre-training, Training, Post-training, Inference, Post-deployment

Automated Labeling: Claude Sonnet 4.5

Two-Stage Pipeline:

Stage 1: Binary Relevance Filtering

Model: Claude Sonnet 4.5
Input: Title + Abstract
Output: Yes/No + Confidence (1-5)
Prompt: Few-shot (2 positive + 2 negative examples)
Result: 380 relevant papers

Stage 2: Deep Taxonomy Extraction

Model: Claude Sonnet 4.5
Input: Title + Abstract
Output: 6 taxonomy fields + 5 keywords
Prompt: Few-shot (3 examples per field)
Result: Complete taxonomy for all relevant papers

Technical Details:

- **API:** Anthropic API
- **Model:** claude-sonnet-4-5-20250929
- **Rate:** 1.5s delay/request
- **Optimization:** Prompt caching (90% cost reduction)

Cost Analysis:

- **Stage 1:** 16€ (3,203 papers)
- **Stage 2:** 7€ (380 papers)
- **Total:** 23€

Stage 1 Prompt: Binary Relevance Filtering

Task: You are a research paper classifier specializing in privacy and large language models (LLMs). Your task is to determine if a paper is relevant to privacy in large language models (LLMs) and language models specifically.

A paper is RELEVANT if it discusses:

- Privacy attacks on LLMs (membership inference, data extraction, prompt injection)
- Privacy defenses in LLMs (differential privacy, federated learning, encryption)
- Training data privacy and memorization in LLMs
- Privacy in fine-tuning, RLHF, or alignment of LLMs
- Privacy in conversational AI and chatbots

A paper is NOT RELEVANT if it:

- Only discusses privacy in general ML models without LLMs
- Focuses on adversarial robustness for accuracy (not privacy)
- Only mentions LLMs or privacy tangentially

IMPORTANT: The paper MUST involve Large Language Models or Language Models. Papers about general machine learning privacy (e.g., privacy in image classifiers, graph neural networks without language models) are NOT relevant.

Here are examples from manually labeled papers:

Few-shot examples: 2 relevant + 2 not relevant papers

Output format: Respond with ONLY a JSON object in this exact format: "relevant": "Yes" or "No", "confidence": 1-5 (1=very unsure, 5=very confident), "reason": "One sentence explaining your decision"

Stage 2 Prompt: Taxonomy Extraction

Task: You are an expert in extracting structured information from privacy and large language model (LLM) research papers. Your task is to extract 7 specific fields from a paper about privacy in LLMs/language models based on its title and abstract.

Field Definitions:

1. **Keywords:** 5 most important technical terms
2. **Privacy Framework:** DP, LDP, f-DP, DP-SGD, Contextual Integrity, None
3. **Threat Model:** Black-box, White-box, Gray-box
4. **Methodology:** Empirical, Theoretical, Both, Survey
5. **Privacy Incident:** Training Leakage, Chat Leakage, Context Leakage, Attribute Inference, Aggregation
6. **Lifecycle Stage:** Pre-training, Training, Post-training, Inference, Post-deployment
7. **Contribution Type:** Attack, Defense, Analysis, Theory, Survey

Few-shot examples: 3 fully annotated papers with all fields

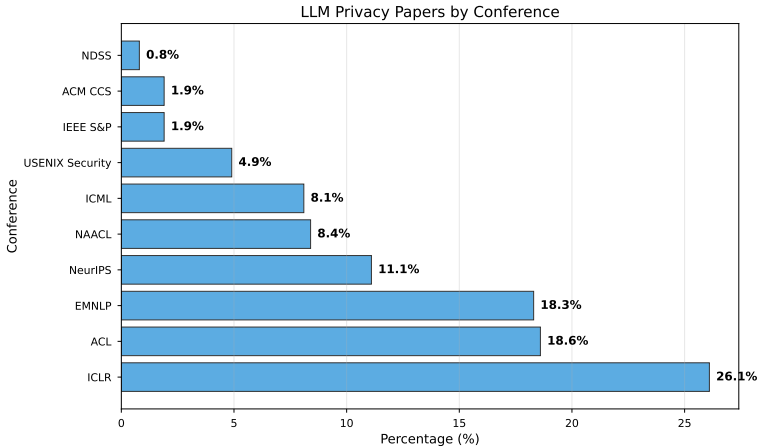
Output format: JSON with all 7 fields using exact category names

