

Towards Automating Differential Privacy:

How can AI help?

Yu-Xiang Wang

UC San Diego

Outline of the Talk

- A brief history of differential privacy (with connection to statistics)
- Example of Private Data Analysis
- LLMs for automating differential privacy



AOL Search Data: Aug. 2006

AOL: "This was a screw up"

500k User Session Collection

 This collection is distributed as
 Any application of this collection

AOL apologizes for release of user search data

The goal of this collection is to provide real query log data that is based on real users. It should be used for personalization, query
 The data we include (AccountID, Query, QueryTime, IPaddress, ClickID, AccountID - an anonymous user ID number, Query - the query entered by the user, name omitted with most punctuation removed, QueryTime - the time of the search, IPaddress - if the user logged on via ClickID - if the user logged on via the client console is listed.

AOL subscribers sue over data leak

AOL's disturbing glimpse into users' lives

Fallout From AOL's Data Leak Just Beginning

The fallout from AOL's unintended release of personal search data of 658,000 subscribers c
 include fines, lawsuits, and changes in law and policy regarding search queries.

On March, 2006 - 31 May, 2006
 Normalized queries:

A Face Is Exposed for AOL Searcher No. 4417749



Record	162000000
Hospital	1621 Sacred Heart Medical Center in Providence
Admit Type	1 Emergency
Type of Stay	1
Length of Stay	6 days
Discharge date	Oct-2013
Discharge Status	under the care of an hospital service organization
Charges	1 Medicare
Payers	6 Commercial insurance
Emergency Code	6235 Other government services patients E8162 motor vehicle traffic accident due to loss of control; lost control - no injury
Diaagnosis Codes	E9107 closed fracture of other specified part of upper extremity S1851 pulmonary insufficiency following trauma & surgery T7671 hypomagnesemia L407 hypotympanum H9051 tinnitus E8101 acute stress disorder M6400 organic amnesia
Age In Years	60
Gender	Male
RIP	98851
Estate Reside	Non-Hispanic

“Only You, Your Doctor, and Many Others May Know”
L. Sweeney. *Technology Science*, 2015
(Documenting re-identification attacks since 1997)

- 3

Around the same time, in a small Italian mountain town...

- Steve Fienberg and “his guys” (e.g., Abowd, Vilhuber et al) vs Cynthia Dwork and “her guys” (e.g., Smith, Kobbi, McSherry et al)
- Focal point of the fight: “Data utility” vs “Formal privacy”



Bertinoro, Italy, July 2005



A taverna in the town square

“The workshop almost broke down on the second day. The cryptographers thought the statisticians were hopelessly ad hoc; the statisticians thought the cryptographers were completely paranoid.”

-- Cynthia Dwork

(Pictures / account from Sesa Slavkovic’s 2013 article “*Steve the Matchmaker: The Marriage of Statistics and Computer Science in the World of Data Privacy*”)

From this workshop, differential privacy was born

Calibrating Noise to Sensitivity in Private Data Analysis

Cynthia Dwork¹, Frank McSherry¹, Kobbi Nissim², and Adam Smith^{3*}

¹ Microsoft Research, Silicon Valley. {dwork,mcsherry}@microsoft.com

² Ben-Gurion University. kobbi@cs.bgu.ac.il

³ Weizmann Institute of Science. adam.smith@weizmann.ac.il

Abstract. We continue a line of research initiated in [10, 11] on privacy-preserving statistical databases. Consider a trusted server that holds a database of sensitive information. Given a query function f mapping databases to reals, the so-called *true answer* is the result of applying f to the database. To protect privacy, the true answer is perturbed by the addition of random noise generated according to a carefully chosen distribution, and this response, the true answer plus noise, is returned to the user.

Fast forward to 2017 - 2018

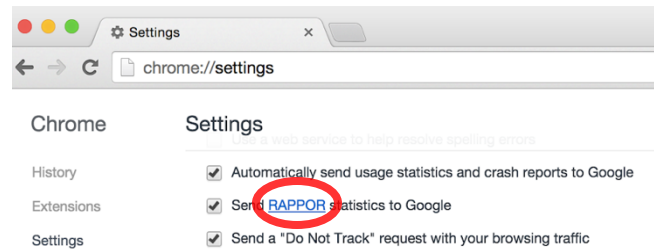
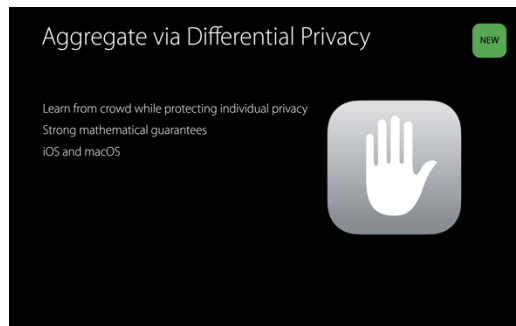
- Creators of DP won Gödel Prize
- “Deep Learning with Differential Privacy” is about to take off (50+ citations then but gaining momentum)
- Census made a pledge to adopt DP for 2020 Census
- A closed-door meeting called “DP-Deployed” was arranged in Harvard University.

Key challenges identified by the 2018 “DP-Deployed” Meeting...

- **Utility loss:** Utility remains the primary issue for small to medium-sized data.
- **Privacy accounting:** There is no standard in selecting, reporting, interpreting privacy parameter ϵ . It is hard to precisely quantify the actual privacy loss due to the slacks in the mathematical analysis.
- **Scalability issue:** The design and analysis of DP mechanisms is delicate and error-prone even for experts [50] — there will never be enough PhDs with DP training to meet the growing demand.
- **Implementation:** There are few high-quality codes for DP, with the exception of PINQ [55], Ektelo[74] and Psi [27], each serving a particular niche.

- Need more refined theory.
- Need better algorithms with more favorable privacy-utility tradeoff.
- Need well-engineered / easy-to-use software.

Differential privacy is transforming into a practical technology!



Privacy-preserving analytics and reporting at LinkedIn



The U.S. Census Bureau
Adopts Differential Privacy

John M. Abowd
Chief Scientist and Associate Director for Research and Methodology
U.S. Census Bureau
2018 International Methodology Symposium
Ottawa, Ontario, Canada
November 9, 2018



As of today, it is becoming one of the hottest topics in ML, Security, Database and in Statistics

Deep learning with differential privacy

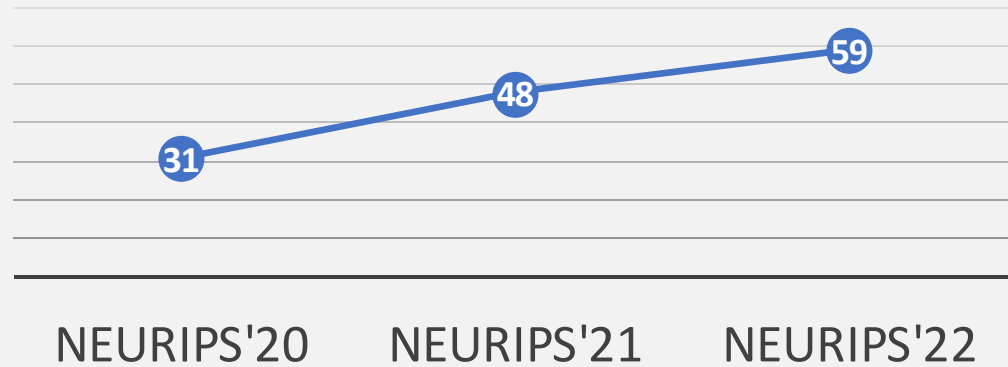
[M Abadi](#), [A Chu](#), [I Goodfellow](#), [HB McMahan](#)... - [Proceedings of the 34th Conference on Neural Information Processing Systems](#)

... We demonstrate the training of **deep** neural network models with a modest total **privacy** loss, computed over entire model.

