Privacy-Preserving Domain Adaptation of Semantic Parsers

Fatemehsadat Mireshghallah, Yu Su, Tatsunori Hashimoto, Jason Eisner, Richard Shin

fatemeh@ucsd.edu

ACL 2023

Problem Definition: Background

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

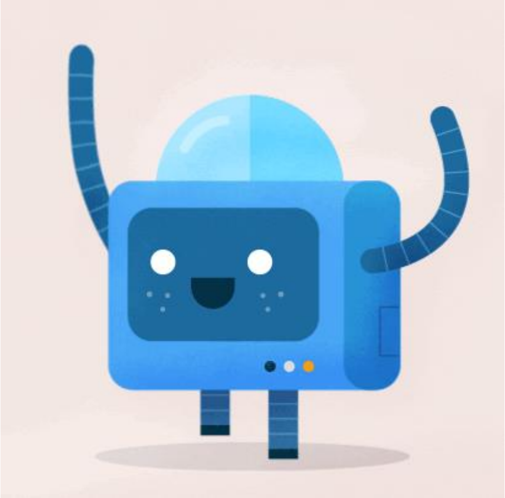
- Data is private and practitioners are not allowed to look at it
- How can we know where the system is failing and needs **more training data** or **new functionality**?

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

Email everyone who declined the invitation, saying …

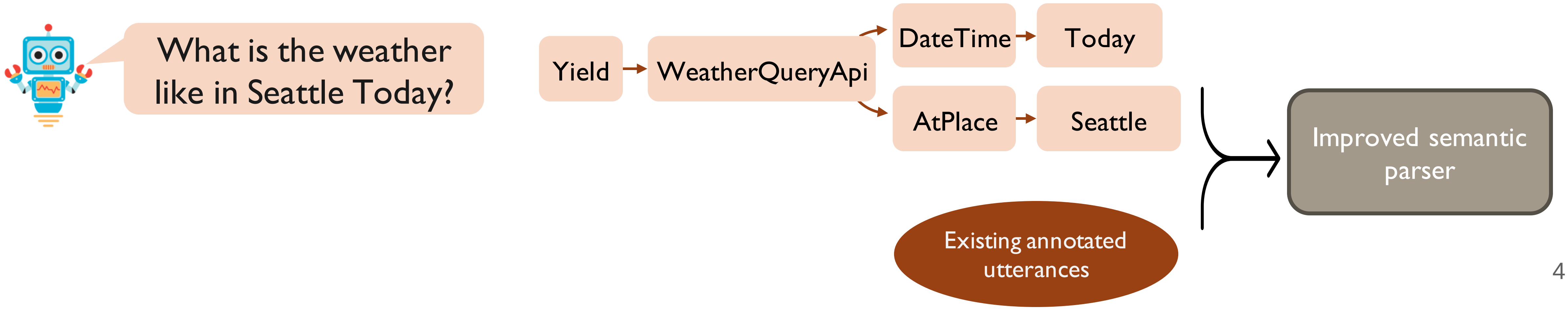
Problem Definition: Adding New Functionality

- 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?

Background: Differential Privacy

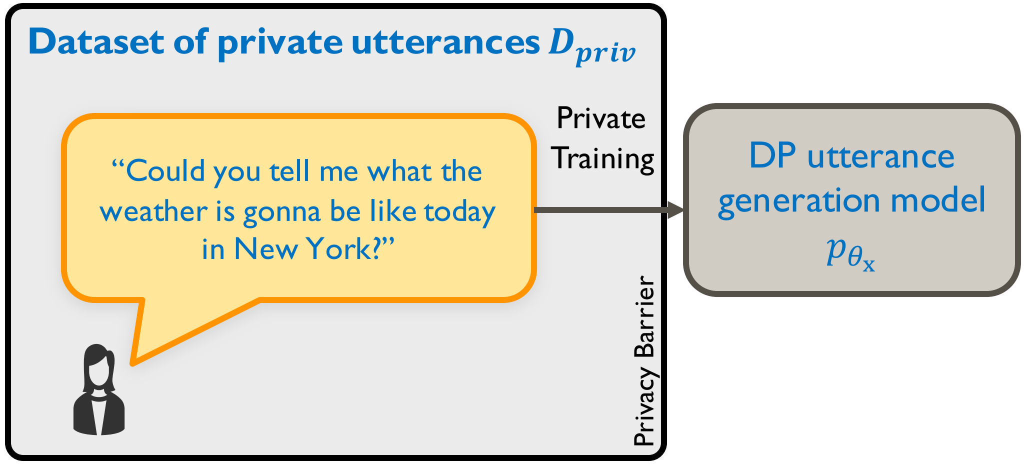
- DP protects the **membership of every single sample** in the training data
- A randomized algorithm A satisfies ϵ-DP, if for all databases D and D' that differ in data pertaining to one user, and for every possible output value Y:

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

- We use DP-SGD, a differentially private variant of SGD:
  - Clipping gradients and adding noise