Important: - Use EXACT category names from the definitions above - Choose the MOST APPLICABLE category for each field - Keywords should be 5 specific technical terms - If unsure, choose the closest match

Analysis

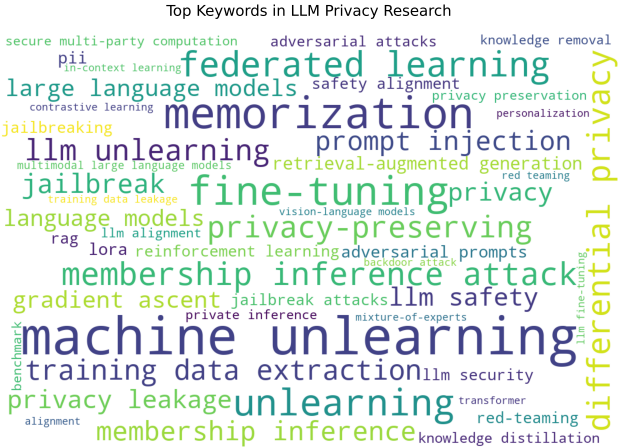
Results & Findings

LLM Privacy Papers by Conference



Key Finding: ICLR leads (26.1%); top 3 venues (ICLR, ACL, EMNLP) account for 63% of all LLM privacy papers

Key Research Themes and Topics

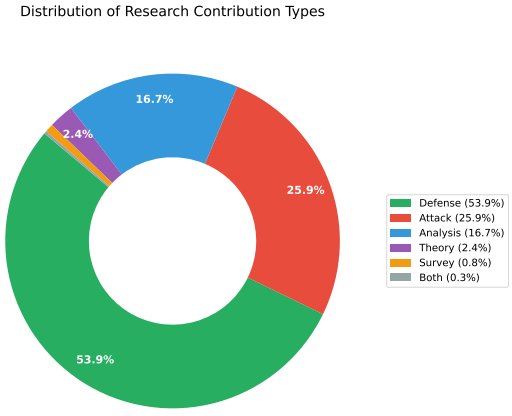


Key Finding: Machine unlearning, membership inference, and memorization dominate; federated learning and fine-tuning emerge as key privacy-preserving approaches

Research Characteristics by Venue Type

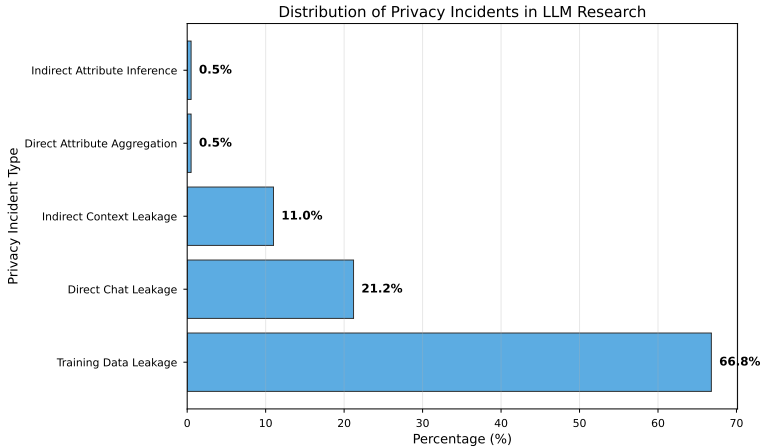
Key Insight: Security venues prioritize attacks and inference; NLP venues lead in defenses and empirical work; ML venues show balanced profile

