Unveiling ML Developers' Perceived and Actual Barriers of Privacy-Preserving Machine Learning Techniques

Odunayo Akinlade
Carnegie Mellon University

Yacoba Oduro-Yeboah
Carnegie Mellon University

Nouba-Asra Goursam Tresor
Carnegie Mellon University

Hana Habib
Carnegie Mellon University

Yuvraj Agarwal
Carnegie Mellon University

Lorrie Cranor
Carnegie Mellon University

1 Introduction

As learning systems are increasingly deployed in sensitive domains such as healthcare and finance, privacy-preserving machine learning (PPML) techniques have become an important consideration in responsible ML development. Among these, differential privacy (DP) has gained particular traction due to its formal mathematical guarantees, and accessible libraries such as Opacus [1]. Prior work has documented that DP implementation is non-trivial even for technically sophisticated users, with recurring challenges around parameter configuration, workflow integration, and interpreting privacy guarantees [2]. There is also evidence that tool accessibility and documentation quality shape adoption decisions as much as technical capability [3, 4].

However, existing studies typically capture either perceived barriers through interviews [2,4] or observed barriers through task-based evaluations [5], and to our knowledge, no prior work directly compares what developers anticipate will be difficult with what actually proves difficult during implementation. This exploratory study addresses that gap by combining structured interviews with a hands-on lab task using Opacus, focusing on Differential Privacy (DP) as the technique of interest. The research questions are:

RQ1: What baseline conceptual understanding do ML developers have of PPML techniques like differential privacy, federated learning, and synthetic data generation, and how do they currently think about incorporating privacy into model development?

RQ2: What barriers do ML developers perceive when implementing differential privacy in machine learning workflows, and what barriers do they face during hands-on implementation tasks using Opacus?

RQ3: How do ML developers interpret and select privacy parameters in DP model training tasks, and what are the consequences of their choices?

Our initial findings show that anticipated barriers did not fully reflect actual challenges.Participants struggled with API-level friction, unclear parameter meanings, and a disconnect in the Opacus API that led six out of ten participants to unknowingly bypass their intended privacy budget. While most recognized that more sensitive data required stronger protections, they lacked the knowledge to translate this into appropriate parameter configurations.

2 Methods

We recruited 10 participants after they had filled out an initial screening survey and met the eligibility requirements. The study included a short interview followed by a hands-on lab task and brief follow-up questions. The interview captured developers' baseline understanding of privacy-preserving machine learning(PPML) techniques and their perceived barriers to implementation. The task component enabled us to observe how participants actually configure and reason about DP during a structured implementation task. Sessions lasted 60-75 minutes, with the interview taking 10–15 minutes, and the task component taking 45-60 minutes. Participants received compensation of $15 after completing the study session in the form of Amazon gift cards. Compensation was not contingent on the performance of the task or the quality of responses. The study obtained Institutional Review Board (IRB) approval prior to data collection and adhered to established ethical principles for human-subjects research. Participation was voluntary, with informed consent obtained from all participants before the start of the sessions.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2025.
August 10–12, 2025, Seattle, WA, United States.

Interview: During the interview, we asked participants about their familiarity with PPML techniques, including differential privacy, federated learning, and synthetic data generation and how they considered privacy in the ML development workflows. Through the interview, we were able to elicit detailed responses regarding participants background, their general privacy practices, awareness of PPML techniques and perceived barriers to their implementation of PPML.

Task Component: The lab portion used a within-subjects design. We gave participants a working PyTorch training script for a simple classification task that had already been trained successfully without privacy protections, and were expected to modify the training process to incorporate Differential Privacy using the Opacus library. To achieve this, we asked participants to integrate the Opacus Privacy Engine, configure key parameters such as noise multiplier, clipping norm, and privacy budget, and retrain the model with Differential Privacy enabled under two scenarios: one involving a low-sensitivity dataset and another involving a highly sensitive dataset. To control for order effects, we randomly assigned participants to one of the two scenarios first.

Data Analysis: We analyzed the qualitative data by conducting thematic analysis on interview transcripts and open-ended responses to identify patterns in reasoning, misunderstandings, and implementation behavior. Two researchers independently coded a shared subset of transcripts using a structured codebook developed iteratively from the data, with disagreements resolved through discussion. Quantitative data included logged parameter values, model accuracy outputs, task completion status, and post-task survey responses, which were examined descriptively across participants.

3 Preliminary Results

Developers conceptualized privacy risk broadly, not as an ML-specific problem: Participants commonly described privacy risk as the exposure or misuse of sensitive data and Personally Identifiable Information (PII). However, none demonstrated awareness of ML-specific privacy attacks such as membership inference or model inversion attacks. Most framed privacy primarily as a data access or preprocessing issue rather than a risk emerging from model training itself.

