PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

Auditing and Mitigating Safety Risks in Large Language Models

Niloofar (Fatemeh) Mireshghallah

niloofar@cs.washington.edu

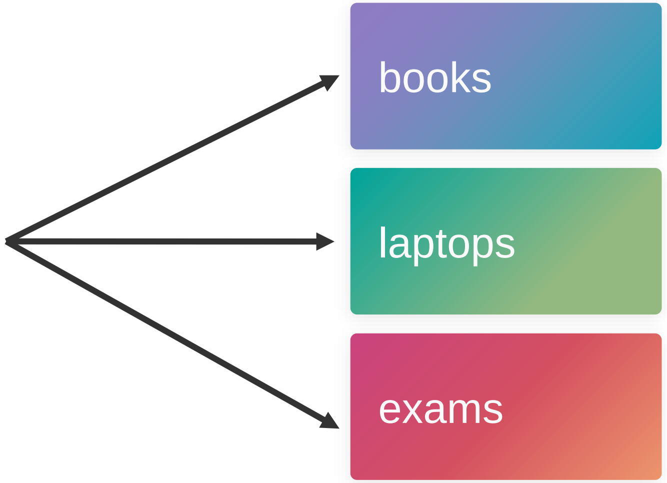
Novermber 2023

Act 1: The LLM Takeover?

What are Language Models?

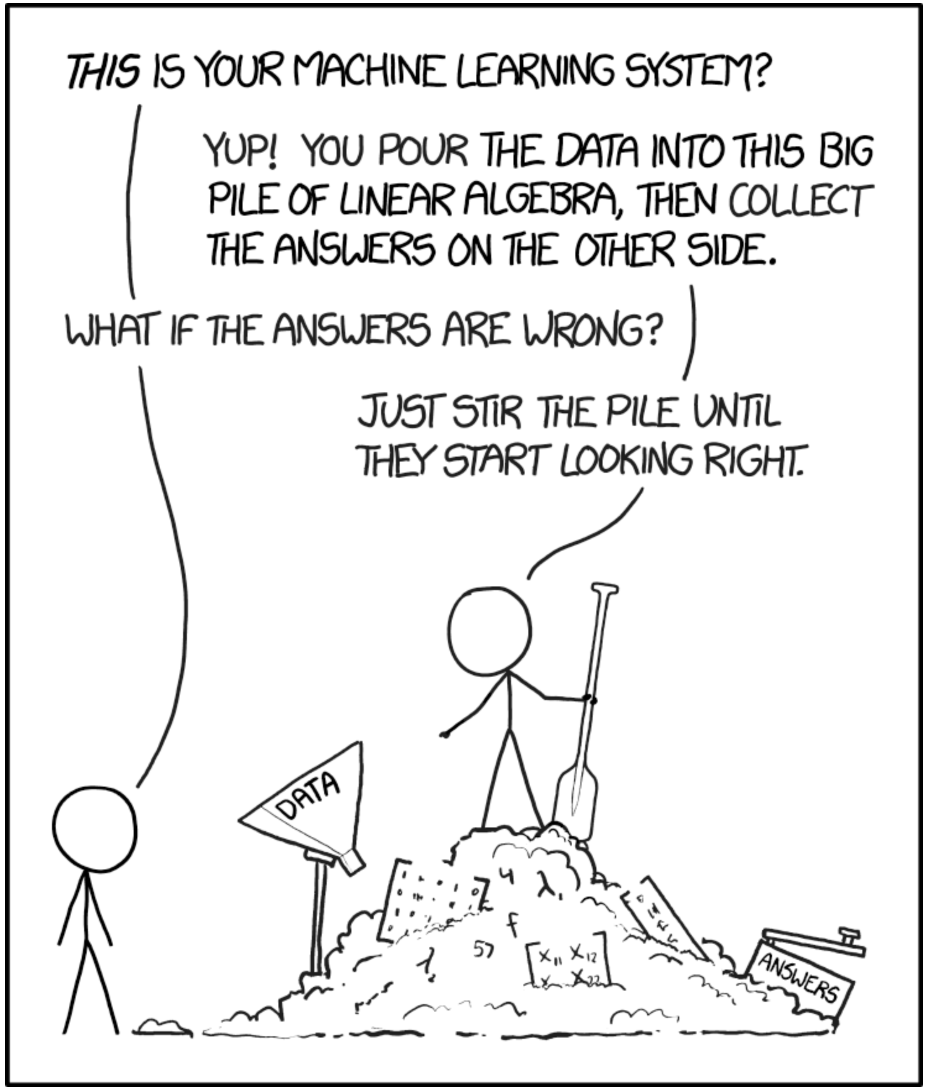
- A language model is a **probability distribution** over sequences of words
- Model what words a given word/context normally appears with

The students opened their ________.



Large Language Models (LLMs)

- Transformer-based language models are often referred to as 'Large LMs' due to their **parameter count** (ranging from 100s of million to billions of parameters)
- Deployed with **Pre-train** and **Fine-tune** paradigm



xkcd.com/1838/

Large Language Models: The Good and the Bad …

- Large language models are very good at **generating text**


ChatGPT passes exams from law and business schools

By Samantha Murphy Kelly, CNN Business
Updated 1:35 PM EST, Thu January 26, 2023

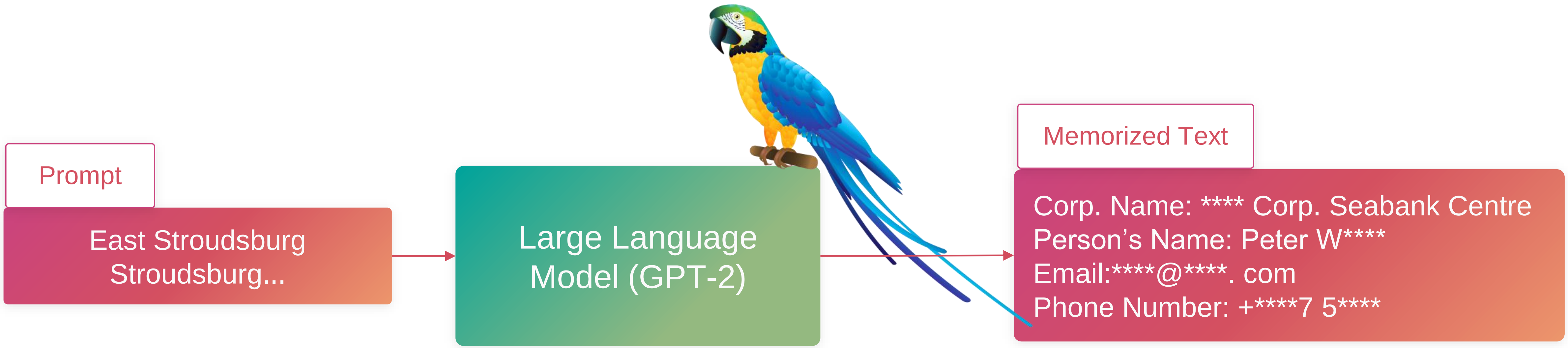
Large Language Models: The Good and the Bad …

- Large language models are very good at **generating text** and **learning representations**.

LLMs: The Bad

xkcd.com/2169/

LLMs: The Bad



Carlini et al. Extracting Training Data from Large Language Models. USENIX SEC 2021.

LLMs: The Bad

- LLMs can also *regurgitate data* they have seen before, creating *privacy risks*.

```
Title:
    Hi everyone, my name is Anish Athalye and I'm a PhD student at
    Stanford University.
```

LLMs: The Bad

- LLMs can also *regurgitate data* they have seen before, creating *privacy risks*.

```
Title:
    Hi everyone, my name is Anish Athalye and I'm a PhD student at
    Stanford University.
```

https://www.anish.io

Anish Athalye

I am a PhD student at MIT in the PDOS group. I'm interested in formal verification, systems, security, and machine learning.

GitHub: @anishathalye   Blog: anishathalye.com

Responses generated by Copilot Feb 8th

LLMs are *not ready* to be widely deployed in *safety critical scenarios* as is!

In this talk:

Question 1: How can we **audit and quantify safety risks** of LLMs?

- [ACL 2023] Membership Inference Attacks via Neighbourhood Comparison
- [EMNLP2022a] Quantifying Privacy Risks of Masked Language Models Using MIAs
- [EMNLP2022b] Memorization in NLP Fine-tuning Methods
- [FAccT2022] What does it mean for language models to preserve privacy?

Question 2: How can we **limit the risks** of LLMs?

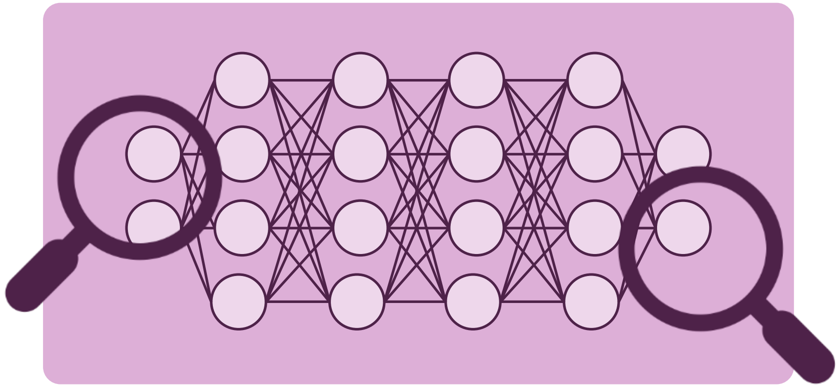
- [ACL2023] Privacy-Preserving Domain Adaptation of Semantic Parsers
- [NeurIPS2022] Differentially private model compression
- [NAACL2021] Joint privacy-utility optimization in language models

Don’t repeat this!!

Act II: Auditing LLMs for Privacy

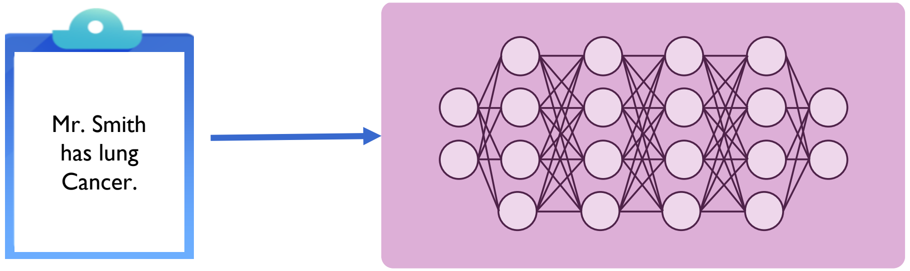
What is information leakage in an ML model?

- ‘Leakage’ is being able to **learn information about the training data**, which cannot be learned from other models/data (from the same distribution)



Measuring Leakage: Membership Inference Attacks

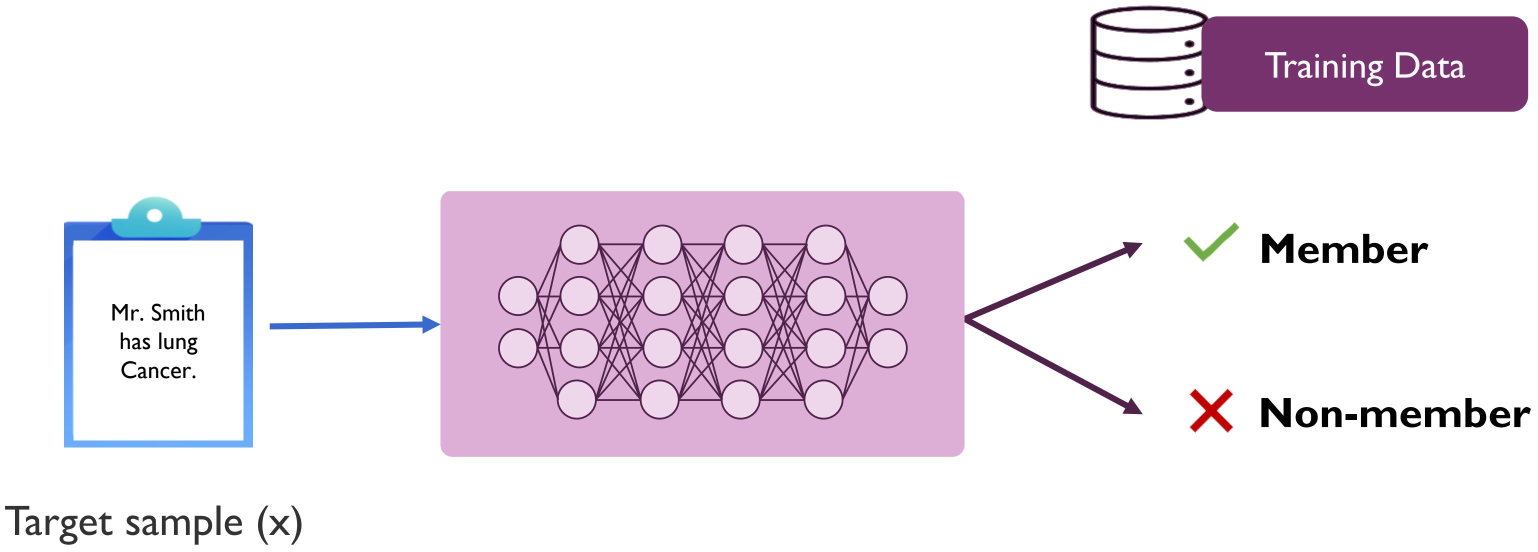
- Can an adversary infer whether a particular data point “x” is part of the training set?