☆ Save 📄 Cite Cited by 3857 Related articles

(7500+ citations in 03/2025)

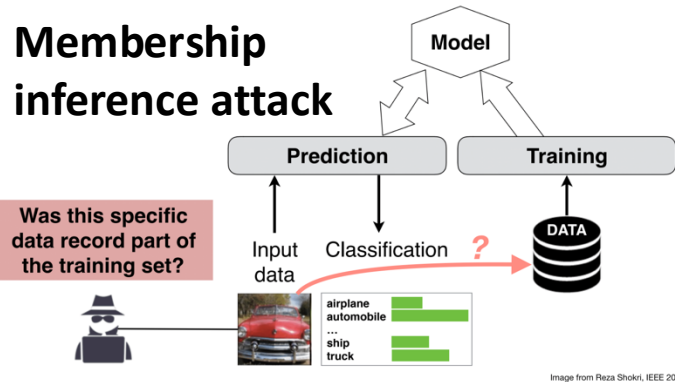
Number of Papers on Differential Privacy



- Selected DP Papers receiving best paper awards
 - **AISTATS'19**: [W., Balle, Kasiviswanathan](#) "Subsampled Renyi DP"
 - **FOCS'20**: [Bun, Livni, Moran](#) on "Private classification"
 - **CCS'21**: [Girgis et al.](#), "Renyi DP of the Shuffle Model"
 - **ICLR'22**: [Steinke and Papernot](#) "Hyperparameter tuning with Renyi DP"
 - Database conferences: **PODS'22**, **CIKM'22**.
 - Also, [Dong et al.](#) "Gaussian Differential Privacy" was a "Discussion paper" at Journal of the Royal Statistical Society

Privacy challenges in the AI Era

Membership inference attack



(Shokri et al., 2016)

Training Set



Caption: Living in the light with Ann Graham Lotz

Generated Image



Prompt: Ann Graham Lotz

(Carlini et al., 2023, "Extracting Training data from Diffusion Models")



Personal health records



Our article with Evidation Health:
<https://www.mobihealthnews.com/news/contributed-when-fitness-data-becomes-research-data-your-privacy-may-be-risk>

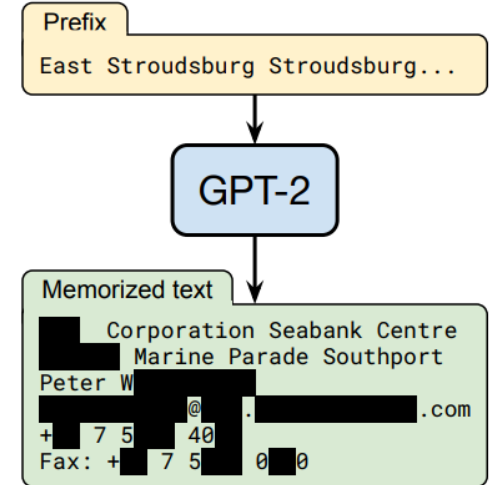
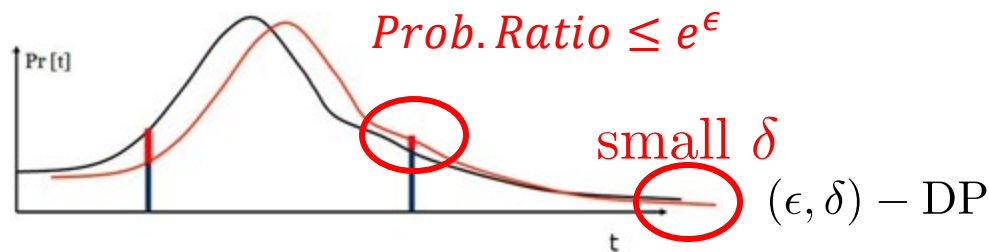


Figure 1: **Our extraction attack.** Given query access to a neural network language model, we extract an individual person's name, email address, phone number, fax number, and physical address. The example in this figure shows information that is all accurate so we redact it to protect privacy.

(Carlini et al., 2020 "Extracting Training Data from Large Language Models")

Differential Privacy provably addresses these challenges.

- GDPR / CCPA, Risk of identifying users, extracting their data *legal compliance of DP is still being debated
- **Differential privacy (DMNS 2006)** is a formal definition of privacy with many good properties.
- The two worlds *with or without* “Alice” are indistinguishable.



Formal definition of DP

- Let Z, Z' be **any two datasets** that differ only by one user, and A is a randomized algorithm. We say A is (ϵ, δ) -DP if for **all output set S**

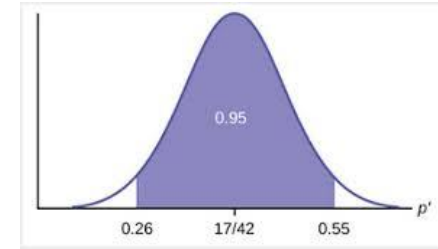
$$P(\mathcal{A}(Z) \in S) \leq e^\epsilon P(\mathcal{A}(Z') \in S) + \delta$$

Example: Take $S = \{\text{Model generates Alice's SSN}\}$.

$Z =$ A dataset *with* Alice's SSN

$Z' =$ A dataset *without* Alice's SSN.

Privacy can be (almost) for free!



- In **mean estimation**, we are interested to see the mean of the population, but all we get is n iid samples.

Statistical error

$$O_P\left(\frac{1}{\sqrt{n}}\right)$$

Error due to DP

$$O_P\left(\frac{1}{n\epsilon}\right)$$

Notes: This is overly simplified

1. Dimension and other constraints may creep in...
2. Utility hurts more for small sample

“DP allows us to learn a lot from population while learning nothing about individuals”. If DP is reliably deployed:

- Solve the unintended memorization
- Solve the copyright issue in AI training
- Allow hospitals to collaborate on medical research
- Allow more automatic data sharing for research
- Many new applications...

20 years later: A DP Workshop in Bertinoro



- Topics on the panel discussion

“Why are people not adopting DP more in practice?”

“How to explain DP guarantees to a layperson?”

“How do we prevent errors in DP proof / DP implementation?”

“Is there another successful family of methods like DP-SGD?”

