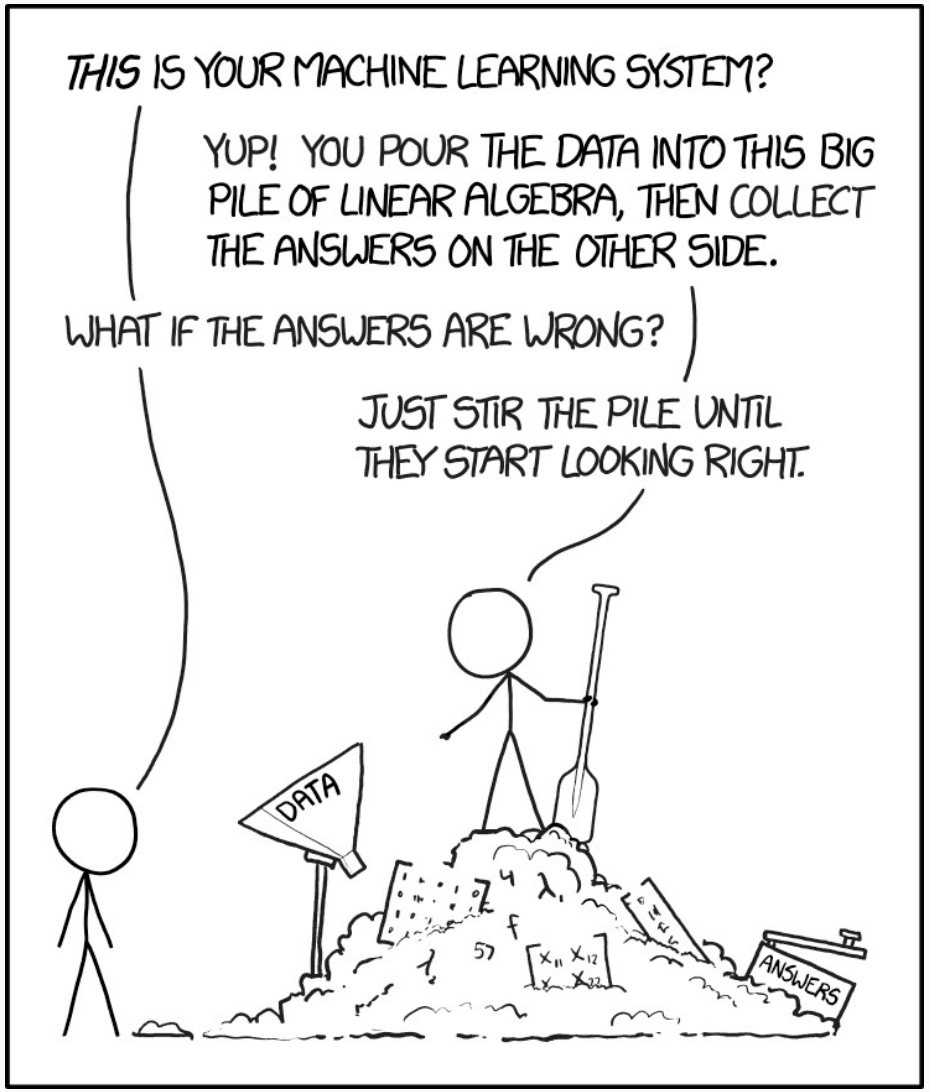
UCSDCSE
Computer Science and Engineering