Target sample (x)

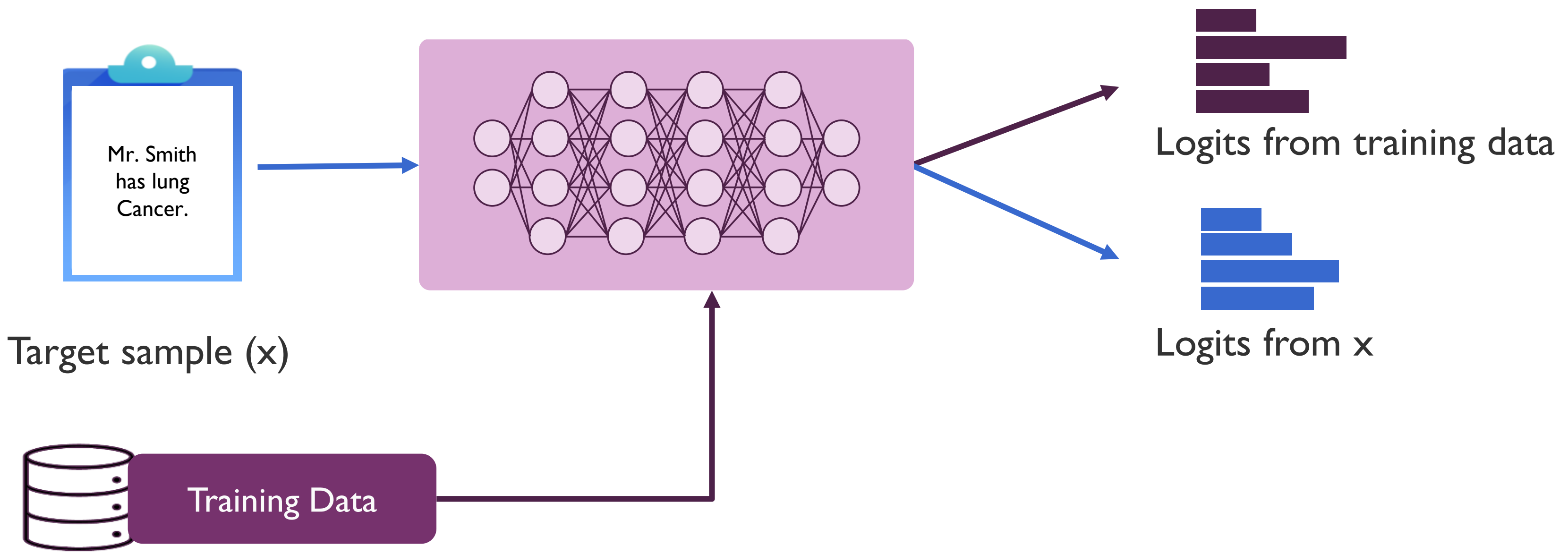
Measuring Leakage: Membership Inference Attacks

- Can an adversary infer whether a particular data point "x" is part of the training set?



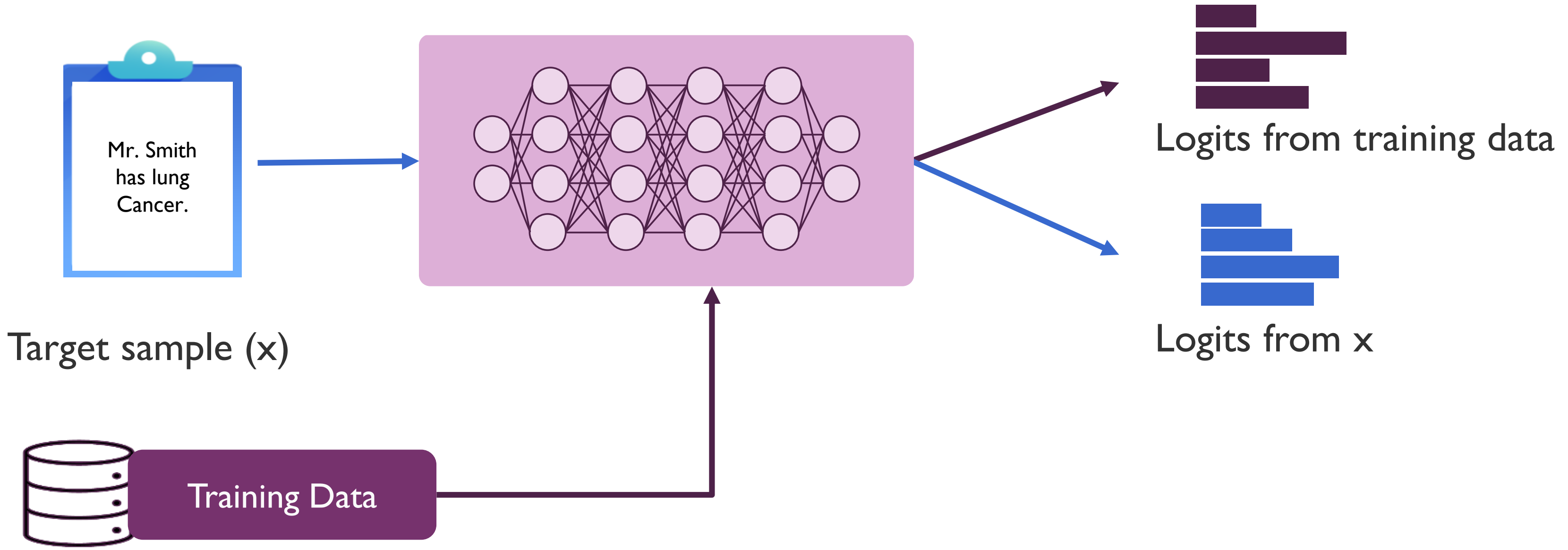
Measuring Leakage: Membership Inference Attacks

- Can an adversary infer whether a particular data point “x” is part of the training set?



Measuring Leakage: Membership Inference Attacks

- Can an adversary infer whether a particular data point “x” is part of the training set?
- Success of attacker is a metric to quantify information leakage of the model about its individual training data



Background: Membership Inference Attacks

- Membership Inference Attacks (MIAs): Loss-based attack
- Stronger MIAs: Reference-based attacks (MIA) [Mireshghallah2022, Ye2021,Carlini2022]
  - A static, absolute threshold does not control for the intrinsic complexity of each utterance
  - We need to calibrate the threshold for each utterance

Reference-based attack

We propose a reference-based attack:

- Complex training points: points that have **higher loss**

Reference-based attack

We propose a reference-based attack:

- Complex training points: points that have **higher loss**

Training data point	Target Model Loss
Mr. Smith has type 2 diabetes.	3
Mr. Smith has fever .	2
Mr. Smith is taking 5 mgs of Haloperidol 2 times a day.	7

Mireshghallah et al. Quantifying Privacy Risks of Masked Language Models Using MIAs. EMNLP 2022

Reference-based attack

We propose a reference-based attack:

- Complex training points: points that have **higher loss**

Training data point	Target Model Loss
Mr. Smith has type 2 diabetes.	3
Mr. Smith has fever .	2
Mr. Smith is taking 5 mgs of Haloperidol 2 times a day.	7

Mireshghallah et al. Quantifying Privacy Risks of Masked Language Models Using MIAs. EMNLP 2022

Reference-based attack

We propose a reference-based attack:

- Complex training points: points that have **higher loss**
- We use a **reference** model, to provide an insight into **how difficult each data point is**

Training data point	Target Model Loss
Mr. Smith has type 2 diabetes.	3
Mr. Smith has fever .	2
Mr. Smith is taking 5 mgs of Haloperidol 2 times a day.	7

Mireshghallah et al. Quantifying Privacy Risks of Masked Language Models Using MIAs. EMNLP 2022

Reference-based attack

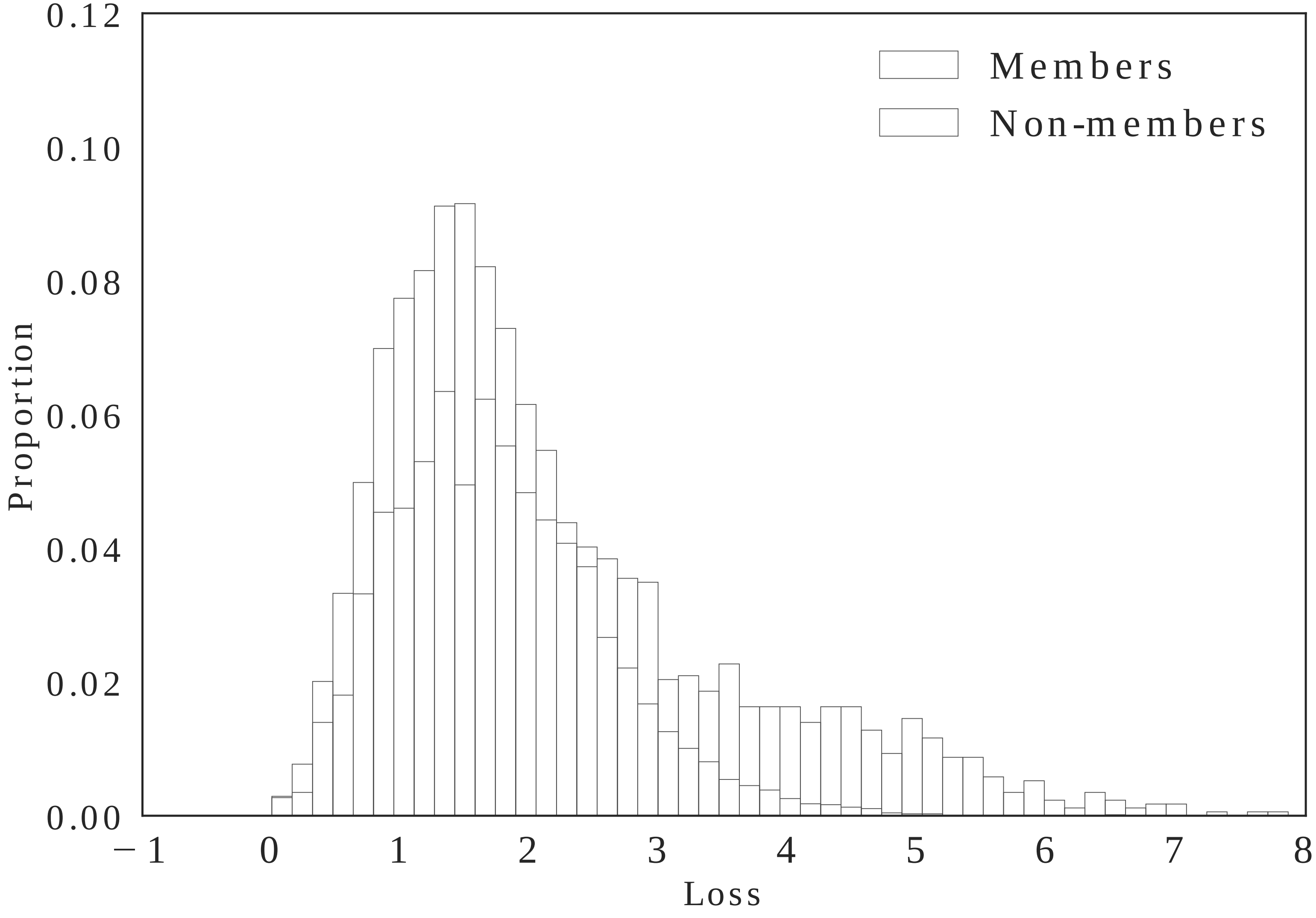
We propose a reference-based attack:

- Complex training points: points that have **higher loss**
- We use a **reference** model, to provide an insight into **how difficult each data point is**

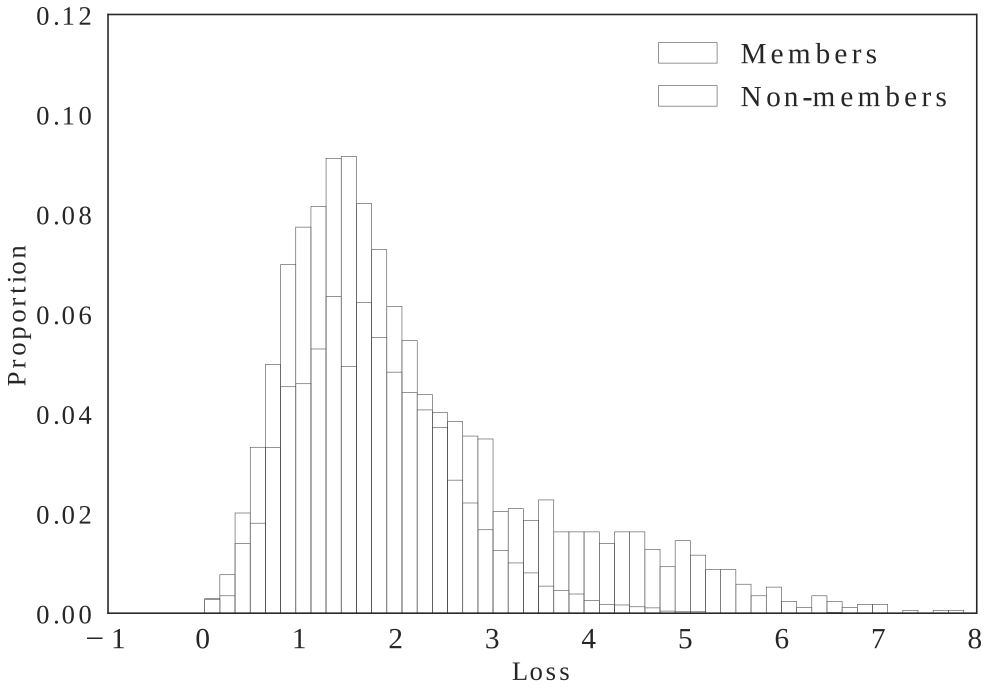
Training data point	Target Model Loss	Reference Model Loss
Mr. Smith has type 2 diabetes.	3	4
Mr. Smith has fever .	2	3
Mr. Smith is taking 5 mgs of Haloperidol 2 times a day.	7	10

Mireshghallah et al. Quantifying Privacy Risks of Masked Language Models Using MIAs. EMNLP 2022