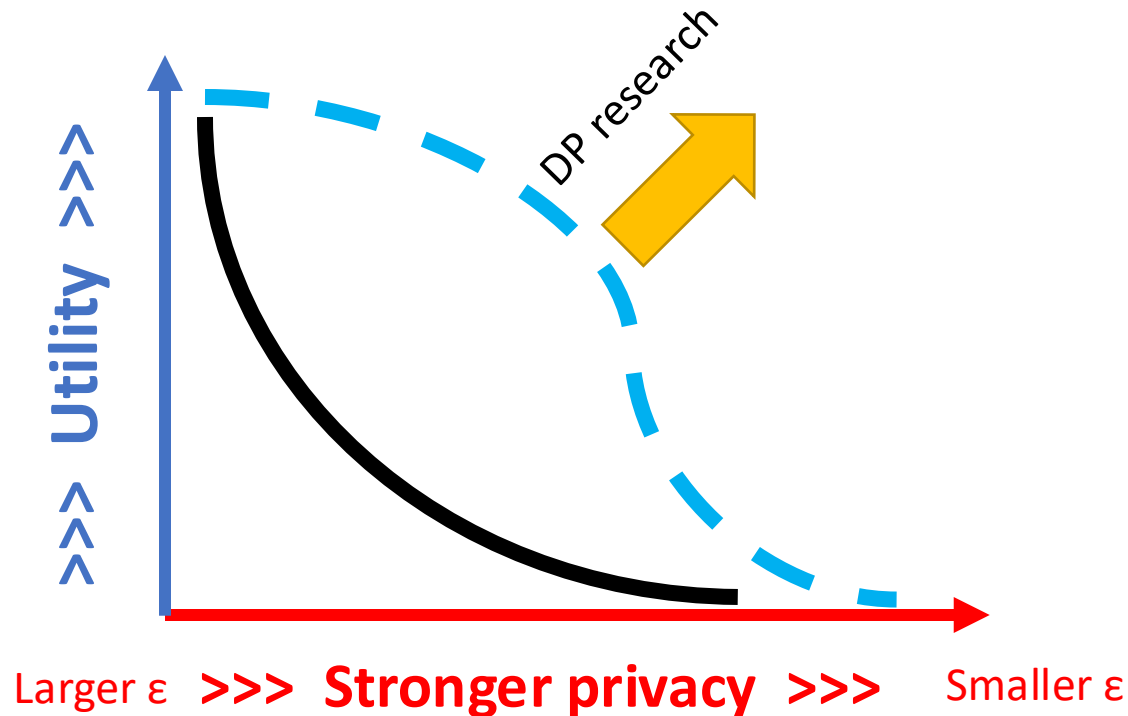
“Is differential privacy failing?”

***Still the same set of problems from 2017/2018.**

To be fair, a lot of progress has been made by the DP community in the last 10 years.

- **More algorithms:** for statistical estimation / inference
- **New primitives:** amplification by X , hyperparameter tuning, etc...
- **Tighter accounting:** zCDP, Renyi DP, PLD, f-DP, GDP, ...
- **New settings:** Private + Public data, attribute/label DP
- **Auditing and attacks:** MIA, Reconstruction etc..
- **More implementations:** tf.privacy, autodp, opacus, OpenDP, etc...
- **Educational materials:** courses, tutorials, workshops

Privacy-Utility Tradeoff remains the central problem in DP research



“How to squeeze out the last drop of utility in DP algorithms?”



- More precise privacy accounting and composition
 - Optimal Gaussian mechanism. [Balle and W., 2018]
 - Sparse Vector technique [Zhu and W., 2020]
 - Amplification by Sampling in RDP: [W., Balle, Kasiviswanathan, 19] [Zhu and W., 19]
 - Analytical Fourier Accountant [Zhu, Dong and W., 22]
- More adaptive: Recipe for Data-dependent DP algorithm design.
 - Exploit the structure of a specific input data. [W., 2018]
 - Generalized Propose-Test-Release [Redberg, Zhu, W., 2022]
- More personalized: Per-instance Privacy certificates.
 - What’s the privacy loss for users in my dataset? [W., 2019]
 - How to privately publish pDP losses [Redberg and W., 2021]

Open-source project:

AutoDP

<https://github.com/yuxiangw/autodp>

pip install autodp

I won't tell you about these today...

All you need to know is that DP is modular there is a growing list of recipes and primitives:

- Calibrating noise to Sensitivity
- Composition
- Amplification by Sampling
- Propose-Test-Release
- Subsample and Aggregate
- Sparse-Vector Techniques

Our new recipe for pure-DP mechanisms

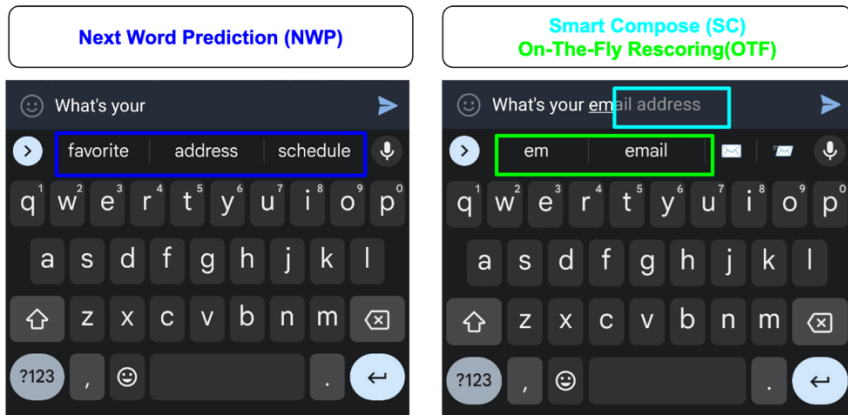
“Purification”: <https://arxiv.org/abs/2503.21071>

(work with Yinyu, Erchi and Yi-An.)

Also not today...

State-of-the-Art DP machine learning methods are impressive but rather involved in their design.

- Key design: Modern accounting with DP-SGD, Large public dataset and pretrained features, adaptive clipping, many other tricks



Spanish Gboard by Google (Xu et al, 2023)

Dataset	Pre-Training Data	Top-1 Accuracy (%)				δ	Section
		$\epsilon = 1$	$\epsilon = 2$	$\epsilon = 4$	$\epsilon = 8$		
CIFAR-10	–	56.8	65.9	73.5	81.4	10^{-5}	3.2
ImageNet	–	–	–	–	32.4	$8 \cdot 10^{-7}$	3.3
CIFAR-10	ImageNet	94.7	95.4	96.1	96.7	10^{-5}	4.1
CIFAR-100	ImageNet	70.3	74.7	79.2	81.8	10^{-5}	4.1
ImageNet	JFT-4B	84.4	85.6	86.0	86.7	$8 \cdot 10^{-7}$	4.2
Places-365	JFT-300M	–	–	–	55.1	$5 \cdot 10^{-7}$	4.3

(De et al. 2021) DeepMind

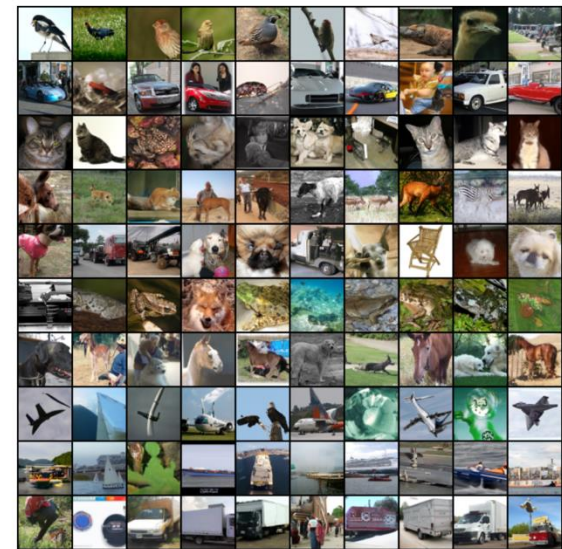


Figure 3: Generated samples on CIFAR10 with $(0.67, 10^{-5})$ -DP. Each row corresponds to one class. FID=7.87. See App. J for real and generated images side-by-side.