Participants showed conceptual awareness without operational understanding: Most participants correctly recognized that medical datasets required stronger privacy protections than lifestyle datasets. However, this awareness did not translate into principled parameter choices during implementation. Participants generally understood that lower epsilon values imply stronger privacy, but lacked a meaningful basis for selecting specific values.

Parameter choices were often heuristic or externally driven: Participants frequently selected DP parameters arbitrarily, copied values from tutorials, or relied on AI tools such as Gemini. Noise multiplier and clipping norm were often left at default values without justification, and several participants could not explain what these parameters controlled.

API design caused silent privacy failures: Six out of ten participants set a target epsilon value expressing their intended privacy goal, then used Opacus' `make_private` API, which ignores epsilon entirely and instead relies on the noise multiplier. Participants were unaware that their stated privacy intent had been decoupled from the actual training configuration, and the API provided no warning or indication of this mismatch.

Observed barriers differed from anticipated barriers: Before the task, participants primarily anticipated challenges related to accuracy-privacy tradeoffs and limited documentation. During implementation, the dominant barriers were API-level friction, uncertainty about parameter meanings, and difficulty identifying which parameters were actually controlling privacy behavior.

4 Discussion

Our findings suggest that current Differential Privacy (DP) tools support implementation at the code level but do not adequately support developers' reasoning about privacy decisions. While participants often recognized the need for stronger privacy protections in sensitive contexts, they struggled to translate this awareness into meaningful parameter configurations. In several cases, participants believed they had configured a target privacy budget but their chosen API method ignored it entirely.

These findings highlight a gap between conceptual understanding and operational implementation in privacy-preserving machine learning workflows. Future PPML tools should provide clearer interpretive guidance for parameters, better distinction between API methods, and more explicit feedback linking developers' privacy intentions to actual training behavior.

5 Next Steps

Future work will expand this study with a larger cohort of developers with prior experience with training models with Opacus to better examine how expertise shapes privacy reasoning, parameter selection, and implementation behavior. We also plan to refine the study protocol to incorporate a deeper evaluation of privacy-utility tradeoff reasoning and post-task reflection on inference attack risks and parameter adjustments.

References

[1] Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, Graham Cormode, and Ilya Mironov. Opacus: User-friendly differential privacy library in PyTorch, August 2022. https://arxiv.org/abs/2109.12298.

[2] Gonzalo Munilla Garrido, Xiaoyuan Liu, Florian Matthes, and Dawn Song. Lessons learned: Surveying the practicality of differential privacy in the industry, November 2022. https://arxiv.org/abs/2211.03898.

[3] Natalia Ponomareva, Hussein Hazimeh, Alex Kurakin, Zheng Xu, Carson Denison, H. Brendan McMahan, Sergei Vassilvitskii, Steve Chien, and Abhradeep Guha Thakurta. How to DP-fy ML: A practical guide to machine learning with differential privacy. *Journal of Artificial Intelligence Research*, 77:1113–1201, July 2023. https://doi.org/10.1613/jair.1.14649.

[4] Patrick Song, Jayshree Sarathy, Michael Shoemate, and Salil Vadhan. "I inherently just trust that it works": Investigating mental models of open-source libraries for differential privacy. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2):1–39, November 2024. https://doi.org/10.1145/3687011.

[5] Ivoline C. Ngong, Brad Stenger, Joseph P. Near, and Yuanyuan Feng. Evaluating the Usability of Differential Privacy Tools with Data Practitioners. In *Twentieth Symposium on Usable Privacy and Security (SOUPS 2024)*, pages 21–40. USENIX Association, 2024.

A Code Book

Below is the thematic codebook used to analyze participant implementation behavior and reasoning patterns.

Table 1: Study Codebook: Categories, Codes, Definitions, and Participant Frequency (*N*)