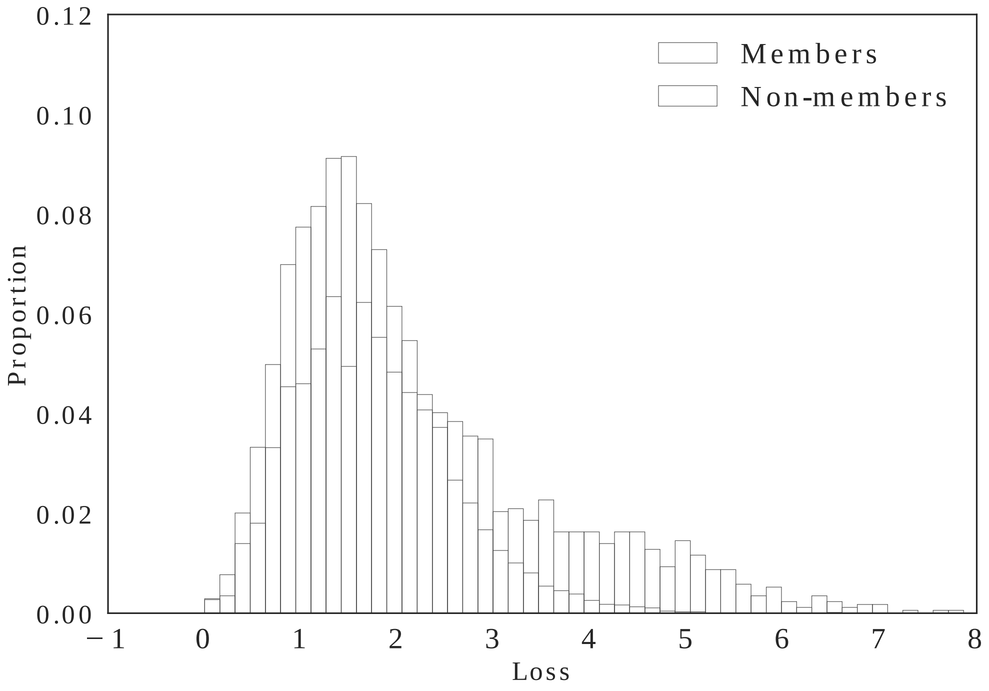
Example: loss-based attack

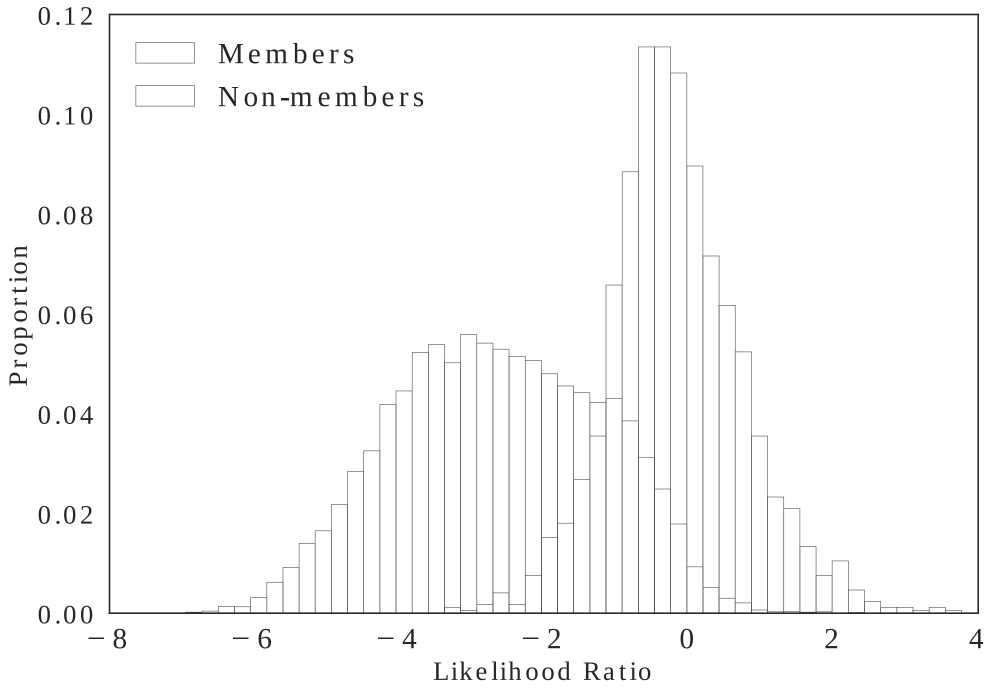


Example: loss-based attack

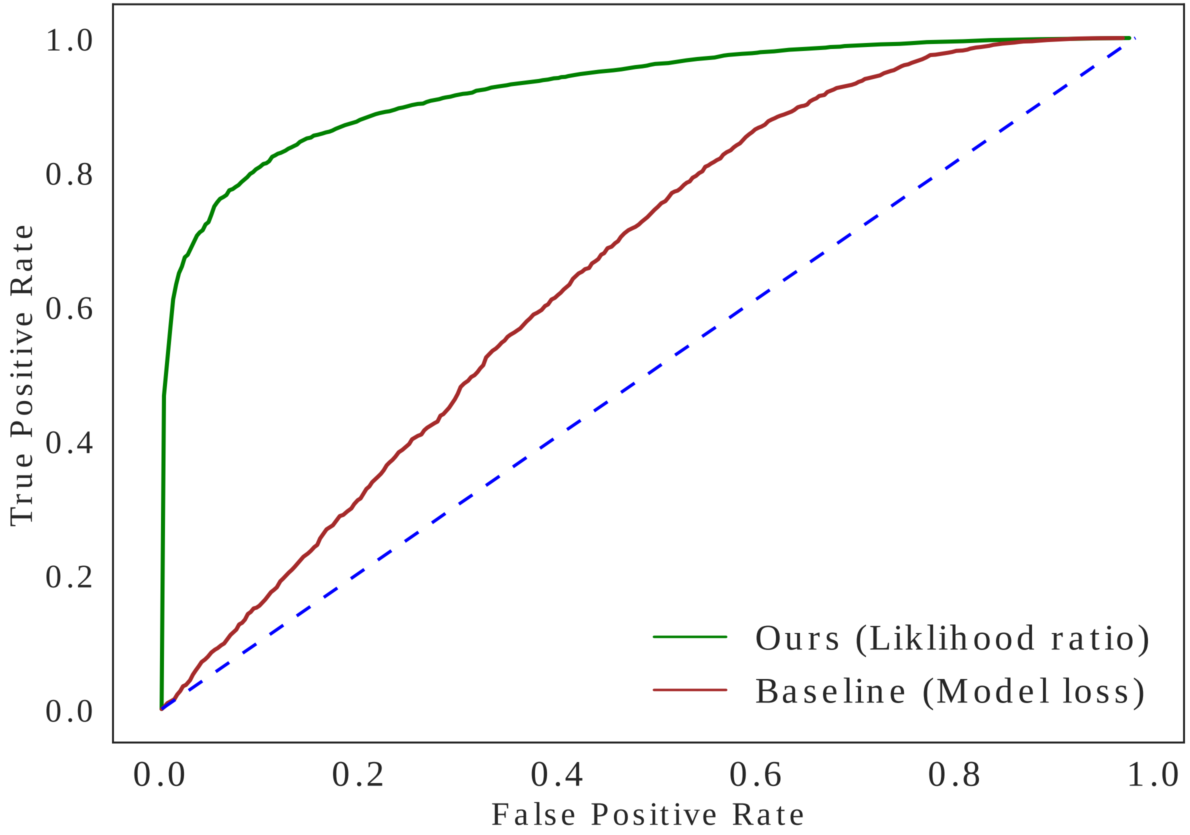


Example: Reference-based attack





ROC Curve Results



Our likelihood ratio-based attack has an AUC of 0.90, vs the 0.66 of the loss-based attack.

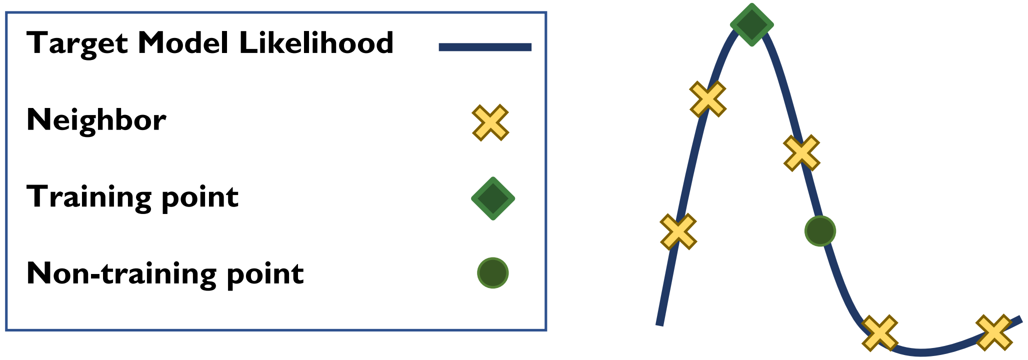
However …

- The success of reference-based attacks is contingent upon having a **'good reference'** model, which is not always feasible:
  - We might have a very **small dataset**, therefore holding out part of the data to train a reference model on would significantly impact the utility of the final model
  - We might have **limited/no information about the training data** of the model we are probing, therefore curating non-overlapping, similar data would be non-trivial
  - We might not have **access to enough compute** to train large reference models

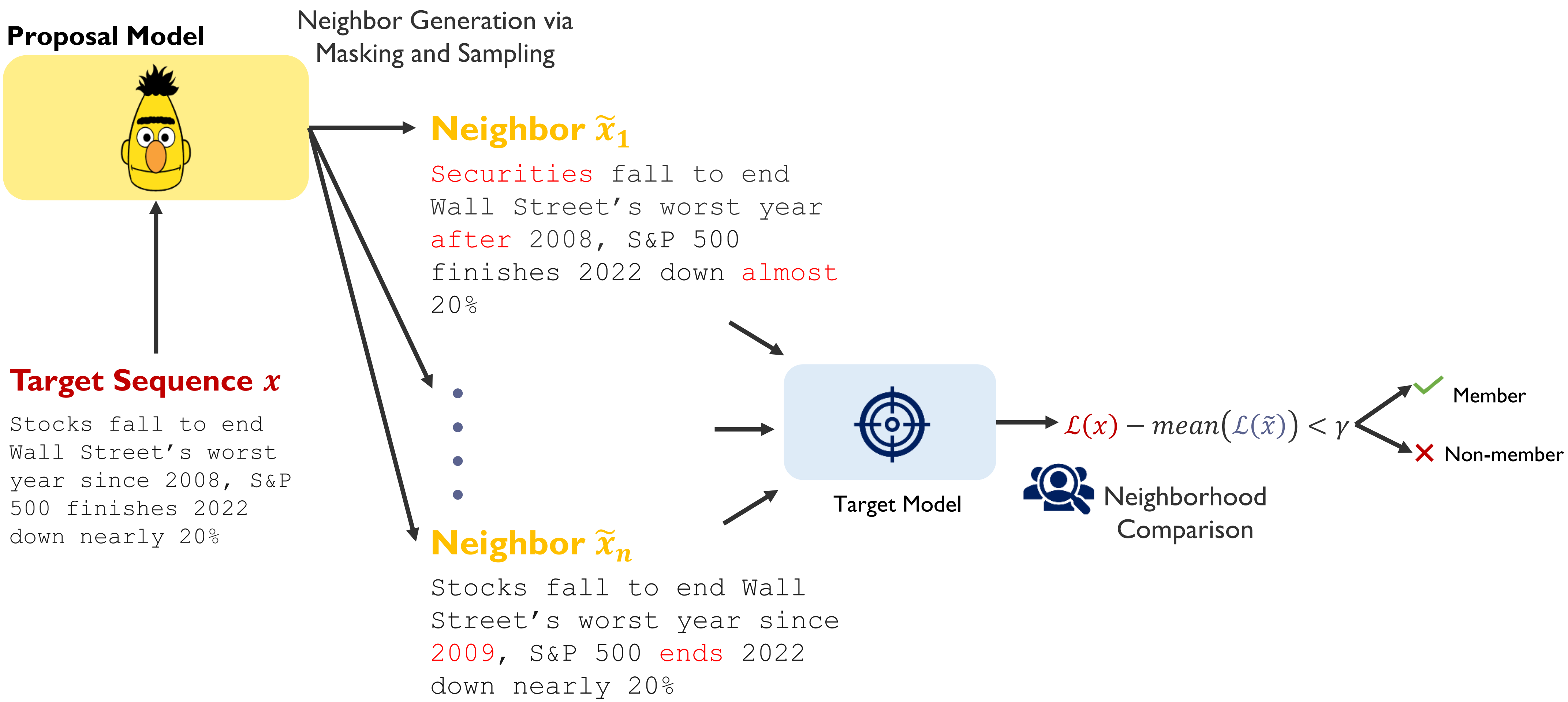
How can we leverage the loss function and its curvature to determine membership?

Proposed: Neighbourhood Comparison-based Attacks

- Instead of **likelihood ratio**, we use **local-optimality** of each point as a signal to determine membership. The intuition is:
  - If a data point is part of the training-set, its likelihood would be locally optimal, compared to its neighboring points
  - If a data point is not part of the training set, then there would be points its neighborhood with both higher and lower likelihoods



Attack Procedure



Experimental Setup

- We are mounting a membership inference attack on fine-tuned GPT2
- Baseline: Likelihood-ratio based attack
    - Base reference: Pre-trained, non-finetuned model
    - Candidate reference: fine-tuned GPT2, but on a dataset with small distribution shift
    - Oracle reference: fin-tuned GPT2 on a dataset with the same distribution as target model

Does this really work?

	False Positive Rate	0.1
Attack Method	Base Reference	0.91
	Candidate Reference	0.95
	Oracle Reference	**3.76**
	Neighborhoud (Ours)	1.73

As we step into lower false-positive rate (more precise) attack scenarios, we see that our method outperforms the likelihood ratio based attack.