Research Contribution Types



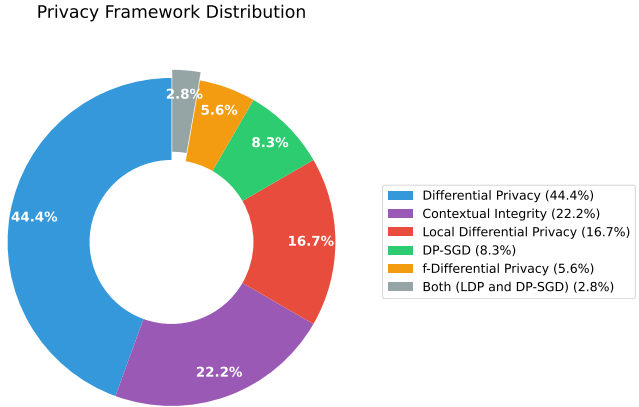
Key Finding: Defense research dominates (53.9%) and outnumbers attacks (25.9%) by 2:1 ratio; theory severely underrepresented (2.4%)

Privacy Incidents Studied



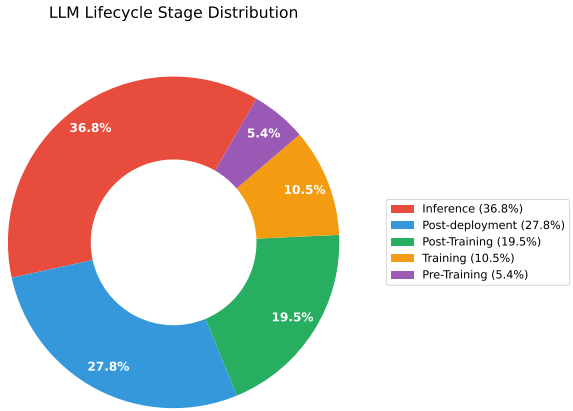
Key Finding: Training data leakage dominates (66.8%); chat leakage emerging as secondary concern (21.2%)

Privacy Frameworks Used



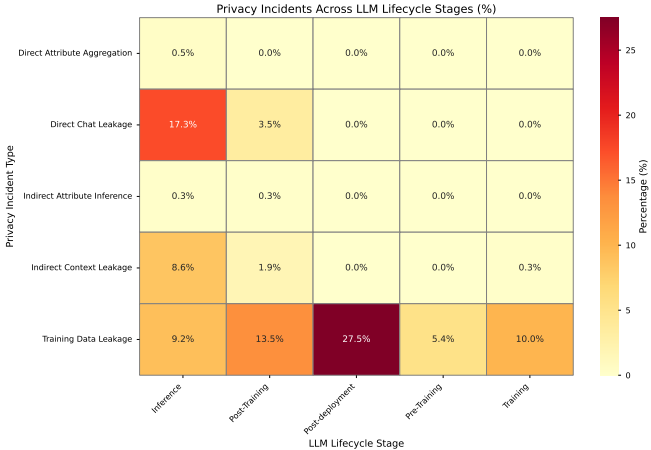
Key Finding: Among papers using formal frameworks (9.7% of all papers), Differential Privacy dominates (44.4%), followed by Contextual Integrity (22.2%)

LLM Lifecycle Stage Coverage



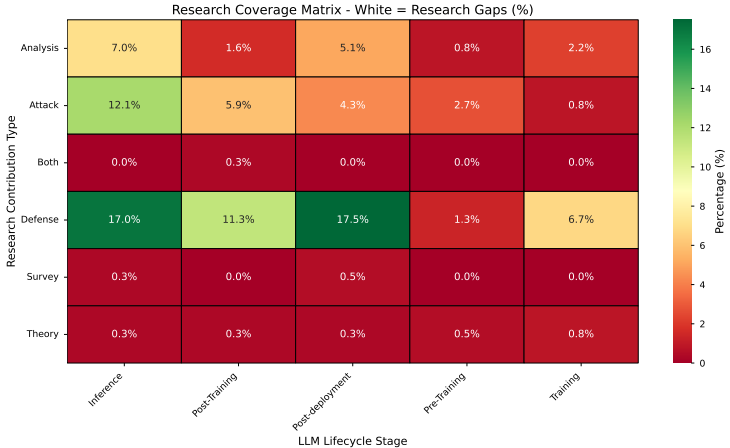
Key Finding: Research heavily focuses on inference (36.8%) and post-deployment (27.8%); training (10.5%) and pre-training (5.4%) severely under-researched