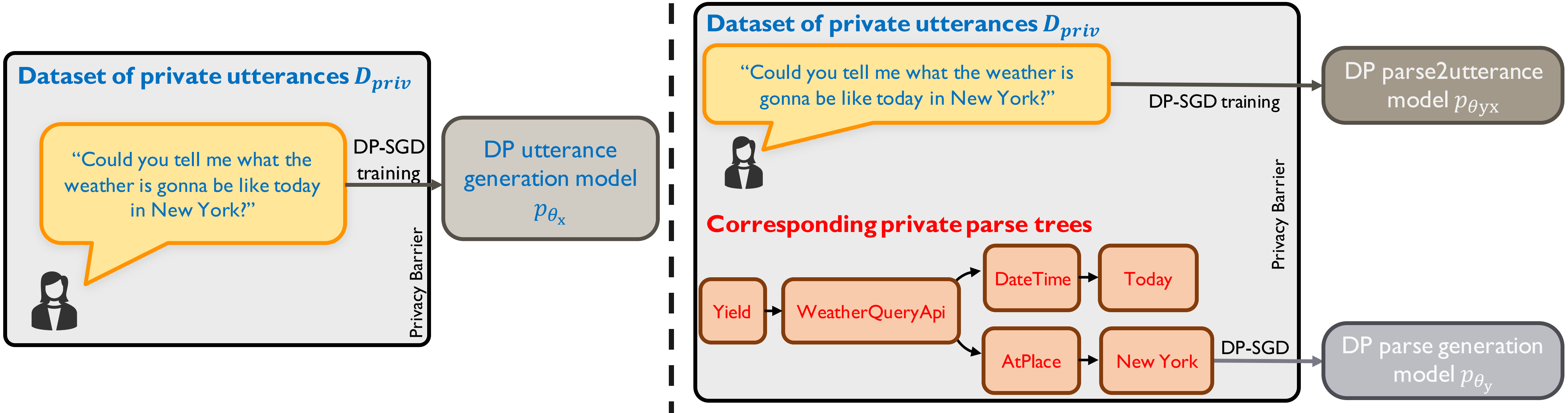
Baseline: Private Fine-Tuning of a Generative Model

- Intuitive Baseline: We model $p(x)$, where $\boldsymbol{x}$ 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,** $\mathbf{p}_{\theta_{yx}}$

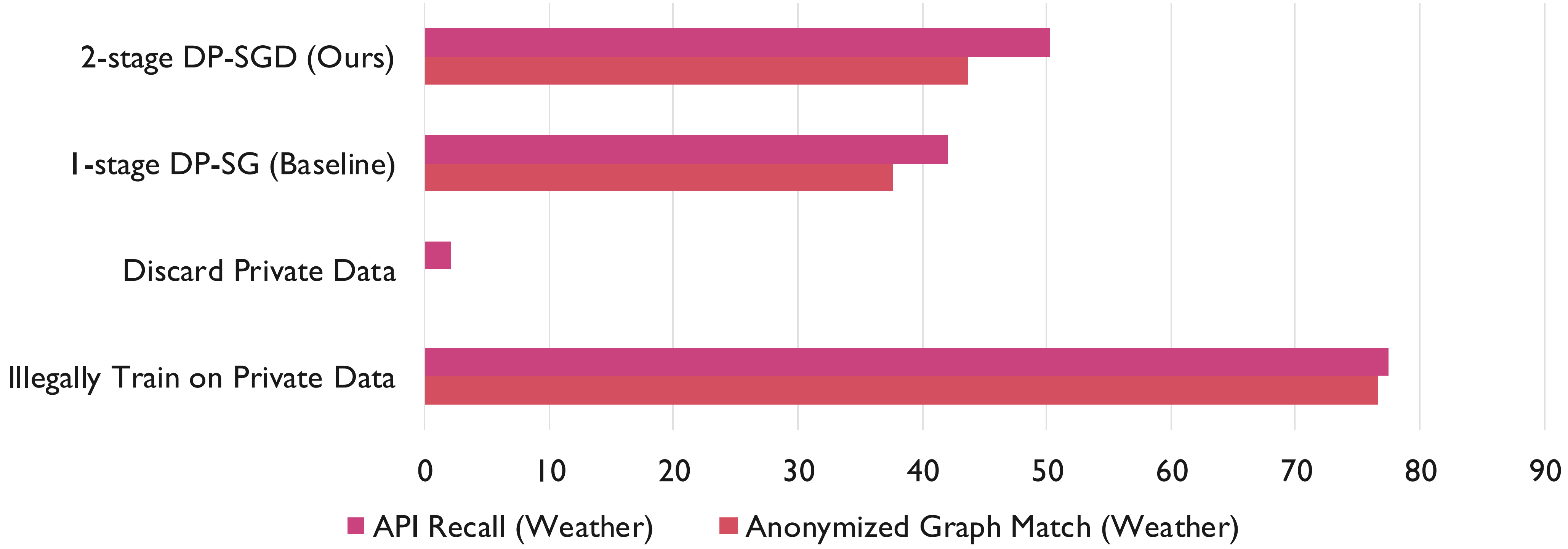


Does This Really Work?

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**

Conclusion and Future Directions

- 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?

Thank you!

fatemeh@ucsd.edu