(Lin et al. 2025) Microsoft Research

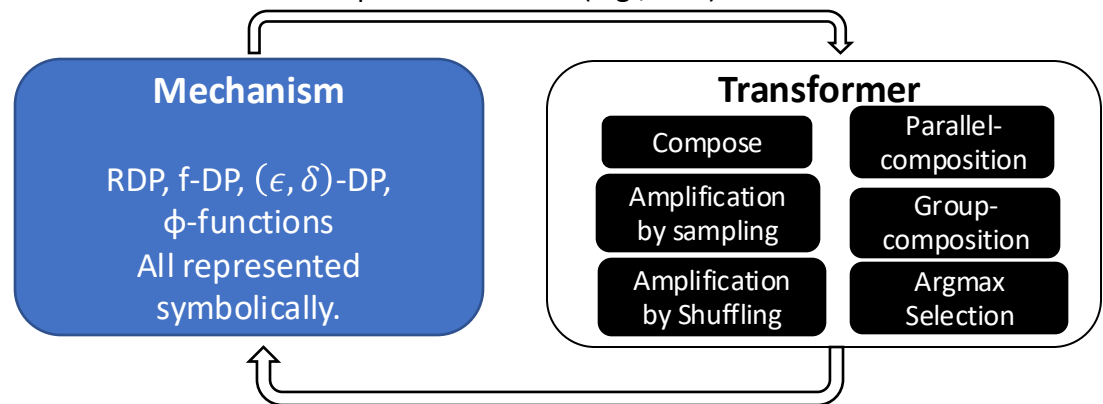
autodp: a flexible and easy-to-use package for differential privacy

Mechanism is the base class that describes a randomized algorithm and its privacy loss.



Calibrator calibrates noise to privacy budget for an arbitrary 'mechanism'

Transformers manipulate functions (e.g., RDP) to create new **Mechanisms**.



1. You bring your mechanism.
2. Describe it in autodp as an Mechanism.

Then autodp takes care of

- Numerical computation of the privacy loss.
- Calibrating noise to privacy requirements.

Open source project:

<https://github.com/yuxiangw/autodp>

pip install autodp

Checkpoint: The State of DP Research

- Crossing the 20 years mark
- Growing community (a vibrant presence among statisticians too)
- Connected to some of the biggest open problems in the AI Era
- High-utility DP in real deployments are possible, but far from being easily accessible.

Outline of the Talk

- A brief history of differential privacy (with connection to statistics)
- Example of Private Data Analysis
- LLMs for automating differential privacy

Can DP-trained students carry out differentially private data analysis?

Differentially Private Telemetry Data Analysis

1. Data Attributes

The telemetry dataset contains synthetic event logs for a variety of product types, with the following attributes:

Attribute	Description
Product Type	A categorical attribute indicating the type of product. Possible values include A, B, C, D, E, F, and Others.
Event Type	The type of event logged. Possible values are: open, close, save, reset, and error.
Time of Event	A timestamp indicating when the event occurred, with values ranging from May 1, 2024, to July 31, 2024.
User ID	Anonymized ID of users who produced these events while using each product.

Goal: Identify products with error rates *significantly higher* than average error rate.

Non-private analysis:

$$\text{ErrorCount}_P = \sum_{i=1}^n \mathbf{1}\{\text{EventType}_i = \text{error and ProductType}_i = P\}$$

$$\text{TotalCount}_P = \sum_{i=1}^n \mathbf{1}\{\text{ProductType}_i = P\}$$

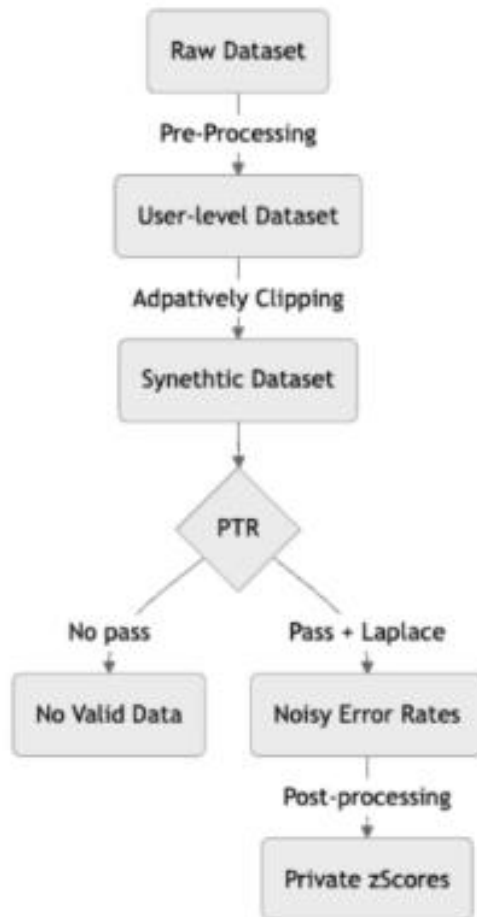
$$\text{ErrorRate}_P = \frac{\text{ErrorCount}_P}{\text{TotalCount}_P}$$

$$\text{MeanErrorRate} = \frac{1}{m} \sum_{P=1}^m \text{ErrorRate}_P$$

$$Z_P = \frac{\text{ErrorRate}_P - \text{MeanErrorRate}}{\text{StandardDeviation}_{\text{ErrorRate}}}$$

Task for students: Release (1) Z-scores (2) products with Z-score > 0. (3) Ensure the algorithm satisfies (2.0, 1e-5)-DP.

4 out of 9 teams came up with valid DP solutions. Here are some examples.



$$\text{ErrorRate}_P = \frac{\text{ErrorCount}_P}{\text{TotalCount}_P}$$

1. Per-product clipping.
2. Laplace / Gaussian mechanism for numerator and denominator separately.

4 out of 9 teams came up with valid DP solutions

	z-score 95%	z-score median	IOU 5%	IOU median	Solution
Team A	0.865	0.1779	0.5	1	per product clipping, laplace noise releasing numerator denominator
Team B	0.553	0.23	0.67	1	adaptive clipping, SVT-PTR, directly add noise to Error-Rates.
Team C	0.145	0.144	1	1	clipping event contribution to 1 across all product type, release noisy histogram, laplace noise. L1 sensitivity
Team D	0.311	0.105	0.67	1	per product clipping, Gaussian noise releasing numerator denominator, nice use of autotp

$$\text{ErrorRate}_P = \frac{\text{ErrorCount}_P}{\text{TotalCount}_P}$$

Team B is the most theoretically interesting, hyperparameter-free

Team C does most aggressive clipping, and the simplest.