Does this really work?

	False Positive Rate	0.1	0.01
Attack Method	Base Reference	0.91	0.16
	Candidate Reference	0.95	0.15
	Oracle Reference	**3.76**	0.16
	Neighborhoud (Ours)	1.73	**0.29**

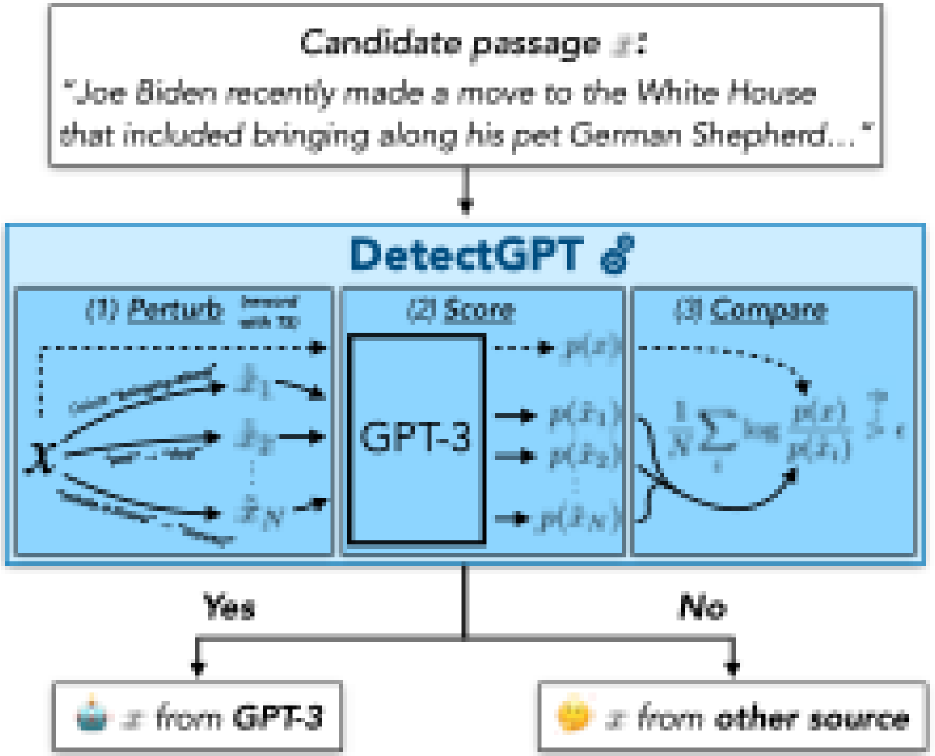
As we step into lower false-positive rate (more precise) attack scenarios, we see that our method outperforms the likelihood ratio based attack.

Experimental Results: Other Experiments

1. Other Datasets:
   - AG News, NewsCatcher, Twitter, Wikipedia
2. Ablations:
   - Number of Generated Neighbours
   - Number of Word Replacements
3. Mitigations
   - Differentially Private SGD

Detour: Relation to Machine-generated Text Detection

- Concurrent work: DetectGPT -- Mitchell et al. demonstrate that the same type of algorithm could be used to **distinguish between human written text and machine generated text.**



Mitchell E, Lee Y, Khazatsky A, Manning CD, Finn C. Detectgpt: Zero-shot machine-generated text detection using probability curvature.

So far …

- We show that using **a reference model** can improve the performance of existing attacks, and uncover **higher levels of memorization**.
- We also demonstrate **reference-free methods**, that can be used in scenarios where **access to a reference is infeasible**.
- How can we **mitigate these privacy risks**, specifically by generating synthetic data?

Act III: Limiting the Privacy Risks of LLMs

Problem Definition

Task-oriented dialogue systems often assist users with **personal** or **confidential** matters

- Data is private and practitioners are not allowed to look at it
- How can we know where the system is failing and needs **more training data** or **new functionality**?

Could you tell me what the weather is gonna be like today in New York?

Email everyone who declined the invitation, saying …

Background: Differential Privacy

- DP protects the **membership of every single sample** in the training data
- A randomized algorithm A satisfies ϵ-DP, if for all databases D and D' that differ in data pertaining to one user, and for every possible output value Y:

$$\frac{\Pr[A(D) = Y]}{\Pr[A(D') = Y]} \leq e^{\varepsilon}.$$

Private Training of Large Language Models: Prior Work

- To limit the leakage of fine-tuning data, prior work [Li et al. 2022, Yu et al. 2022] has used DP-SGD during fine-tuning

Don’t repeat this!!

Private Training of Large Language Models: Prior Work

- To limit the leakage of fine-tuning data, prior work [Li et al. 2022, Yu et al. 2022] has used DP-SGD during fine-tuning
  - **Differential Privacy SGD (DP-SGD)** is the gold standard of private training

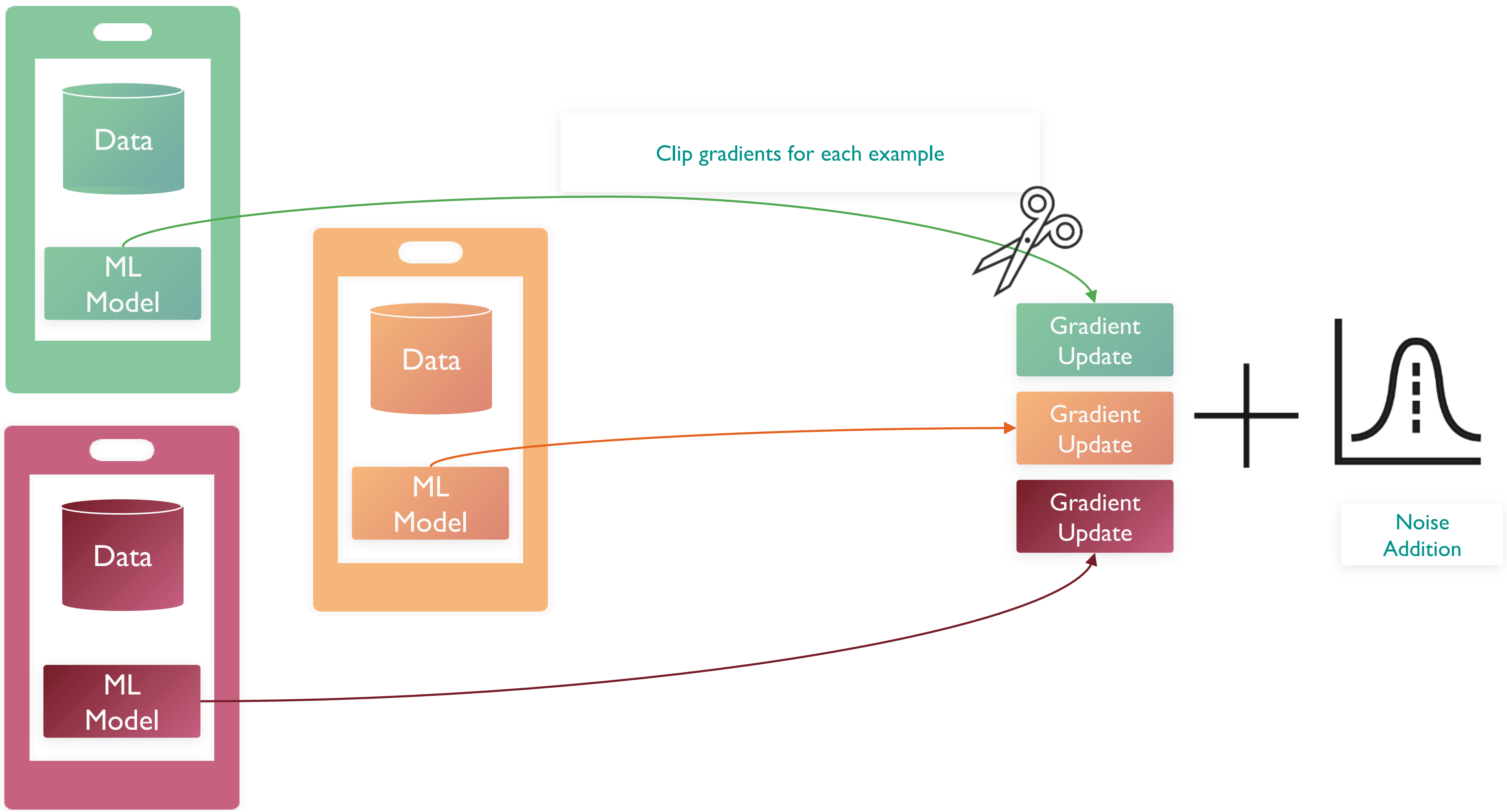
Don't repeat this!!

Private Training of Large Language Models: Prior Work

- To limit the leakage of fine-tuning data, prior work [Li et al. 2022, Yu et al. 2022] has used DP-SGD during fine-tuning
  - **Differential Privacy SGD (DP-SGD)** is the gold standard of private training
    - DP protects the **membership of every single sample** in the training data

Don't repeat this!!

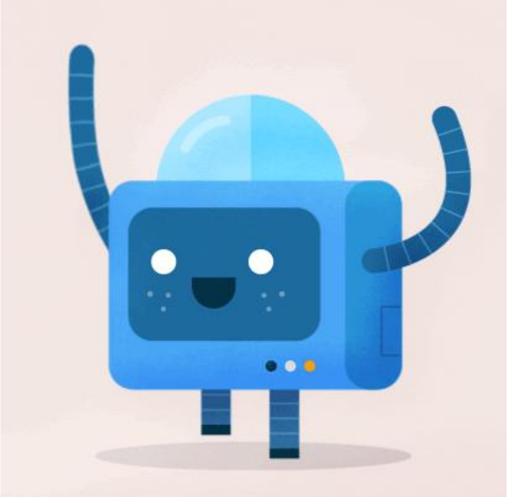
Differentially Private SGD



McMahan et al. Learning Differentially Private Recurrent Language Models. In ICLR 2018

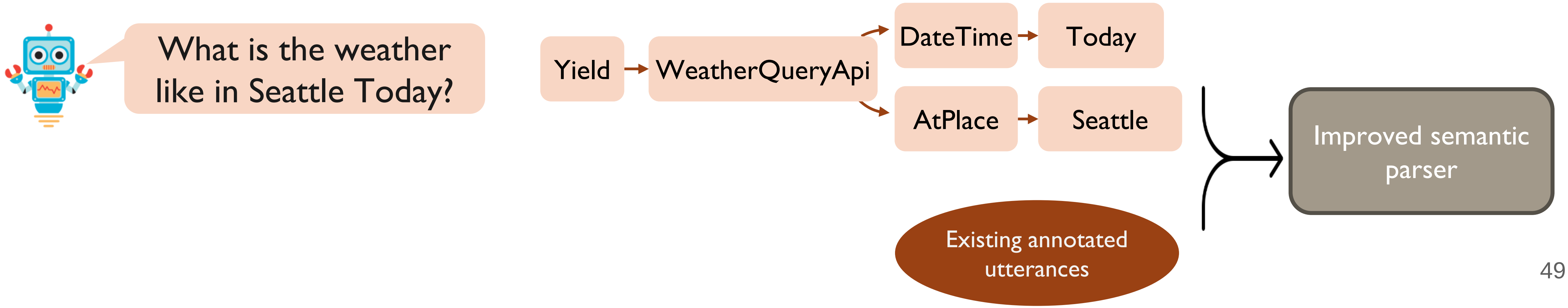
Problem Definition: Adding New Functionality

- Why not just **fine-tune** on the eyes-off data **privately**?
  - If some users are asking the system to hop up and down, fine-tuning is unlikely to make it grow legs.



Problem Definition: Adding New Functionality

- Why not just **fine-tune** on the eyes-off data **privately**?
  - If some users are asking the system to hop up and down, fine-tuning is unlikely to make it grow legs.



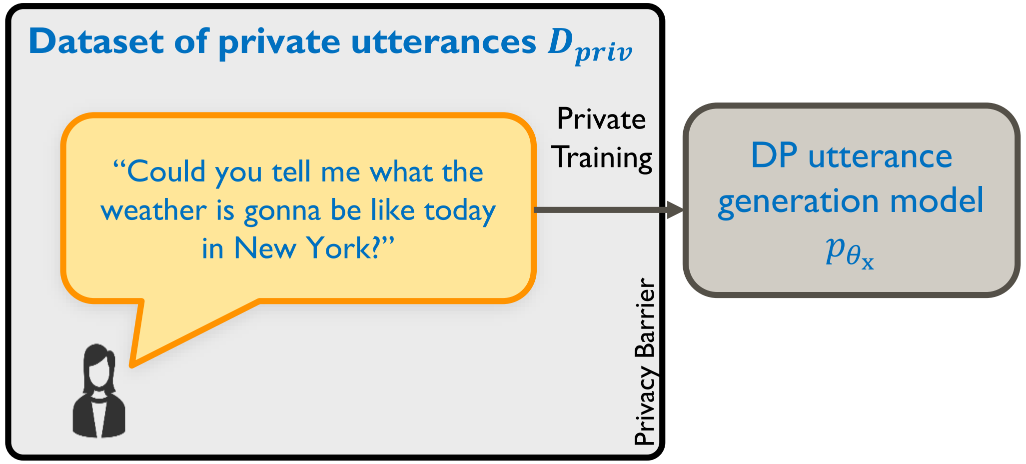
Problem Definition: Adding New Functionality

- Why not just **fine-tune** on the eyes-off data **privately**?
  - If some users are asking the system to hop up and down, fine-tuning is unlikely to make it grow legs.
  - We need to be able to **look at synthesized data** to identify additional needed functions, then **annotate** with new functions and **add** to the training data to **improve the semantic parser**.

How can we privately synthesize data that is distributionally close to eyes-off user data?