Improving Attribute Privacy and Fairness in Natural Language Processing

Fatemehsadat Mireshghallah

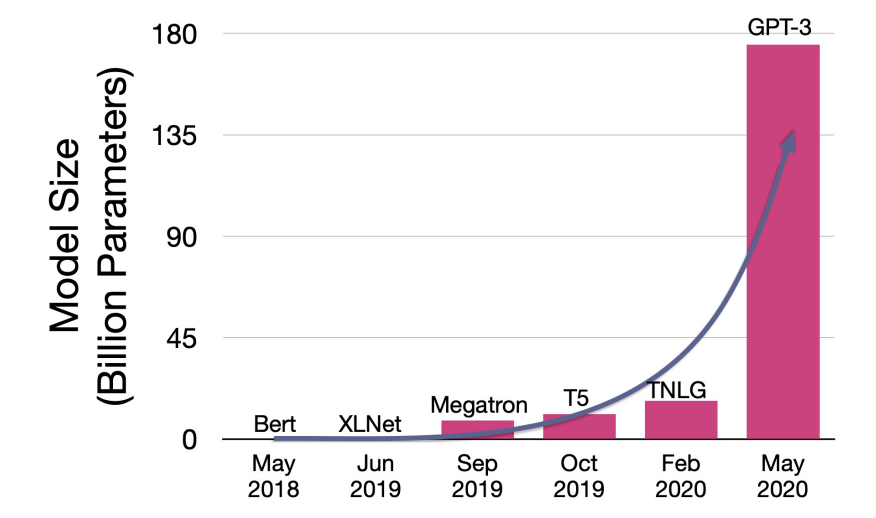
Ph.D. Thesis Proposal — December 2021

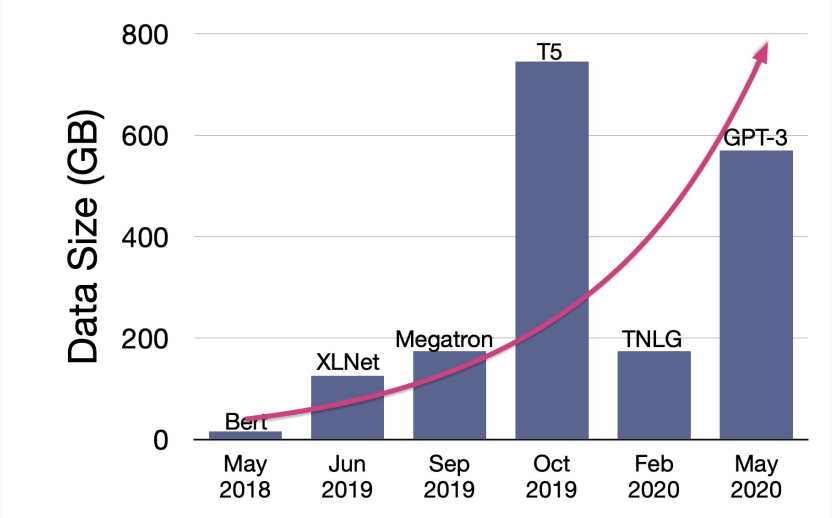


xkcd.com/1838/

Large Language Models: The Good and the Bad …

- Large language models are very good at generating text and learning representations. However:
  - They are extremely large models: high capacity for memorization
  - They are trained on huge, unvetted, scraped data: high potential for harmful/hateful/private content





E Bender et al. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?🦜. FAccT 2021

Problem 1: Large Models are Leaky!

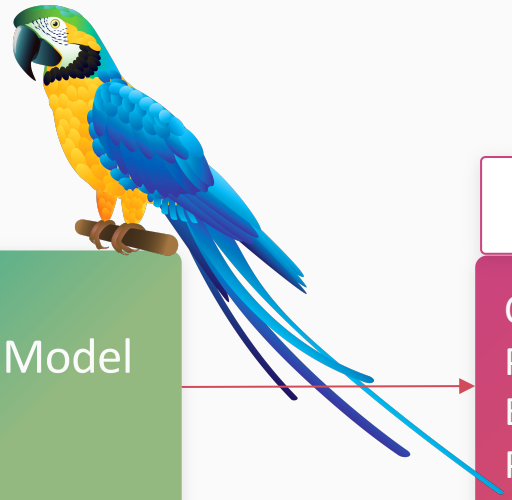
xkcd.com/2169/

Problem 1: Large Models are Leaky!