Winning strategy is simple and clever

Clipping max-contribution of user (across all product category) to 1.



	Errors	Non-Errors
Product A	125	23096
Product B	767	38325
Product C	30	1581

Add Laplace Noise to this table, then release

$$\frac{\widehat{error}_p}{\widehat{error}_p + \text{non} - \widehat{error}_p}$$

Pure-DP.

Avoid composition all together.

Less numerical issues.

Lessons learned

- Many design choices. Hard to get the best strategy with one-strike even for experts.
- Simpler method works better.
- Constant here and there to be saved and there through
 - Correct representation of the data
 - Modern privacy accounting (e.g., using autotp) helps.

Key challenges identified by the 2018 “DP-Deployed” Meeting...

- **Utility loss:** Utility remains the primary issue for small to medium-sized data.
- **Privacy accounting:** There is no standard in selecting, reporting, interpreting privacy parameter ϵ . It is hard to precisely quantify the actual privacy loss due to the slacks in the mathematical analysis.
- **Scalability issue:** The design and analysis of DP mechanisms is delicate and error-prone even for experts [50] — there will never be enough PhDs with DP training to meet the growing demand.
- **Implementation:** There are few high-quality codes for DP, with the exception of PINQ [55], Ektelo[74] and Psi [27], each serving a particular niche.

~~—Need more refined theory.~~

~~—Need better algorithms with more favorable privacy-utility tradeoff.~~

~~—Need well-engineered / easy-to-use software.~~

We can never train enough PhD students to meet the growing demand.

Even experts can easily get things wrong in DP: 6 versions of SVT from [Lyu et al, 16]

Algorithm 1 An instantiation of the SVT proposed in this paper.

Input: $D, Q, \Delta, T = T_1, T_2, \dots, c.$

```

1:  $\epsilon_1 = \epsilon/2, \rho = \text{Lap}(\Delta/\epsilon_1)$ 
2:  $\epsilon_2 = \epsilon - \epsilon_1, \text{count} = 0$ 
3: for each query  $q_i \in Q$  do
4:    $\nu_i = \text{Lap}(2c\Delta/\epsilon_2)$ 
5:   if  $q_i(D) + \nu_i \geq T_i + \rho$  then
6:     Output  $a_i = \top$ 
7:     count = count + 1, Abort if count  $\geq c.$ 
8:   else
9:     Output  $a_i = \perp$ 
```

Algorithm 2 SVT in Dwork and Roth 2014 [8].

Input: $D, Q, \Delta, T, c.$

```

1:  $\epsilon_1 = \epsilon/2, \rho = \text{Lap}(c\Delta/\epsilon_1)$ 
2:  $\epsilon_2 = \epsilon - \epsilon_1, \text{count} = 0$ 
```

Correct but adding more noise
than needed

```

/:   count = count + 1, Abort if count  $\geq c.$ 
8:   else
9:     Output  $a_i = \perp$ 
```

Algorithm 3 SVT in Roth's 2011 Lecture Notes [15].

Input: $D, Q, \Delta, T, c.$

```

1:  $\epsilon_1 = \epsilon/2, \rho = \text{Lap}(\Delta/\epsilon_1),$ 
2:  $\epsilon_2 = \epsilon - \epsilon_1, \text{count} = 0$ 
3: for each query  $q_i \in Q$  do
4:    $\nu_i = \text{Lap}(c\Delta/\epsilon_2)$ 
5:   if  $q_i(D) + \nu_i \geq T_i + \rho$  then
6:     Output  $a_i = q_i(D) + \nu_i$ 
7:     count = count + 1, Abort if count  $\geq c.$ 
8:   else
9:     Output  $a_i = \perp$ 
```

Algorithm 4 SVT in Lee and Clifton 2014 [13].

Input: $D, Q, \Delta, T, c.$

```

1:  $\epsilon_1 = \epsilon/4, \rho = \text{Lap}(\Delta/\epsilon_1)$ 
2:  $\epsilon_2 = \epsilon - \epsilon_1, \text{count} = 0$ 
3: for each query  $q_i \in Q$  do
4:    $\nu_i = \text{Lap}(\Delta/\epsilon_2)$ 
5:   if  $q_i(D) + \nu_i \geq T_i + \rho$  then
6:     Output  $a_i = \top$ 
7:     count = count + 1, Abort if count  $\geq c.$ 
8:   else
9:     Output  $a_i = \perp$ 
```

Algorithm 5 SVT in Stoddard et al. 2014 [18].

Input: $D, Q, \Delta, T.$

```

1:  $\epsilon_1 = \epsilon/2, \rho = \text{Lap}(\Delta/\epsilon_1)$ 
2:  $\epsilon_2 = \epsilon - \epsilon_1$ 
3: for each query  $q_i \in Q$  do
4:    $\nu_i = 0$ 
5:   if  $q_i(D) + \nu_i \geq T_i + \rho$  then
6:     Output  $a_i = \top$ 
7:
8:   else
9:     Output  $a_i = \perp$ 
```

Algorithm 6 SVT in Chen et al. 2015 [1].

Input: $D, Q, \Delta, T = T_1, T_2, \dots.$

```

1:  $\epsilon_1 = \epsilon/2, \rho = \text{Lap}(\Delta/\epsilon_1)$ 
2:  $\epsilon_2 = \epsilon - \epsilon_1$ 
3: for each query  $q_i \in Q$  do
4:    $\nu_i = \text{Lap}(\Delta/\epsilon_2)$ 
5:   if  $q_i(D) + \nu_i \geq T_i + \rho$  then
6:     Output  $a_i = \top$ 
7:
8:   else
9:     Output  $a_i = \perp$ 
```

Typical modes of errors in published DP papers

- Errors in proofs
- Errors in the algorithms
- Errors in code repo
 - Not actually adding noise
 - Adding too little noise
 - Used private data in the clear some way
- Dependence on previously published papers / code repo that have issues above

Fun facts: These errors (appear to) boost performance!

Challenges in maintaining an open-source community

The OpenDP White Paper*

The OpenDP Team[†]

May 11, 2020

A framework for ensuring that measurements and transformations have their claimed privacy or stability properties, by only allowing procedures that have been vetted (e.g. through the OpenDP editorial board reviewing a written proof) the right to directly construct measurements and transformations.

Outline of the Talk

- A brief history of differential privacy (with connection to statistics)
- Example of Private Data Analysis
- LLMs for automating differential privacy

Can we use Large Language Models for automating differential privacy?

Task: Is Minecraft is *more addictive* than an average video game?