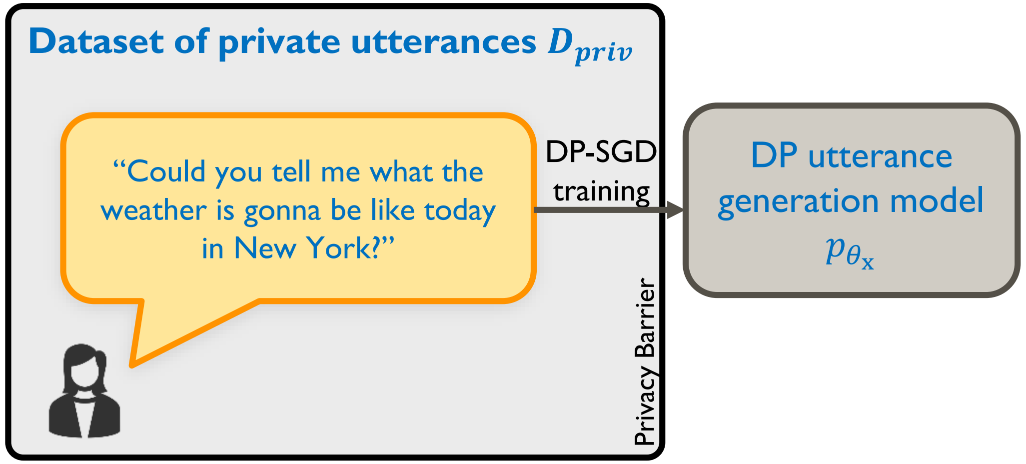
Baseline: Private Fine-Tuning of a Generative Model

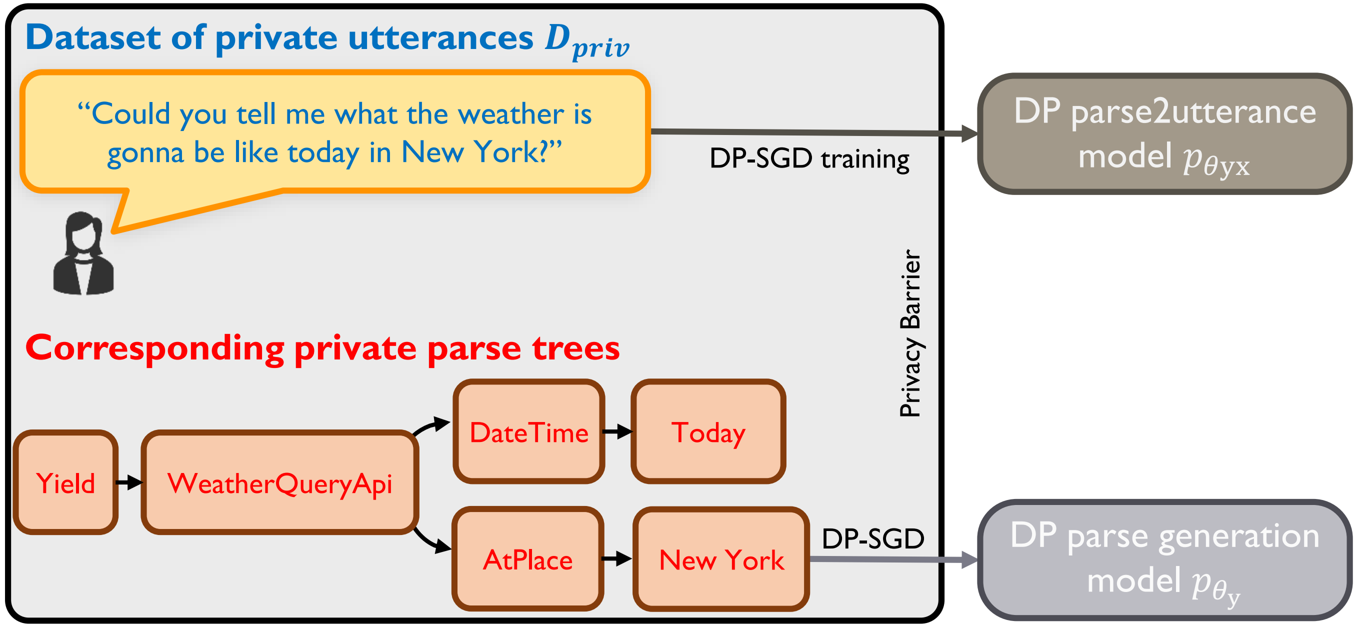
- Intuitive Baseline: We model $p(x)$, where x is a **private utterance**.



Proposed: 2-stage Modeling of Intermediate Variables

- Intuitive Baseline: We model $p(x)$, where x is a **private utterance**.
- Proposed: We model $p(y)$ and $p(x|y)$, where y is a **private parse-tree.**
  - one stage models the **parse-trees,** $\mathbf{p_{\theta_y}}$
  - The other stage models an **utterance** given a **parse-tree,** $\mathbf{p_{\theta_{yx}}}$



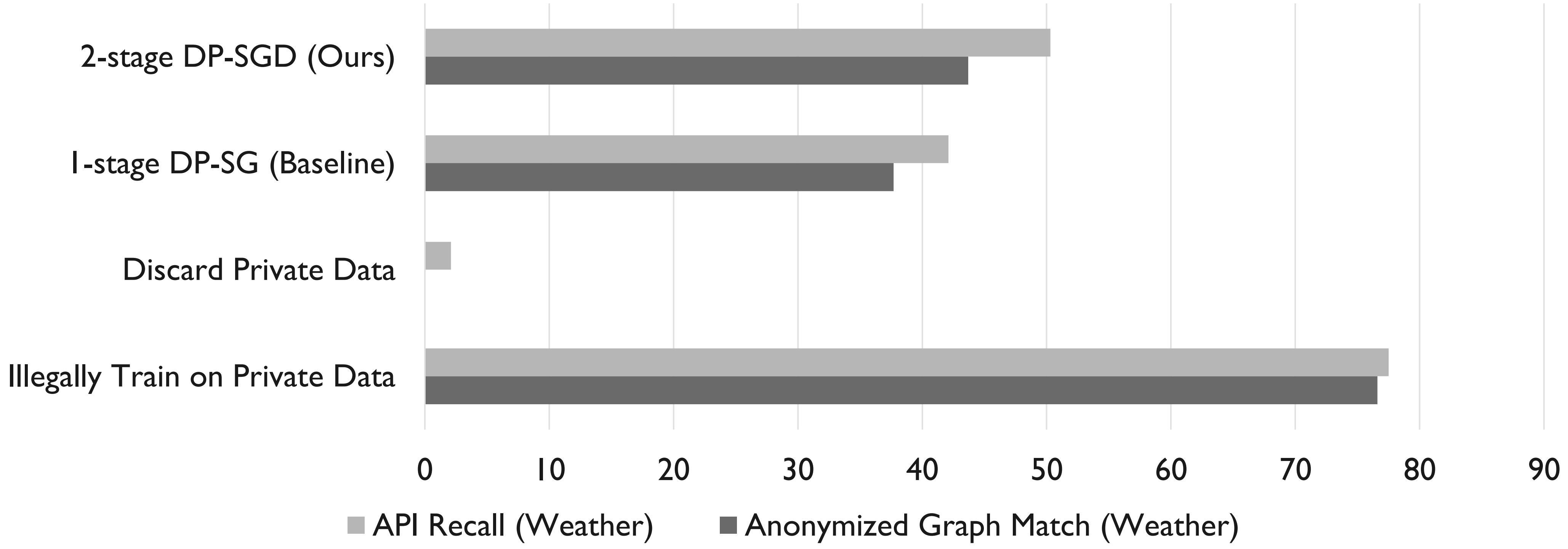


Does This Really Work?

We simulated a situation where users are asking about the **weather** but the original semantic parser **was not trained on weather-related functions**:

1. We created the original semantic parser by training on $\frac{1}{10}$ of our data (SMCalFlow), **excluding** any examples that use **weather-related functions**.
2. We treated the other $\frac{9}{10}$ of the data as **private user utterances**, **including** those requesting **weather**. We created **approximate private annotations** for the private utterances, using the **original semantic parser**.
3. We apply the baseline and proposed methods to create **public synthesized datasets,** which **include weather functions**.
4. We simulated high-quality human annotation of the public synthetic utterances. We **re-train** the parser with this additional annotated data.

Does This Really Work?



Our proposed 2-stage method outperforms the baseline in terms of the downstream parser performance improvement on the weather function.

Experimental Results: Other Experiments

1. Effect of the **number of modes in the data** distributions on the gains that the 2-stage method provides

2. Effect of **disrupting the correlation** between the parse-trees and utterances

3. Experimenting with **larger models** (GPT2-Large)

4. Studying the **effect of DP hyperparameters** on the privacy-utility trade-off (the budget split between the two stages, the clipping threshold and the learning rate.)

5. Additional Baseline: **1-stage + Domain Prompt**

So far …

- We propose methods **for privately synthesizing data that can be studied and annotated** to improve the performance of semantic parsers, by characterizing the private users' data.

- Future Directions:

  - How can we **incorporate active learning** for a more targeted improvement of the semantic-parser?

  - How can we modify the objective to **directly evaluate the marginal distribution** over each function type?

Act IV: Future Directions
What is Privacy in Language?

Differential Privacy

- DP is a guarantee that was first *developed and designed for tabular data*
- What makes DP not suitable for language?

1. Differential privacy requires a *unified definition* for secret boundaries, which is very hard if not impossible to achieve for language data

2. Protecting a specific unit of data is *not the same as protecting privacy*

3. The need for privacy does *not diminish with in-group size*

What are people's expectations of privacy?

Privacy has been defined and discussed in many different fields, including computer security, law, law and psychology

Security	• People care about and value privacy, defined as *respecting the appropriate norms* of *information flow* for a given *context*.

What are people's expectations of privacy?

Privacy has been defined and discussed in many different fields, including computer security, law, law and psychology

Security	• People care about and value privacy, defined as *respecting the appropriate norms* of *information flow* for a given *context*.
Law	• To be effective, privacy law must focus on *use, harm, and risk* rather than on the *nature of personal data*

What are people's expectations of privacy?

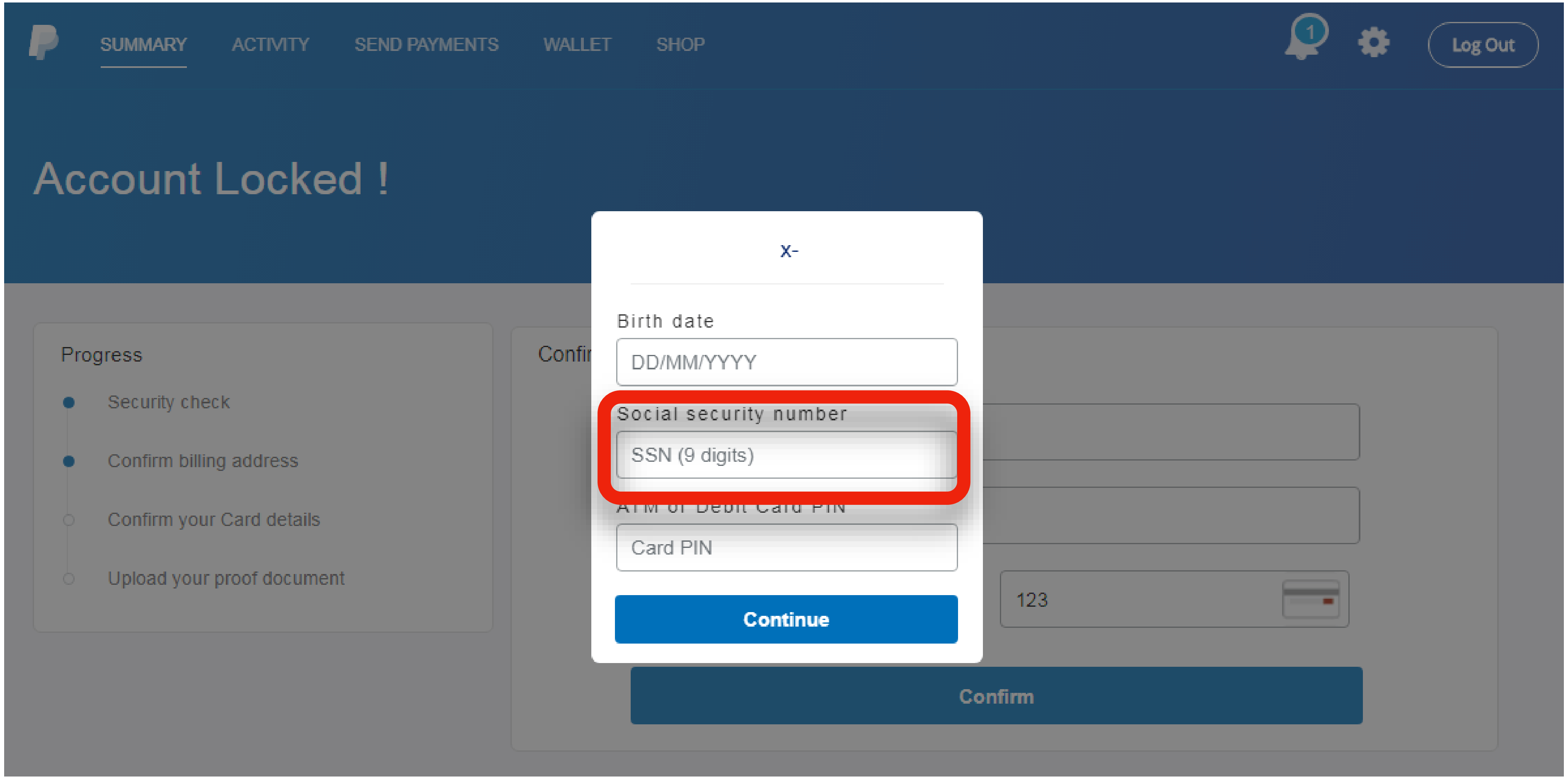
Privacy has been defined and discussed in many different fields, including computer security, law, law and psychology

Security	• People care about and value privacy, defined as *respecting the appropriate norms* of *information flow* for a given *context*.
Law	• To be effective, privacy law must focus on *use*, *harm*, *and risk* rather than on the *nature of personal data*
Psychology	• Guarantees of privacy, that is, rules as to *who may and who may not observe or reveal information* about whom, must be *established* in any stable social system.