Category	Code	Definition / Inclusion Criteria	N
ML experience level	Intermediate	Has done ML in courses, tutorials, or small projects; understands basic concepts but limited real-world workflow experience.	6
	Advanced	Has trained and evaluated models end-to-end, understands workflows, may have deployed or worked on real projects.	4
Data type	Sensitive	Participants worked on a dataset classified as private, such as healthcare or finance.	6
	Non sensitive	Participant has not worked with data that required privacy protections.	7
Privacy risk mental model	Unauthorized access	frames privacy risk as the possibility of data being accessed by unauthorized parties, without reference to ML-specific attack vectors.	3
	ML-specific attack awareness	demonstrates awareness of attacks like membership inference, model inversion, or re-identification.	2
	Sensitive Data Exposure	sees privacy risk as the use or exposure of personal or sensitive data (e.g., health data, PII, financial data) during training.	5
	Vague	uses privacy language but cannot articulate a concrete mechanism of harm in the ML context.	5
Privacy integration point	Pre-processing	believes privacy should be addressed before training, through anonymization or data minimization.	3
	Training stage	identifies model training as the point where privacy-preserving techniques should be applied.	2
	All stages	views privacy as a concern that spans the entire ML pipeline.	2
	No clear integration point	does not identify a specific pipeline stage where privacy should be addressed.	1
Privacy protection methods used	Anonymization	removed or masked identifying fields before using data in a model.	5
	Data Filtering	selectively restricted access to, or removed, specific sensitive data points.	2
	Access control	relied on infrastructure-level protections such as restricted access or encryption at rest.	3
	Differential privacy	applied differential privacy (noise addition) as a protection mechanism.	3
	Synthetic data generation	applied synthetic data generation as a protection mechanism.	2

Category	Code	Definition / Inclusion Criteria	N
Differential privacy knowledge	Accurate understanding	correctly explains DP as a mathematical guarantee limiting what can be inferred about individuals.	1
	Partial understanding	has some correct intuition about DP but cannot fully articulate the mechanism or guarantees.	1
	Misconception	conflates DP with other techniques (e.g., anonymization) or holds an inaccurate definition.	3
	No awareness	has not encountered the term differential privacy before this study.	3
Federated learning knowledge	Accurate understanding	correctly describes FL as training models across distributed sources without centralizing raw data.	3
	Partial understanding	has some correct intuition about FL but cannot fully articulate the mechanism.	4
	Misconception	confuses FL with other distributed or privacy-related approaches.	2
Synthetic data generation knowledge	Accurate understanding	describes SDG as creating artificial data that preserves statistical properties without exposing real records.	5
	Partial understanding	knows SDG involves artificial generation but unclear on how privacy guarantees are achieved.	2
	No awareness	has not encountered SDG as a privacy-preserving technique before this study.	1
Self-reported implementation	Production implementation	has deployed a PPML technique in a live system or real-world project.	1
	Experimental	has tried a PPML technique in a personal project, course, or sandbox.	2
	No implementation	has conceptual awareness but has never implemented any technique.	4
Perceived Barriers	Implementation difficulty	reports or anticipates technical friction when integrating PPML into ML workflows.	2
	Lack of examples	identifies the absence of realistic, worked examples as a barrier.	3
	Documentation	finds official documentation insufficient, unclear, or too theoretically oriented.	2
	Accuracy tradeoffs	recognizes or anticipates degradation in model performance.	4
	High Computation Cost	mentions high computation cost as a challenge.	3
	Privacy Reasoning	struggle with knowing what needs to be private or not.	3
	Understanding parameters	does not understand what parameters represent in terms of privacy.	1
Task Approach	Read documentation first	Read the Opacus documentation before implementation.	9
	Trial and error	Iteratively changes parameters/code without a clear plan.	2
Epsilon Choice	Justified	Mentions the privacy-utility tradeoff explicitly.	6
	Heuristic-based	Based choice on values seen before (e.g., 1.0 or 8.0).	1
	Copied value	Taken from documentation/tutorial without reasoning.	1
	Random	No clear rationale provided.	1

Category	Code	Definition / Inclusion Criteria	N
Other parameters Choice	Understood relationship	correctly links the parameter to its role (e.g., noise increases privacy).	1
	Default value	uses example values without modification or justification.	1
	Heuristic-based	choice based on personal experience or intuition.	3
	Used value from AI	used values generated by AI without modification.	2
	Choose a random value	selects values arbitrarily.	4
	No understanding	cannot explain the parameter or selects values arbitrarily.	2
Privacy Reasoning	Explicit tradeoff	articulates the relationship between privacy and accuracy.	5
	Accuracy prioritization	ignores privacy implications to favor accuracy.	2
	Surface-level reasoning	uses terms without conceptual depth.	5
Resource Used	Official Documentation	consults Opacus or PyTorch docs/API references.	7
	AI summary	uses Google AI overviews or snippets rather than clicking sources.	5
	AI tools	directly queries ChatGPT, Gemini, etc., for guidance or code.	7
	Prior knowledge	relies on memory without consulting external resources.	1
	Blogs/Articles	uses Medium, Stack Overflow, or other community guides.	4
Error Faced	Syntax error	formatting or language issues (e.g., missing brackets).	6
	Library misuse	incorrect use of APIs or Opacus/PyTorch mismatch.	2
	Conceptual error	misunderstanding DP concepts (e.g., wrong parameter logic).	3
Parameter Adaptation	Adjusted for sensitivity	explicitly links parameter changes to dataset sensitivity/context.	1
	Reused previous values	keeps the same values across scenarios without reconsideration.	3
	Inconsistent changes	arbitrary or contradictory changes.	1
Sensitivity Awareness	Explicit recognition	states one dataset is more sensitive and uses that to guide decisions.	7
	Implicit recognition	behaves differently across scenarios suggesting awareness but does not articulate it.	1
Task Completion	Completed independently	finishes without heavy reliance on external help.	4
	Heavy assistance	depends significantly on AI tools or step-by-step guidance.	3
	Not completed	unable to finish the task within the session.	2
Time Use	Within expected time	completes task within the allocated window.	5
	Over time	exceeds the expected time or needs extra time.	3