Data Format: logs of a users playing each type of games

Non-private code:

```
Events.filter(game in "game_list").map(x ->  
x.duration >  
7200).groupby(game).average().zscore()
```

Request: Could you design and implement an ϵ -differentially private version of this data analysis? Set $\epsilon = 1.0$.



Sure! Let me help you with that!

Here's an 1.0-DP algorithm implemented in Python

[... Code block in python ...]

I have prepared a proof of DP in latex. [[Download](#)]

What do we do with this LLM generated DP algorithm?

- Does it even compile?
- Does it really satisfy 1.0-DP?
- How do I check the proof?
- How do I make sure the proof is about this algorithm?
- Even if it does satisfy 1.0-DP, does it provide any utility?



Doing “Human Verification” defeats the purpose.



Sure! Let me help you with that!

Here's an 1.0-DP algorithm
implemented in Python

[\[... Code block in python ...\]](#)

I have prepared a proof of DP in
latex. [\[Download\]](#)

How about using LLM to checks its own proof?

- i.e., LLM-as-a-judge
- It was shown to work, e.g. in winning Gold at the International Mathematics Olympiad.

Huang and Yang (2025) Gemini 2.5 Pro Capable of Winning Gold at IMO 2025: <https://arxiv.org/abs/2507.15855>

- Our ongoing work: Can LLMs check DP?
 - with Ruihan, Erchi, Pengrun and Kamalika



Preliminary work: Can LLMs check if an algorithm satisfies DP?

- Prompt:

f``[Description of the function / algorithm $A(X)$]
Does $A(x)$ Satisfy ϵ -DP?``

- Random draw from a function bank of interest:
 - Crazy-looking composition of a variety of functions (mostly statistical queries)
- The actual actual mechanism are
 - “Laplace Mechanism”: Pure-DP
 - “Gaussian Mechanism”: GDP and zCDP
 - “Private Selection”: Gumbel, Laplace, Exponential
- Negative examples are created by misspecifying the **sensitivity**

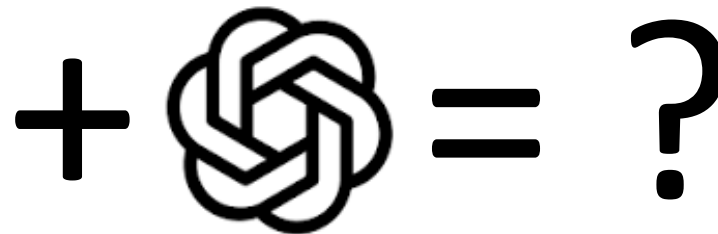
Preliminary work: Can LLMs check if an algorithm satisfies DP?

		Laplace Mechanism	Gaussian Mechanism (GDP)	Gaussian Mechanism (z-CDP)	Private Selection (gumbel noise)	Private Selection (laplace noise)	Private Selection (exponential noise)
open-source models with thinking / reasoning	Qwen3-30B-Think	98.0%	66.3%	81.6%	57.1%	51.0%	56.1%
open-source models trained towards math	DeepSeek-Prover-V2	96.9%	83.7%	93.9%	45.9%	87.8%	39.8%
Closed-source models without thinking	gpt5 (reasoning=minimal)	94.9%	75.5%	81.6%	49.0%	52.0%	83.7%
	gemini2.5 (flash)	98.0%	76.5%	96.9%	67.3%	94.9%	73.5%
	Claude (Sonnet)	95.9%	82.7%	92.9%	57.1%	85.7%	58.2%
Closed-source models with thinking / reasoning	gpt-5 (reasoning=high)	98.0%	92.9%	98.0%	98.0%	98.0%	98.0%

Key messages from these experiments

- GPT5 in high-thinking mode is doing quite well (but expensive!)
- As the mechanisms become less common (from the literature LLMs trained on) performance drops significantly.
- This is a tiny building block of checking DP-proofs (basically deriving sensitivity)

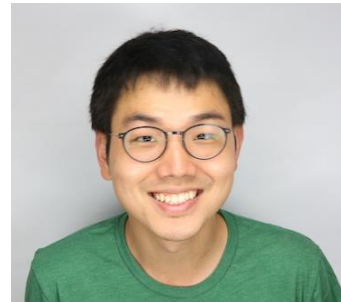
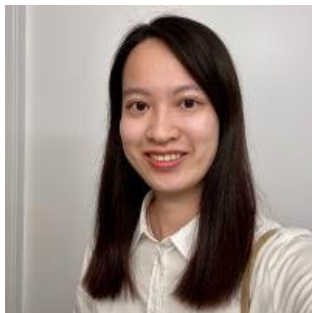
How about semi-automated verification?



Preliminary work: “*DPCheatSheet*: Using Worked and Erroneous LLM-usage Examples to Scaffold Differential Privacy Implementation”

- Q1: Can a layperson with no DP training use LLM *interactively* to carry out private data analysis?
- Q2: Can an LLM helps novices to learn DP?

Work with Shao-Yu Chu, Yuhe Tian, Haojian Jin



We recruited 8 participants to prompt the LLMs to code in DP for there statistical/data science tasks.

Three tasks were considered: (1) “Average” (2) “Z-score”, (3) “ANOVA”

Table 1. Summary of participants’ demographics of our needfinding study.

ID	Group	Occupation	Dept.	Age	Gender	Aware DP	Small ϵ	DP Tech.	DP Res. Exp.
N1	Novice	Undergrad	Data Sci.	18–24	Female	✓	×	×	Never
N2	Novice	Master’s	Data Sci.	18–24	Male	×	×	×	Never
N3	Novice	Undergrad	Data Sci.	18–24	Male	✓	×	×	Never
N4	Novice	Undergrad	Data Sci.	18–24	Male	✓	✓	×	Never
N5	Novice	Master’s	Com. Sci.	18–24	Male	✓	×	×	Never
N6	Novice	Undergrad	Data Sci.	18–24	Male	×	×	×	Never
E1	Expert	PhD	Data Sci.	18–24	Male	✓	✓	✓	2–3 years
E2	Expert	PhD	Data Sci.	25–34	Male	✓	✓	✓	2–3 years
E3	Expert	Professor	Data Sci.	35–44	Male	✓	✓	✓	>3 years

All six novices failed, while all experts succeeded.

Table 2. We identified three main barriers participants encountered when programming differentially private data analysis with LLMs. We present the complete codebook in Appendix A.3.

Barrier Category	Issues	Examples
Struggles in setting DP configurations	Vague description for privacy requirements	(Prompt ChatGPT:) “Propose a solution using differential privacy” - N2
Unable to verify LLM-generated solutions	Require further clarification	“I should compute the maximum change (to set the sensitivity). But should it be a theoretical maximum change or the maximum change when one unit is added or removed from our current dataset?” - N1
	Information overload	“There are too many things to figure out when implementing the program, so I could not find answers to all the questions in my mind.” - N3
	Confused by program outputs	“Why are all the error rates 0s or 1s?” - N1 “The noisy z-score is pretty close to the actual z-score. Is that protecting user privacy?” - N3
Difficulty in guiding LLMs to effective solutions	Lost in back-and-forth exploration	“Wait, why are the recapped steps adding noise to the counts. It was previously adding noise to the error rates.” - N1

On the other hand, experts were able to ping point the issues using the relevant “jargon”, and efficiently helps the LLM to fix the code.