Privacy Incidents Across LLM Lifecycle



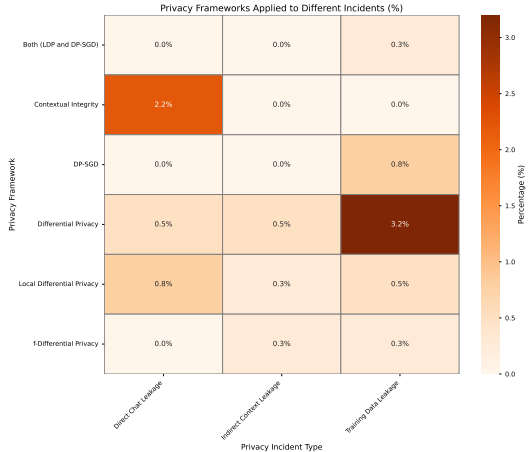
Key Insight: Training data leakage studied across all stages (peak at post-deployment: 27.5%); chat leakage concentrated at inference (17.3%)

Research Gap Analysis



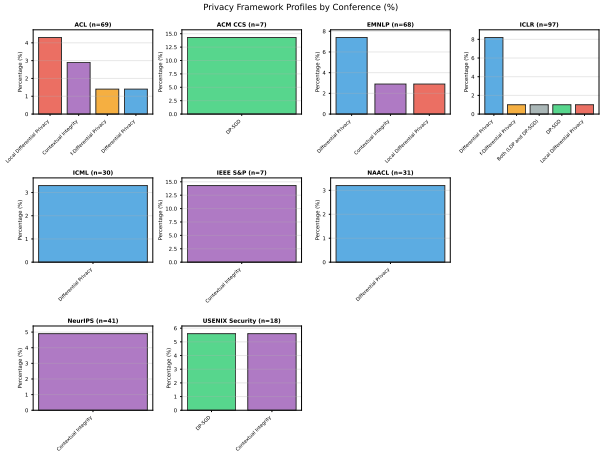
Key Insight: Critical gaps in pre-training stage (<3% across all types); theory severely lacking (<1%); defenses concentrate on inference (17.0%) and post-deployment (17.5%)

Framework Coverage for Privacy Incidents



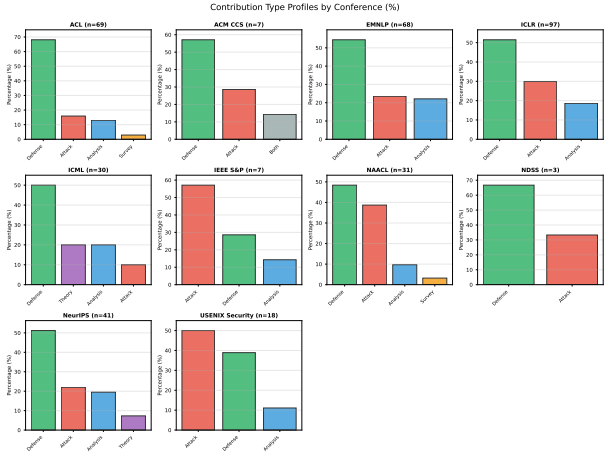
Key Insight: Differential Privacy primarily addresses training data leakage (3.2%); Contextual Integrity targets chat leakage (2.2%); major coverage gaps remain across all incident types