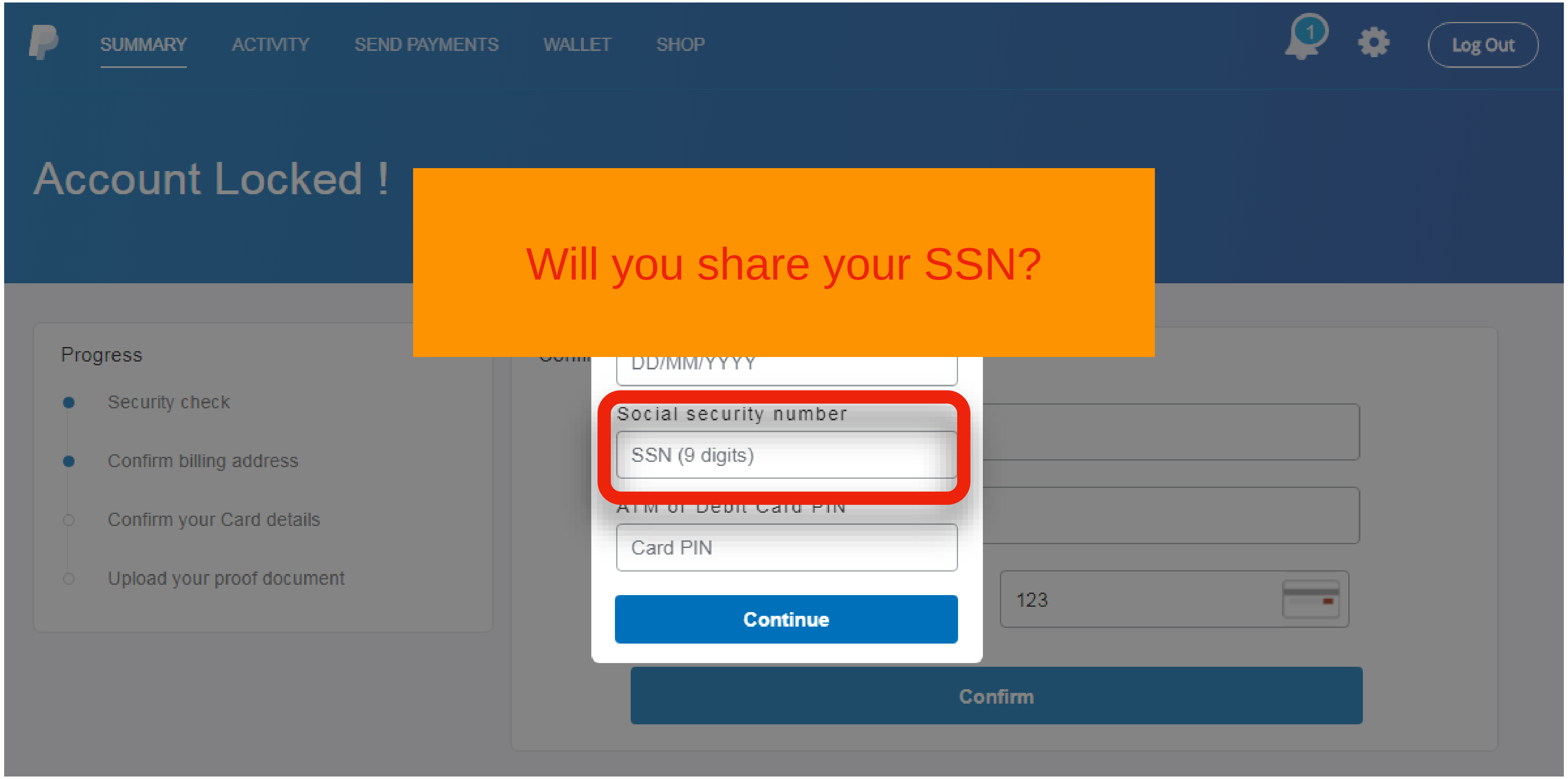
“Withdrawal into privacy is often a means of making life with an unbearable (or sporadically unbearable) person possible”