We also tried to ask the LLM to check its own implementation for mistakes.

Table 3. LLMs are error-prone when implementing differentially private analysis. We tested GPT-4o on three tasks (5 trials each), and none of the programs satisfied DP. Even after explicitly listing common implementation mistakes and instructing the model to review and correct its code, the revised z-score and ANOVA programs still failed to satisfy DP.

Trial	Average				Z-score				ANOVA			
	Init. Impl.		Post Self-check		Init. Impl.		Post Self-check		Init. Impl.		Post Self-check	
	Corr.	Mistake	Corr.	Mistake	Corr.	Mistake	Corr.	Mistake	Corr.	Mistake	Corr.	Mistake
#1	×	1, 2	✓	–	×	5	×	5	×	5	×	5
#2	×	3	✓	–	×	1, 5	×	1, 5	×	2	×	5
#3	×	3	✓	–	×	1, 5	×	1, 5	×	5	×	5
#4	×	1, 4	✓	–	×	5	×	5	×	1, 5	×	1, 5
#5	×	2, 4	✓	–	×	5	×	5	×	3	×	2

1. Misused Sensitivity 2. Unprivatized data-dependent hyper-parameters 3. Only partially privatized
 4. Overly noisy result 5. Exceeded privacy budget

- No initial implementations are correct.
- Self-check only provides limited benefit, even after we provide a list of common mistakes.

How can we help the novices?

- Idea 1: debugging Checklist
- This makes sure that the problems raised by novices are relevant.

Debugging Checklist

Execution

☐ Does executing the program result in any errors?

Analysis

☐ Does the program perform the analysis incorrectly?

Privacy Protection

Does the program fail to achieve differential privacy?

☒ Is any path of the computation graph unprotected by noise?

``unbounded_sum`` is not protected by noise

☐ Is the sensitivity incorrectly set, failing to reflect the maximum possible change in the result when a single privacy unit is added or removed?

☐ If noise is injected at multiple points in the computation, does the total privacy budget exceed the overall budget you specified?

☐ Is any hyperparameter (e.g., privacy budget) set by directly querying the data?

Does the program fail to produce good utility under the specified privacy budget?

☐ Are the added noise overwhelming the true result, leading to poor accuracy?

Fix these errors

Idea 2: Training the Novices with Positive and Negative Examples

DPCheatSheet: Using Worked and Erroneous LLM-usage Examples to Scaffold Differential Privacy Implementation¹

Worked Example

(a) Initial Code Snippet

```
sensitivity = visits_per_visitor.max()
```

(b) Expert's Reasoning

"Now, the sensitivity is the maximum number of visitors per visitor. However, it is set by querying the data, which would violate the privacy guarantee!"

"Should I spend some privacy budget on this? Or should I set a data-independent value?"

"This dataset covers the data for a week. Maybe setting the threshold at five would be reasonable. Few people should visit the same restaurant more than 5 times a week."

(c) Expert's Prompt to LLM

ChatGPT

(What is wrong) (How to fix it)

The sensitivity should not depend on the data. Apply clipping to ensure a bound on the maximum change from adding or removing one visitor. Let's clip it to 5.

Exactly — excellent point again!
In proper differential privacy, sensitivity must not depend on the dataset. We need a fixed, public bound — not one computed from private data.
✓ So let's clip the number of visits per visitor to 5. If a visitor appears more than 5 times, we randomly keep only 5 of their visits. This way, the sensitivity becomes exactly 5 — no matter what.

```
# Step 1: Clip visits per visitor to a maximum of 5 visits
max_visits_per_visitor = 5

# Shuffle visits to randomly select up to 5 per visitor
df_shuffled = df.sample(frac=1)

# Keep at most 5 visits per VisitorId
df_clipped = df_shuffled.groupby('VisitorId').head(max_visits_per_visitor)

# Step 2: Compute counts after clipping
total_visits = len(df_clipped)
long_visits = len(df_clipped[df_clipped['Time spent (minutes)'] > 60])

# Step 3: Apply Laplace noise with known sensitivity
epsilon = 1.0
sensitivity = max_visits_per_visitor # Now public and fixed
noisy_long_visits = long_visits + np.random.laplace(loc=0, scale=sensitivity / epsilon)
noisy_total_visits = total_visits + np.random.laplace(loc=0, scale=sensitivity / epsilon)

# Clip to valid counts
noisy_long_visits = max(noisy_long_visits, 0)
noisy_total_visits = max(noisy_total_visits, 0)

# Step 4: Compute private proportion
private_proportion = noisy_long_visits / noisy_total_visits
```

(d) LLM's Output

Erroneous Example

(a) Incorrect Code Snippet

```
import pandas as pd
import numpy as np

# Load dataset
df = pd.read_csv('restaurant.csv')

# Step 1: Identify long visits
df['LongVisit'] = df['Time spent (minutes)'] > 60

# Step 2: Compute sensitivity: max number of visits by any visitor
visits_per_visitor = df.groupby('VisitorId').size()
sensitivity = visits_per_visitor.max()

# Step 3: True counts
true_count_long = df['LongVisit'].sum()
true_count_total = len(df)

# Step 4: Differential privacy with epsilon split
epsilon = 1.0
epsilon_long = epsilon_total = epsilon / 2 # Split budget equally

# Add Laplace noise
noisy_count_long = true_count_long + np.random.laplace(loc=0, scale=sensitivity / epsilon_long)
noisy_count_total = true_count_total + np.random.laplace(loc=0, scale=sensitivity / epsilon_total)

# Step 5: Compute private proportion (clip result)
if noisy_count_total <= 0:
    private_proportion = 0
else:
    private_proportion = noisy_count_long / noisy_count_total
private_proportion = min(max(private_proportion, 0), 1)

# Output
print(f"Private proportion of long visits (>60 minutes): {private_proportion:.4f}")
```

(b) Error Choices

The sensitivity does not account for the maximum change in the query result when a single privacy unit's data is added or removed. ☐

The sensitivity is set by querying the data without privacy protection. ☐

Exceeded privacy budget. ☐

(c) Solution & Explanation

Answer: The sensitivity is set by querying the data without privacy protection.

The sensitivity is set by querying the maximum number of visits a single visitor contributes. Setting hyperparameters by querying the data without privacy protection can leak sensitive information!

A recommended fix is to set the sensitivities using data-independent values.

Fig. 2. Inspired by learning science principles, DPCheatSheet includes a **Worked Example** (left), which demonstrates an expert's step-by-step problem-solving approach, and **Erroneous Examples** (right), which let learners practice evaluating choices at each key decision point.

We compare students learning DP using “DPCheatSheet” vs students using conventional materials.