Privacy Framework Adoption by Conference



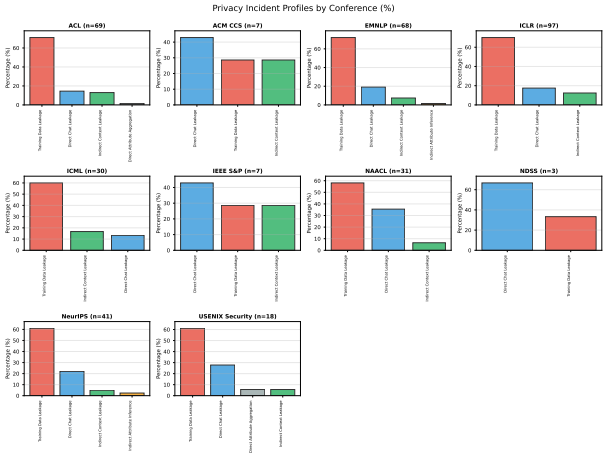
Key Insight: Formal framework adoption extremely low across all venues (<15%); Security venues show highest rates (CCS: DP-SGD 14%, S&P: Contextual Integrity 14%); ML/NLP venues minimal adoption

Contribution Type Profiles by Conference



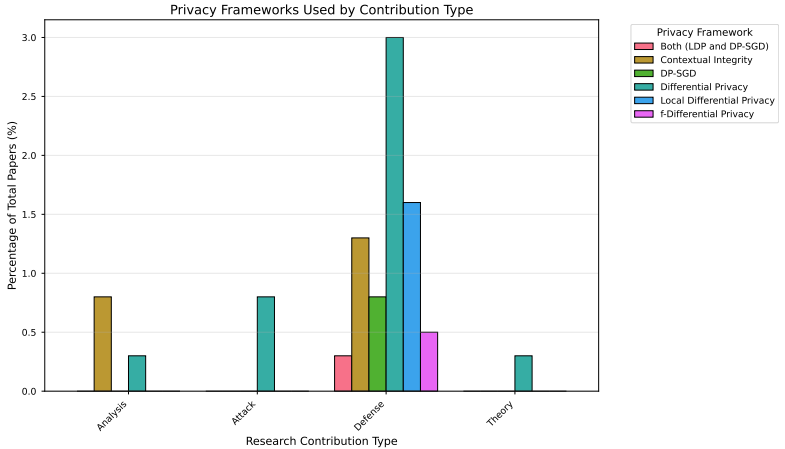
Key Insight: NLP venues strongly defense-oriented (ACL 70%, NAACL 65%); Security venues favor attacks (USENIX 45% attack vs 30% defense); ML venues balanced but defense-leaning

Privacy Incident by Conference



Key Insight: ML venues prioritize training data leakage (ICLR 70%); Security venues show highest chat leakage focus (NDSS 65%, CCS 40%); NLP venues more balanced between both incident types

Privacy Frameworks by Contribution Type



Key Insight: Formal framework adoption critically low across all types (<3%); defenses slightly favor Differential Privacy; attacks and analysis largely framework-agnostic

Key Findings (1/2)

1. **Empirical Work Dominates, Theory Critically Lacking**
   86.8% of papers are purely empirical with only 0.8% using theoretical methodology, revealing severe gap in formal foundations
2. **Training Data Leakage is the Primary Privacy Threat**
   66.8% of papers focus on Training Data Leakage, with Direct Chat Leakage (21.2%) as a secondary concern
3. **Defense Research Dominates LLM Privacy**
   Defense papers (53.9%) outnumber attacks (25.9%) by 2:1 ratio, unlike traditional ML privacy where attacks dominated
4. **Research Heavily Focuses on Inference and Post-Deployment**
   Inference (36.8%) and post-deployment (27.8%) stages dominate, while training phase severely under-researched (10.5%)

Key Findings (2/2)

5 **Machine Unlearning Emerges as Hottest Research Topic**

Machine unlearning appears in 15% of papers (56 papers), indicating growing focus on data deletion and GDPR compliance

6 **Black-Box Threat Models Dominate LLM Privacy Research**

73.3% assume black-box adversaries with only API access, reflecting practical deployment scenarios

7 **Formal Privacy Frameworks Rarely Applied**

Only 9.7% use formal frameworks (DP, LDP, DP-SGD); 90% of LLM privacy research operates without mathematical guarantees

Questions?

Thank you!