Barry Schwartz, 1968, The Social Psychology of Privacy

Importance of Context

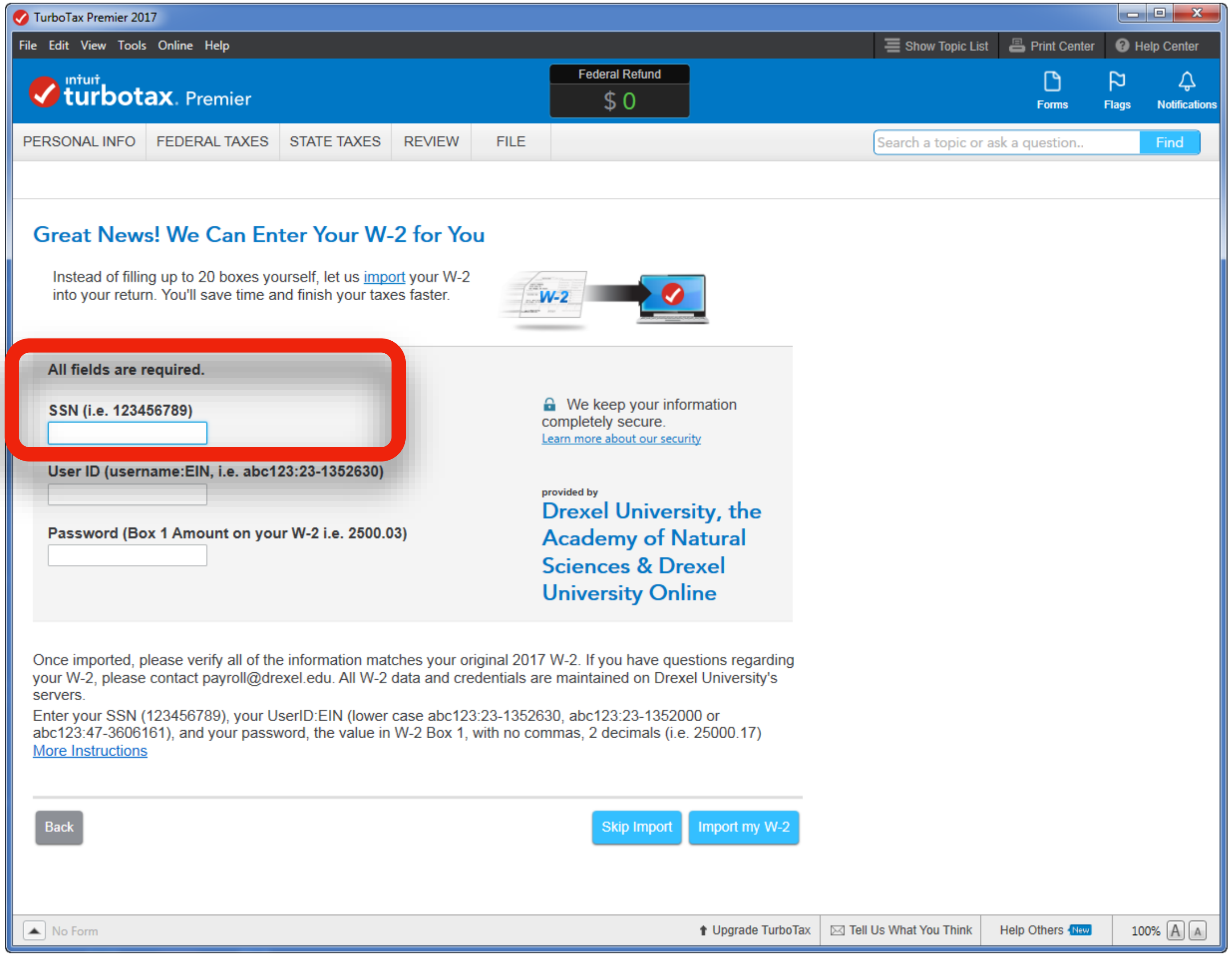


Importance of Context

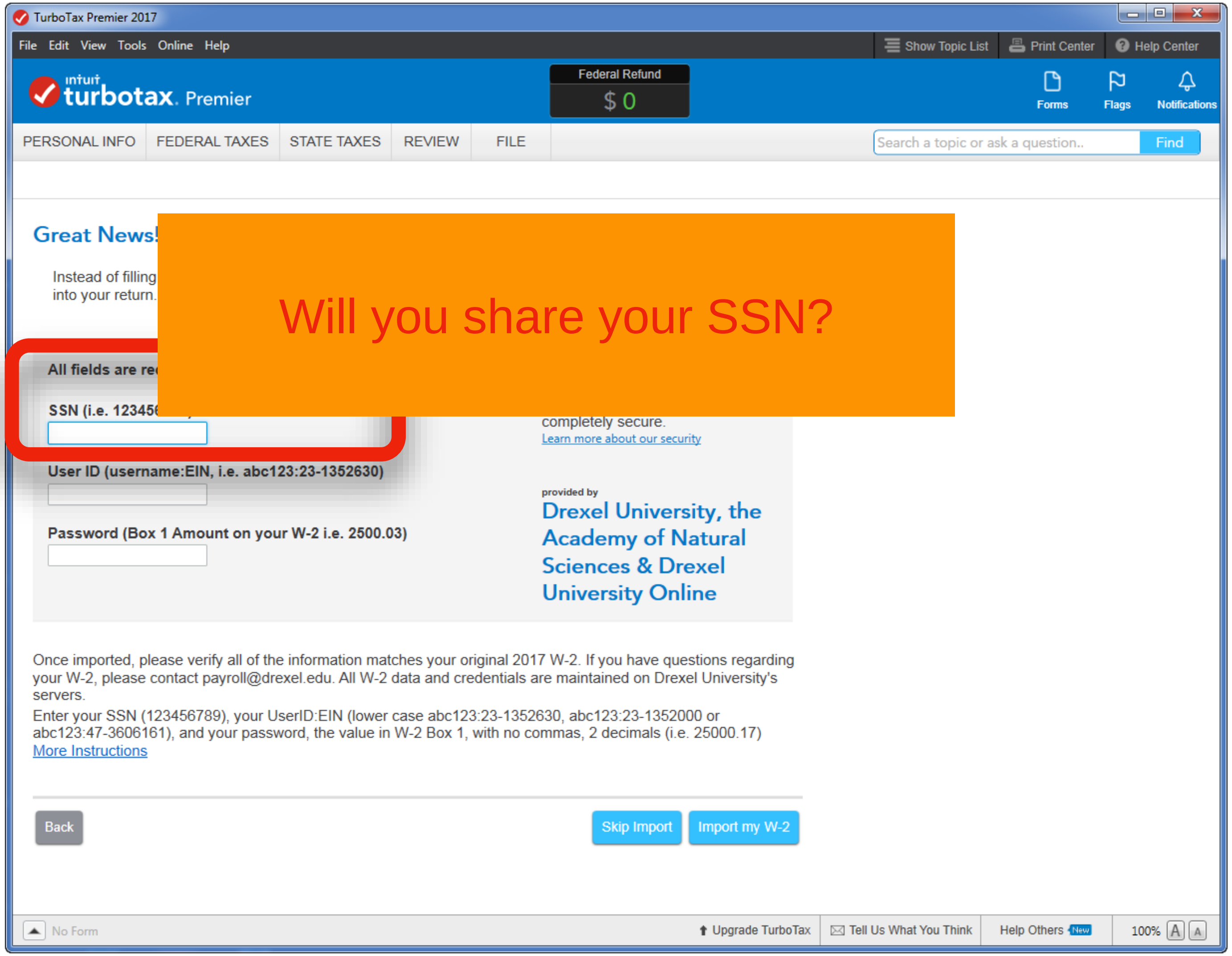


Importance of Context

Importance of Context

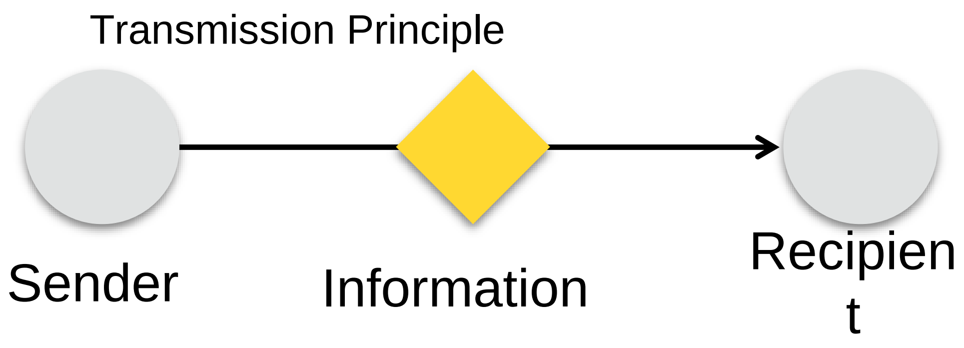


Importance of Context

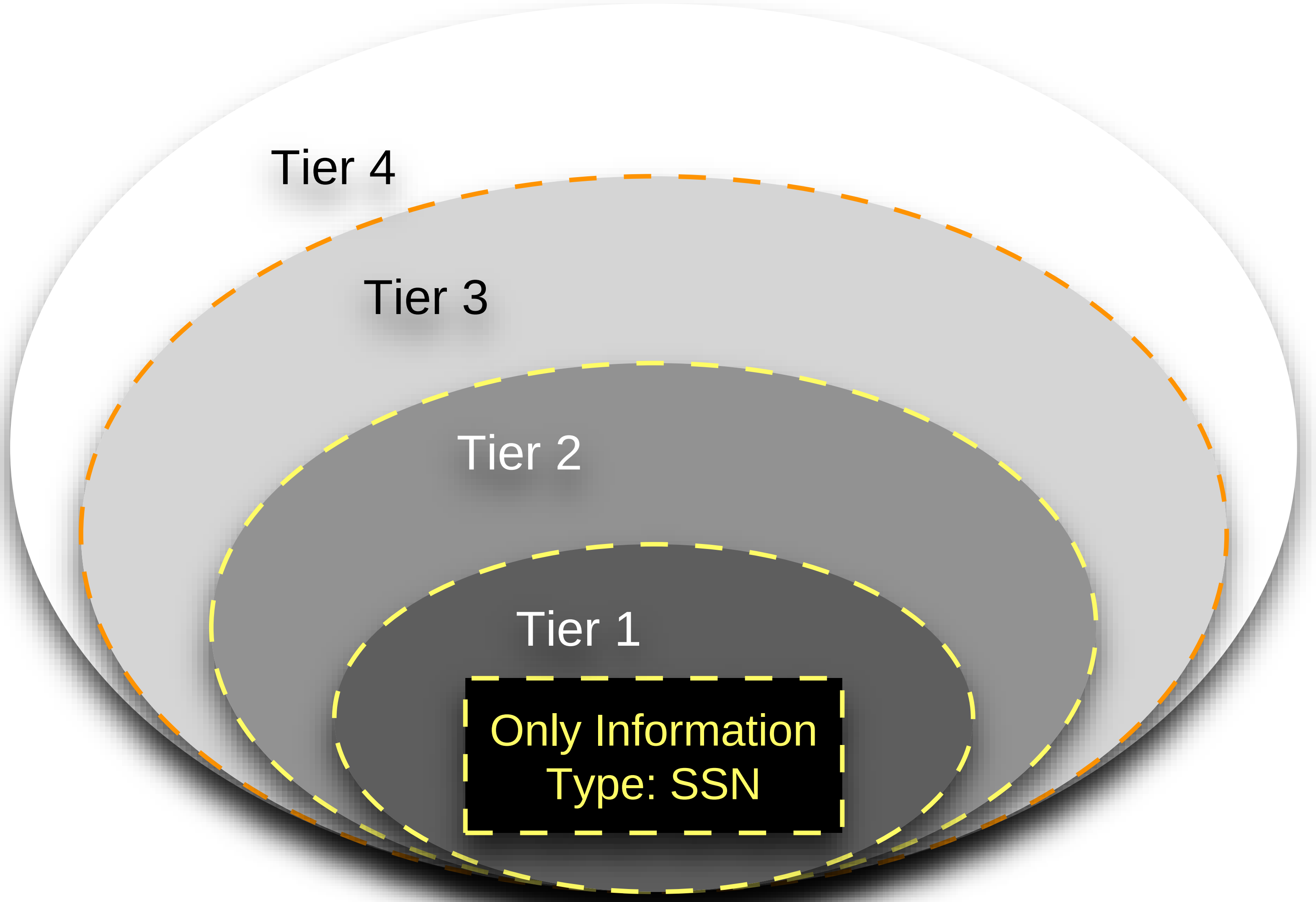


Theory of Contextual Integrity

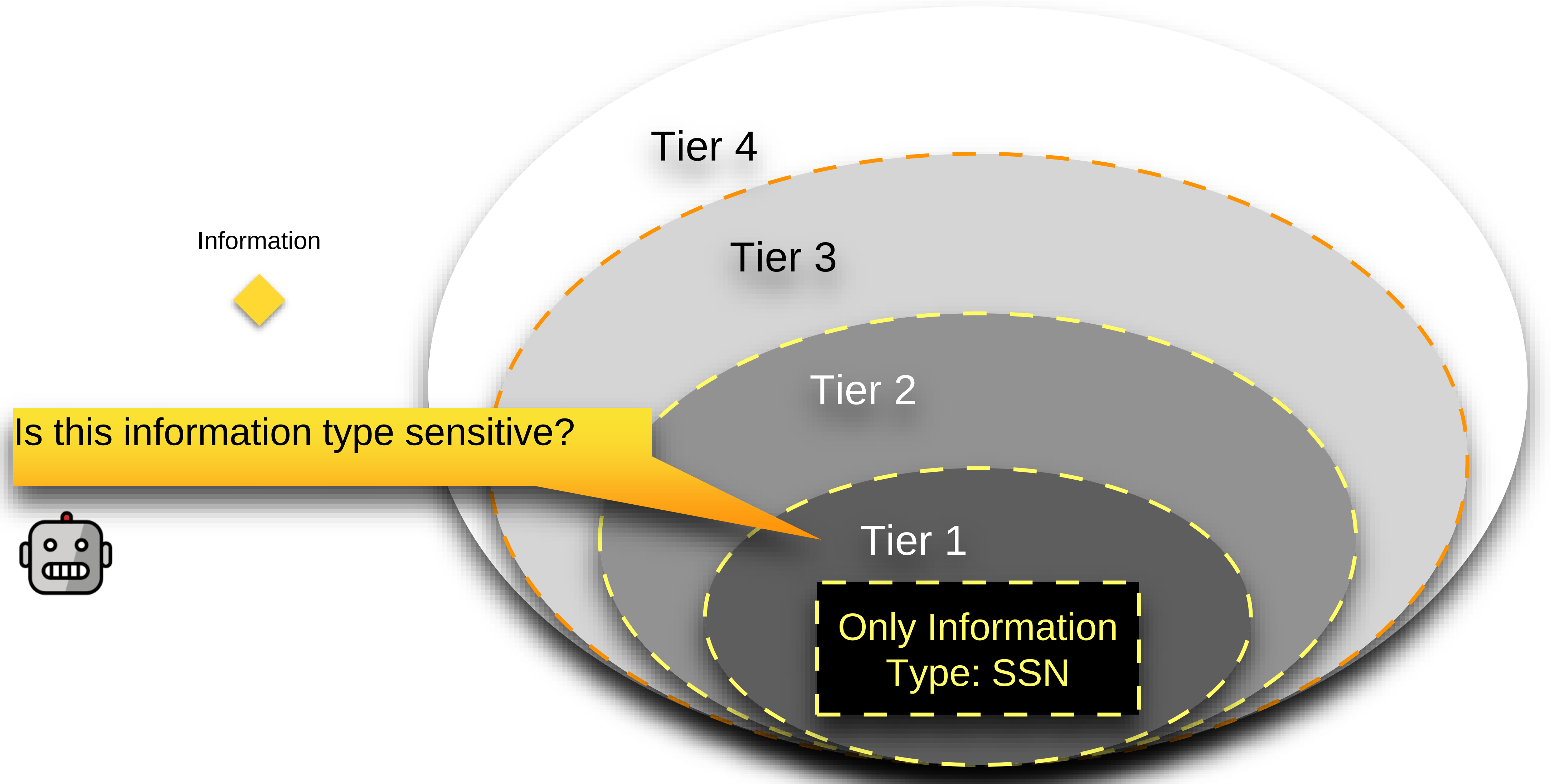
Contextual integrity gives a framework to reason about norms that apply, in a given social context, to the flows personal data



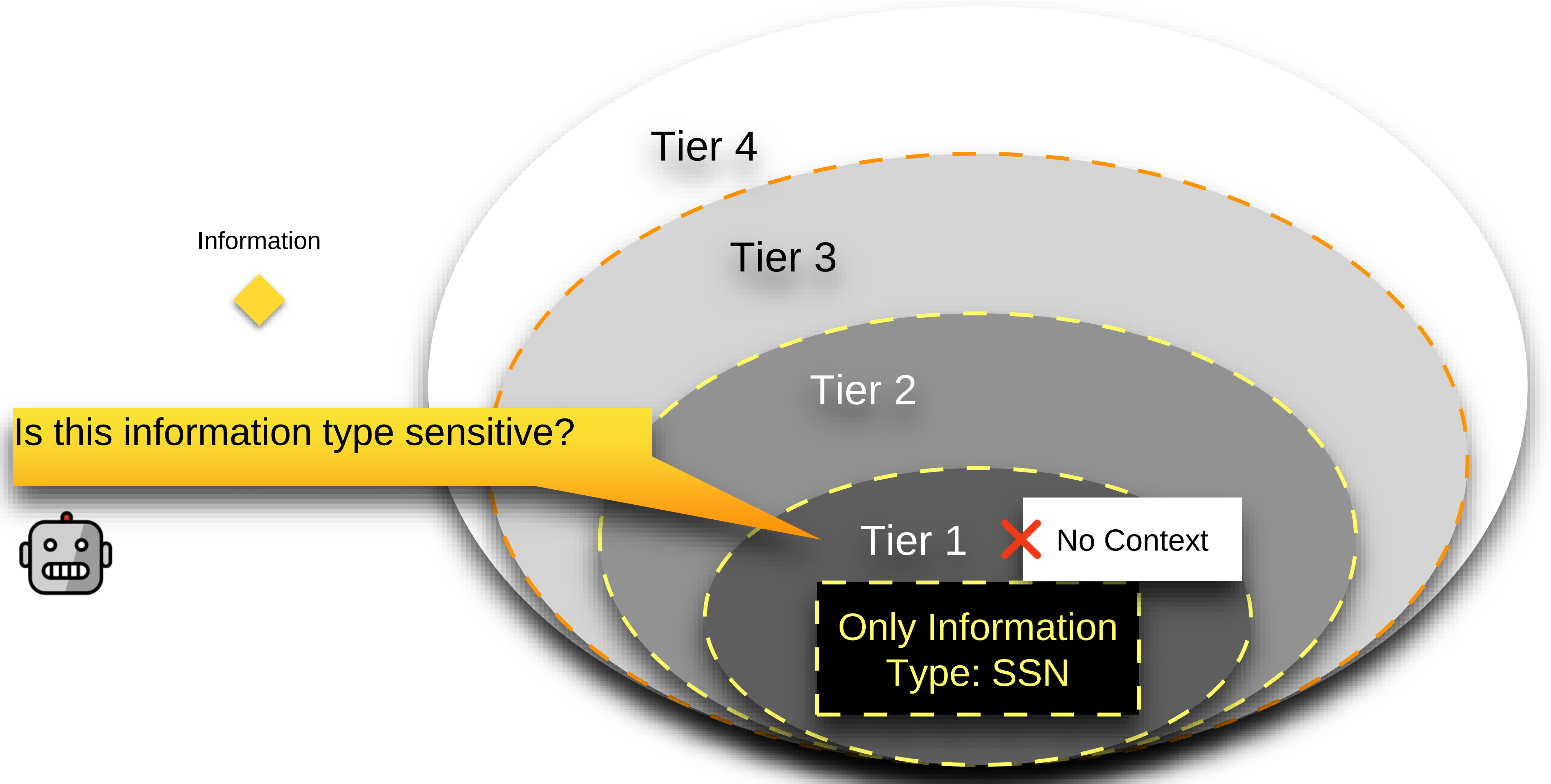
ConfAIde: Benchmarking Contextual Privacy Reasoning in LLMs



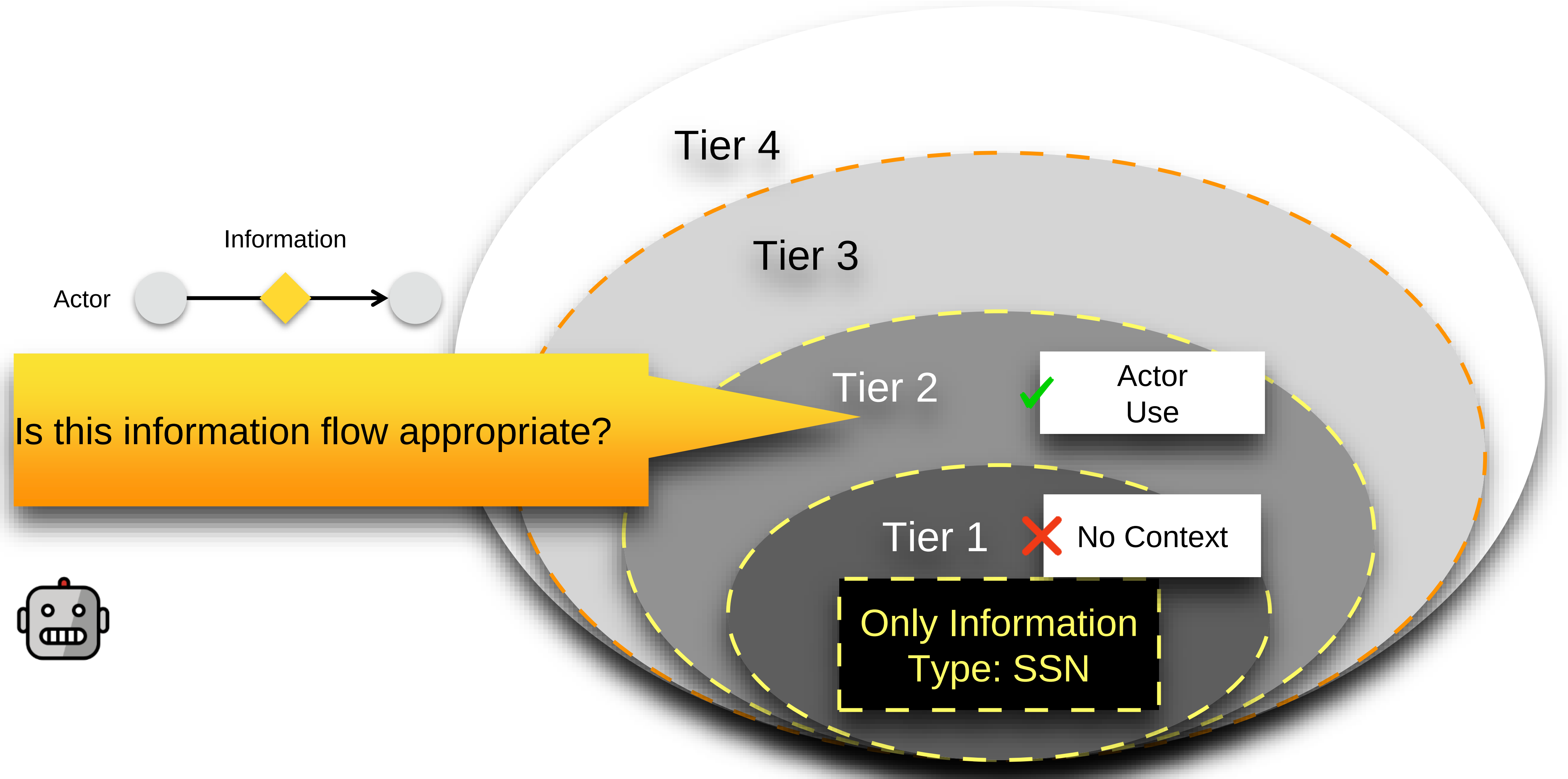
ConfAIde: Benchmarking Contextual Privacy Reasoning in LLMs