B Interview Script

INTERVIEW SCRIPT

"Thanks so much [participant name] for taking the time to talk with us today. My name is [your name], and I am a Master's student at Carnegie Mellon University studying how machine learning developers think about privacy when building and training models. I am here with my colleagues, [Other people's names].

There are no right or wrong answers. We are genuinely interested in your experience and perspective, whatever they entail. The interview should take about 10 to 15 minutes, and there is a lab study after that takes 45-60 minutes. We will be taking notes during the session. The interview will also be recorded to capture your responses for research purposes accurately. Nothing you say will be shared with anyone outside the research team, and your name won't be attached to any findings. As a reminder, please avoid sharing any private or personally identifiable information about yourself or others in your responses. If a question prompts you to think of a specific person or organization, please keep your answers general.

For this study, we ask that you please turn off your camera during the interview, as we only require audio participation. Please also ensure you are in a private space for the duration of the session so that no one nearby is inadvertently recorded. Before we begin, I would like to remind you of your rights as a participant briefly:
Your participation is completely voluntary You may choose not to answer any question You may take a break at any time You may stop the interview at any time without any negative consequences. You may also ask for the recording to be paused or stopped at any point. Do you confirm that you agree to participate in this interview? Do you agree to have this interview audio recorded?
If no: That is completely fine. Thank you for your time. We will stop here
If yes: Thank you. I will now begin the recording.

For documentation purposes, could you please confirm again that you consent to participate in this interview and to have it audio recorded?

(Wait for verbal confirmation on the recording)

SECTION 1: Background

Q1. **To start, can you tell me a little about your background in machine learning and how long you've been working in this space?**

Follow-up probes:
What kinds of projects have you worked on?
What kinds of datasets do you typically work with?
Do you use PyTorch, TensorFlow, Scikit-learn, XGBoost, MONAI, or any other framework?

SECTION 2: General Privacy Practices

Q2. **In simple terms, what would you say is "privacy risks" in the context of training a machine learning model?**

Follow-up probes:
How might you go about protecting the data you're using when building a machine learning model?
At what stages of the development pipeline do you think privacy should be considered (eg, data ingestion, training, or inference)?

Q3. **Tell me about a time you worked with data that might be considered sensitive. What made that dataset sensitive?**

Follow-up probes:
Were any steps taken to protect the data?
If Yes
•Were those steps your decision or part of team guidelines?
If No
• Why wasn't any step taken to protect the data?

SECTION 3: Awareness of PPML Techniques

Q4. **In your own work, have you ever actually integrated a PPML technique at any stage of a pipeline you were building?**

Follow-up probes:
•What has been the main source of your knowledge in PPML?

Q5. **I'm going to mention a few terms. For each one, just tell me how you would explain it to a fellow developer, and whether you've ever used it. There's no pressure, and it's completely fine if these are unfamiliar.** (If any term was mentioned in Q6 above, skip it)
•Differential privacy
• Federated learning
• Synthetic data generation

Probes (if recognized)
Do you know any tools or libraries that support it?
Which of them do you prefer and why? (For people familiar with more than one)

SECTION 4: Perceived Barriers
(For inexperienced participants)
Q6. **Imagine your team decides that your next ML model must include stronger privacy protections during training.**
What challenges do you expect you would run into?