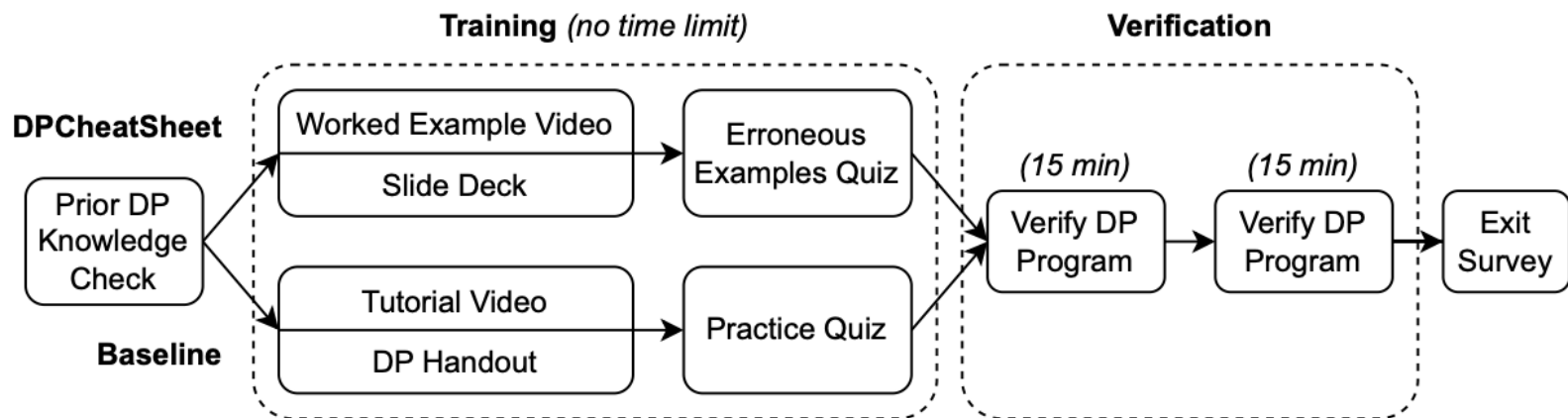
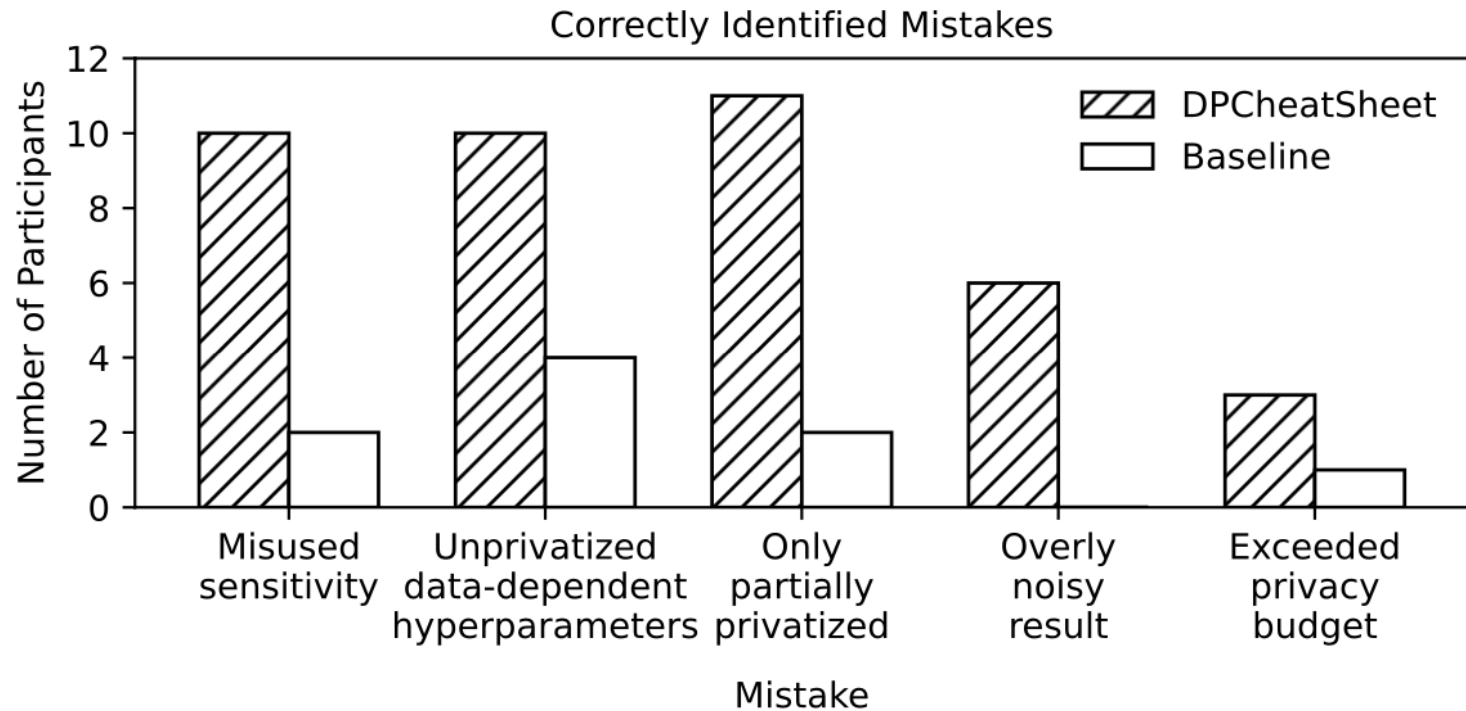


Fig. 3. Error identification study procedure overview.

DPHeatSheet helped novices to more effectively identify mistakes.



DP CheatSheet also helped students to more effectively use LLMs for DP.



ID	Research Question	Analysis Method
P1	Does different most-used platforms (specifically Tiktok and Instagram) have different effect on academic performance?	Z-test
P2	Does longer sleep duration positively influence mental health among students?	Linear Regression
P3	Does a higher addiction score negatively influence academic performance among students?	Linear Regression
P4	How does social media usage relate to students' mental health across different academic levels?	Multiple Linear Regression
P5	Does higher average social media usage negatively affect average sleep hours per night?	Linear Regression

ID	Time	# Turns	Fixed Errors	Remaining Errors	% Errors Fixed
P1	23.1	6	Only partially privatized	Exceeded privacy budget	50%
P2	44.3	19	Misused sensitivity, Overly noisy, Exceeded privacy budget	-	100%
P3	45.2	14	Only partially privatized, Exceeded privacy budget	-	100%
P4	25.8	8	-	Misused sensitivity	0%
P5	45.5	6	Overly noisy result, Exceeded privacy budget	Exceeded privacy budget	67%

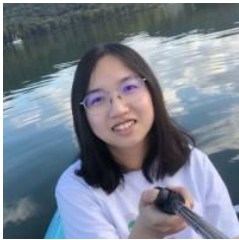
Future work

- Using LLMs for automatically check research-level DP papers on Arxiv and the corresponding code repo.
- Finetune (or RLVR) LLMs for this task. Check if the AI-Agent idea are possible.
- Look into formal verification with Lean.
- Automatically explore the DP-design space in search for high utility algorithms.

Thank you for your attention!

References and Acknowledgment

- “Can LLMs check differential privacy?”



- “DPCheatSheet: Using Worked and Erroneous LLM-usage Examples to Scaffold Differential Privacy Implementation”