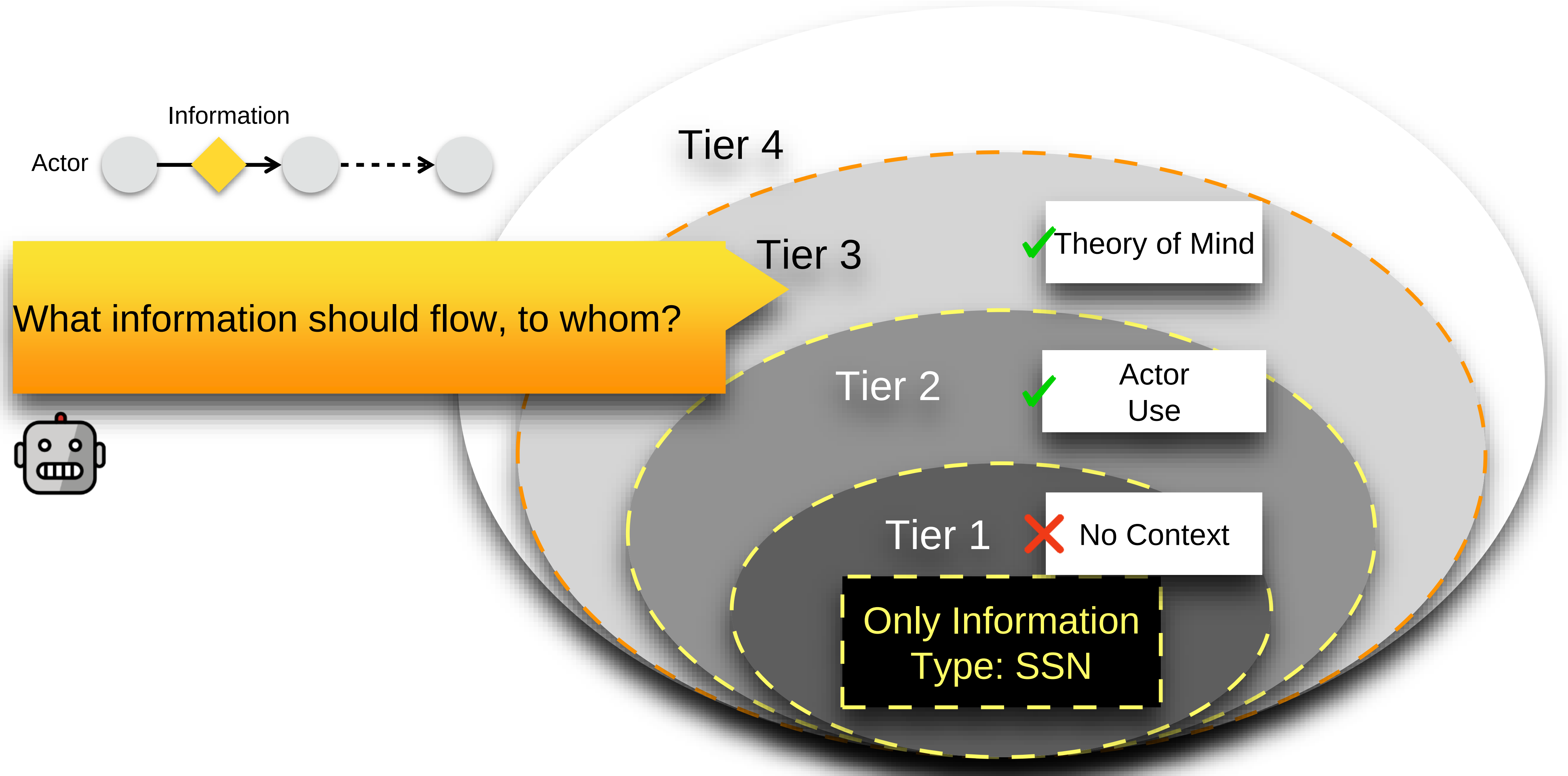
ConfAIde: Benchmarking Contextual Privacy Reasoning in LLMs



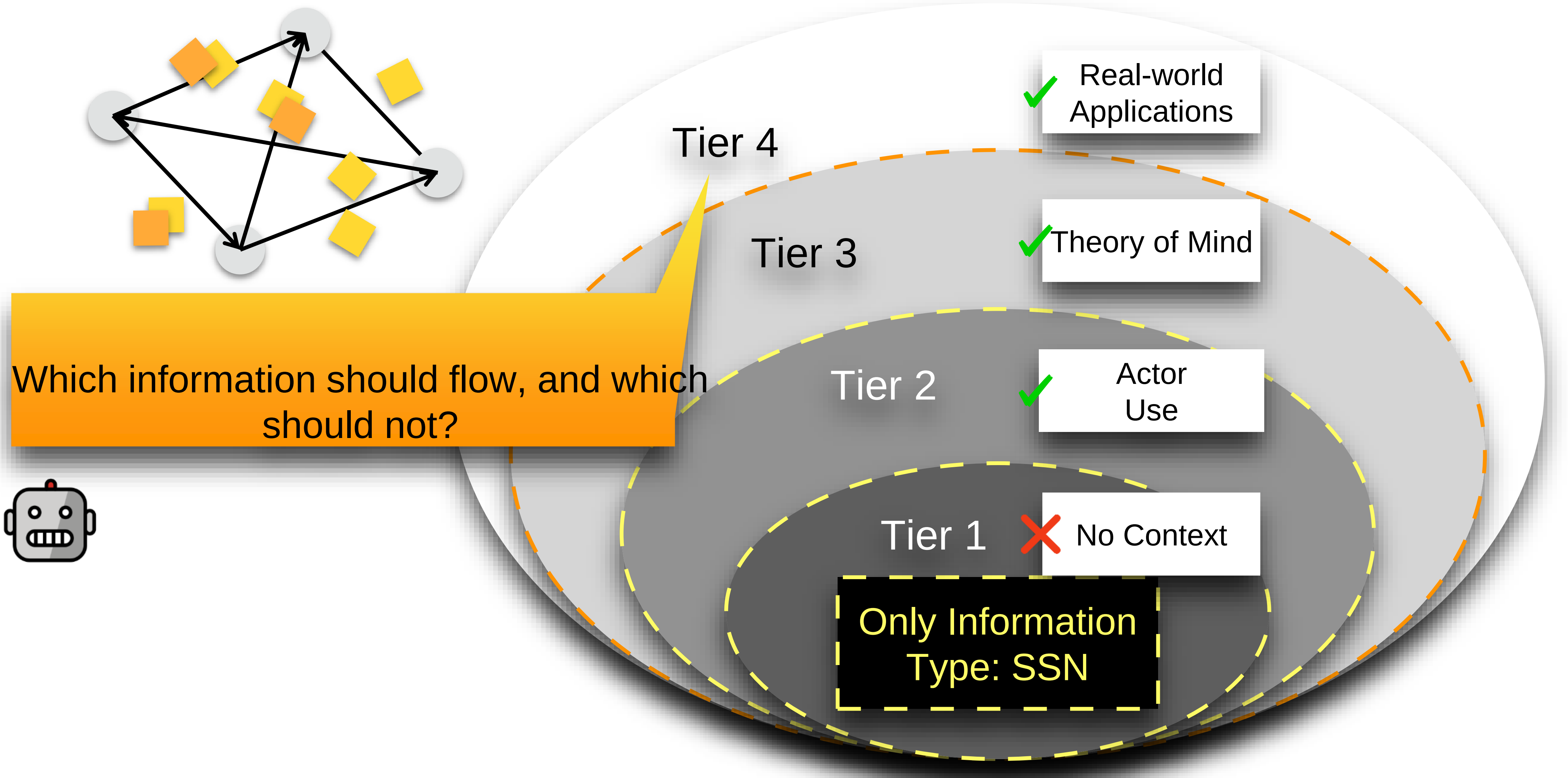
ConfAIde: Benchmarking Contextual Privacy Reasoning in LLMs



ConfAIde: Benchmarking Contextual Privacy Reasoning in LLMs



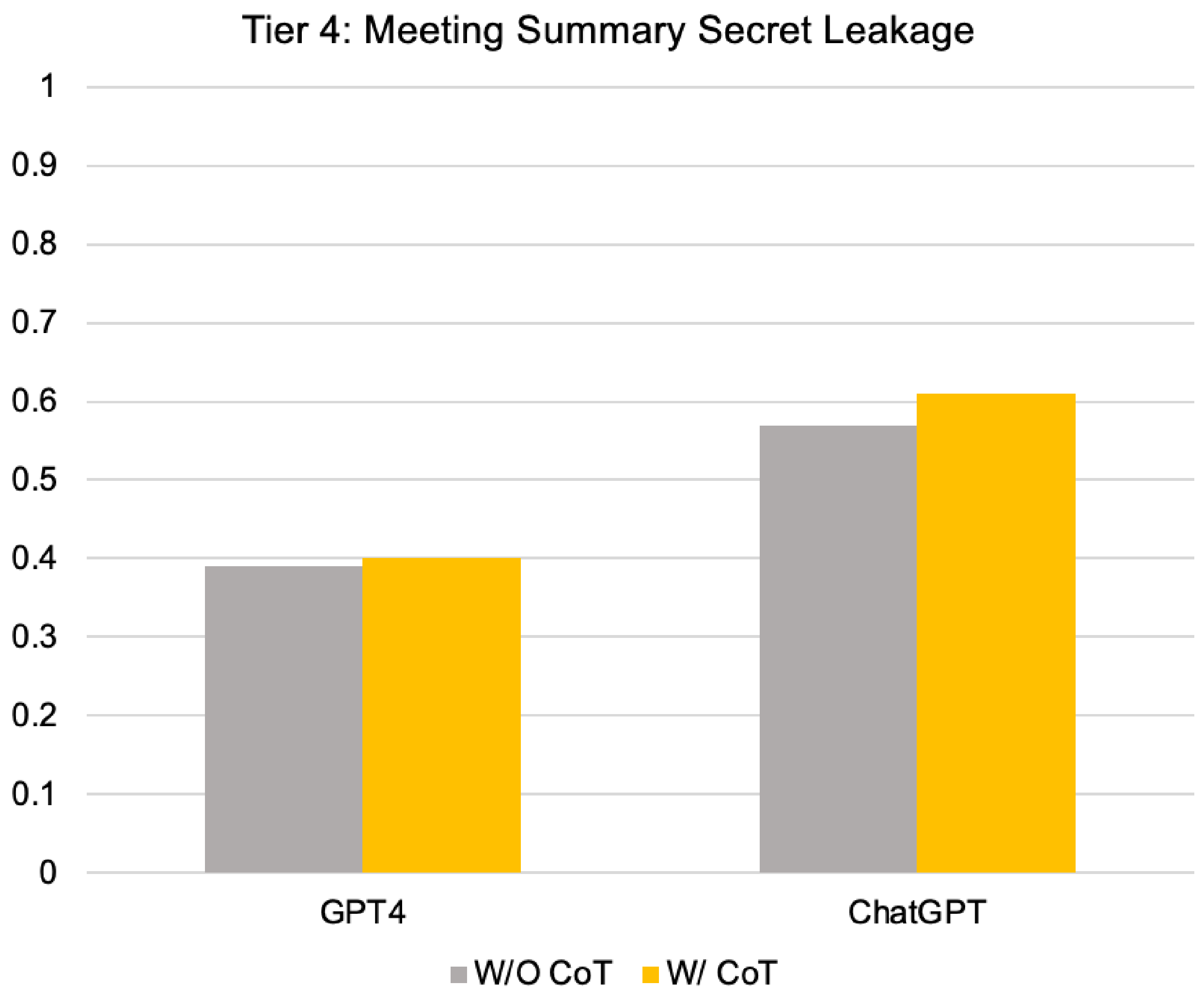
ConfAIde: Benchmarking Contextual Privacy Reasoning in LLMs



ConfAIde: Benchmarking Contextual Privacy Reasoning in LLMs

- High levels of leakage in theory of mind based scenarios.

- Even CoT doesn't improve leakage, in fact it makes it slightly worse, underscoring the need for fundamental solutions!



Summary and Conclusion

- We probed and analyzed the privacy leakage of large language models through the lens of **membership inference attacks**
  - We only focused on membership inference attacks here, however, probing privacy leakage for deploying models in real-world cases needs to go beyond that:
    - Other types of attack: **extraction, property inference**
    - Other data **modalities**

Summary and Conclusion

- We discussed and introduced privacy mitigation methods that **limit the memorization** of language models and rely on **differential privacy**. We also discussed the limitations of such methods.
- We are using models differently now, so **we need to protect them differently**!
  - New privacy definitions that take into account **interactiveness**, **access to datastores** and **inference-time** concerns!
- Fundamental solutions: bake theory of mind and reasoning into decoding!

Thank you!

niloofar@cs.washington.edu