(Wait for a few seconds for their initial response(s). If not mentioned, list the items below to guide participants.
•Implementation difficulty
• Understanding parameters
• Performance or accuracy tradeoffs
• Documentation
• Lack of examples or tutorials

(For experienced participants)
Q6b. **What challenges have you run into implementing PPML? (Wait for a few seconds for their initial response(s), then list those below to guide participants)**
•Implementation difficulty
• Understanding parameters
• Performance or accuracy tradeoffs
• Documentation
• Lack of examples or tutorials

CLOSING
"Those are all my questions. Before I wrap up, is there anything else you'd want to share about how you think about privacy in your work, or anything I didn't ask about that feels relevant?"

C Post-Task Survey

Instructions: Please indicate your level of agreement on each statement using the scales provided.

Dataset for Lifestyle Habit

1. **I was confident in my parameter choices for this scenario.**
   Likert scale: 1 = Strongly Disagree, 5 = Strongly Agree
2. **The scenario was difficult to complete.**
   Likert scale: 1 = Strongly Disagree, 5 = Strongly Agree
3. **I had a clear privacy goal in mind for this scenario.**
   Likert scale: 1 = Strongly Disagree, 5 = Strongly Agree

Dataset for Hospital

4. **I was confident in my parameter choices for this scenario.**
   Likert scale: 1 = Strongly Disagree, 5 = Strongly Agree
5. **This scenario was difficult to complete.**
   Likert scale: 1 = Very Easy, 5 = Very Hard
6. **I had a clear privacy goal in mind for this scenario.**
   Likert scale: 1 = Strongly Disagree, 5 = Strongly Agree

Overall Evaluation

7. **The Opacus documentation was clear and helpful.**
   Likert scale: 1 = Strongly Disagree, 5 = Strongly Agree
8. **I feel I understand what DP parameters do after this task.**
   Likert scale: 1 = Strongly Disagree, 5 = Strongly Agree
9. **I would use DP in a real ML project.**
   Likert scale: 1 = Strongly Disagree, 5 = Strongly Agree

Multi-Select Question

10. **Which resources did you use during this task?** *(Select all that apply)*
    - Opacus documentation
    - PyTorch documentation
    - Web search
    - AI assistant
    - Other tutorials/guides

Open-Ended Questions

11. **If you used other resources, please describe them.** *(Text Response)*
12. **If you used AI tools, please specify which ones.** *(Text Response, e.g., ChatGPT, Copilot, Claude)*

D Screening Survey

We are researchers at Carnegie Mellon University conducting a study on machine learning development practices. This short survey helps us determine if you are eligible to participate in a paid study session. It should take approximately 5 minutes to complete. Participation is voluntary and your responses will be kept confidential and used only for research purposes.

1. **What is your age group?**
   - Under 18
   - 18–24
   - 25–34
   - 35–44
   - 45–54
   - 55 or older
   - Prefer not to answer
2. **What is your gender?**

- Male
- Female
- Prefer not to say

3. **Which of the following best describes your current primary role?**
   - Data Scientist
   - Machine learning Engineer/ Researcher
   - Software Engineer
   - Student
   - Other
4. **What is your current degree program?**
   - Computer Science
   - Data Science / Artificial Intelligence
   - Electrical Engineering
   - Other(please specify)
5. **How many machine learning courses have you taken?**
   - 1 - 2 courses
   - 3 - 5 courses
   - 6 - 10 courses
   - + 10 courses
   - None
6. **Which Programming Languages have you used? (Check all that apply)**
   - Python
   - R
   - Java
   - C++
   - Julia
   - Other
7. **Rate your PyTorch proficiency**
   - Expert
   - Advanced
   - Intermediate
   - Beginner
   - No experience
8. **Are you aware of techniques designed to protect privacy when training machine learning models?**
   - Yes I have used them
   - Yes i know about them, but i have not used them
   - I have heard of them, but don't know what they are
   - No, I was not aware of them
   - Not sure
9. **Which of the following privacy-preserving techniques have you heard of? (Select all that apply)**
   - Federated learning
   - Differential privacy
   - Adaptive Tokenization
   - Synthetic data generation
   - Homomorphic Encryption
   - Secure Multiparty Computation
   - Gradient Compression
   - None
   - Other
10. **For each technique below, indicate your level of familiarity**
    - Never heard of them
    - Heard of them
    - Know what it does
    - Have used it in practice
11. **Which PPML tools or frameworks have you used?(Select all that apply)**
    - Tensorflow Privacy
    - Opacus(PyTorch)
    - PySyft / PyGrid
    - Click IBM Differential Privacy Library
    - None
    - Other

Thank you for completing the screening questions! If you are selected for the study, we will contact you to schedule your session. To do so, we will need your contact details below. This information will only be used for scheduling purposes. If you are not selected, your contact information will not be retained and will be permanently deleted within 30 days of the screening period closing