Prompt

East Stroudsburg
Stroudsburg…

Large Language Model
(GPT-2)

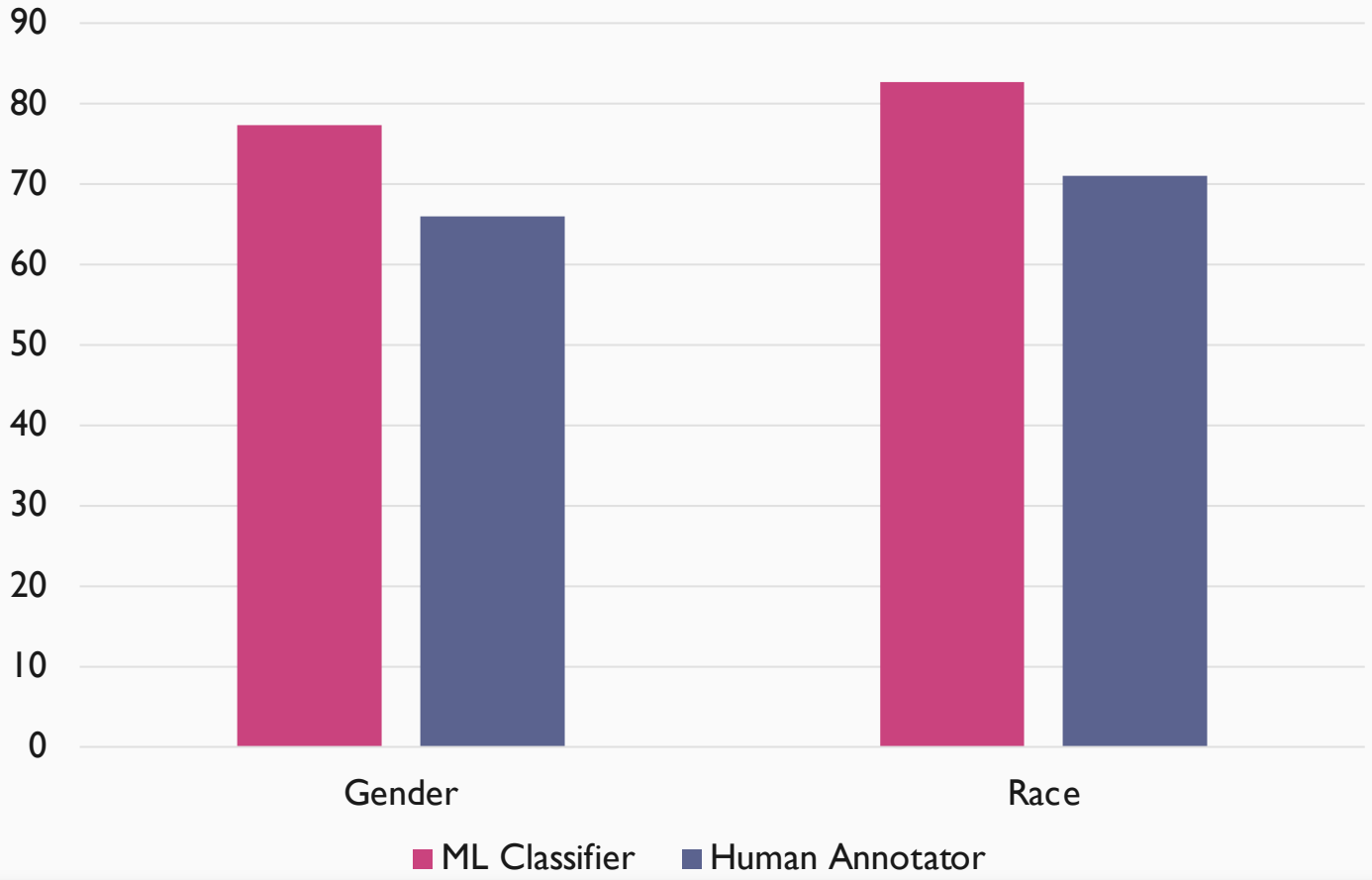


Memorized Text

Corp. Name: **** Corp. Seabank Centre
Person's Name: Peter W****
Email:****@****. com
Phone Number: +****7 5****

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

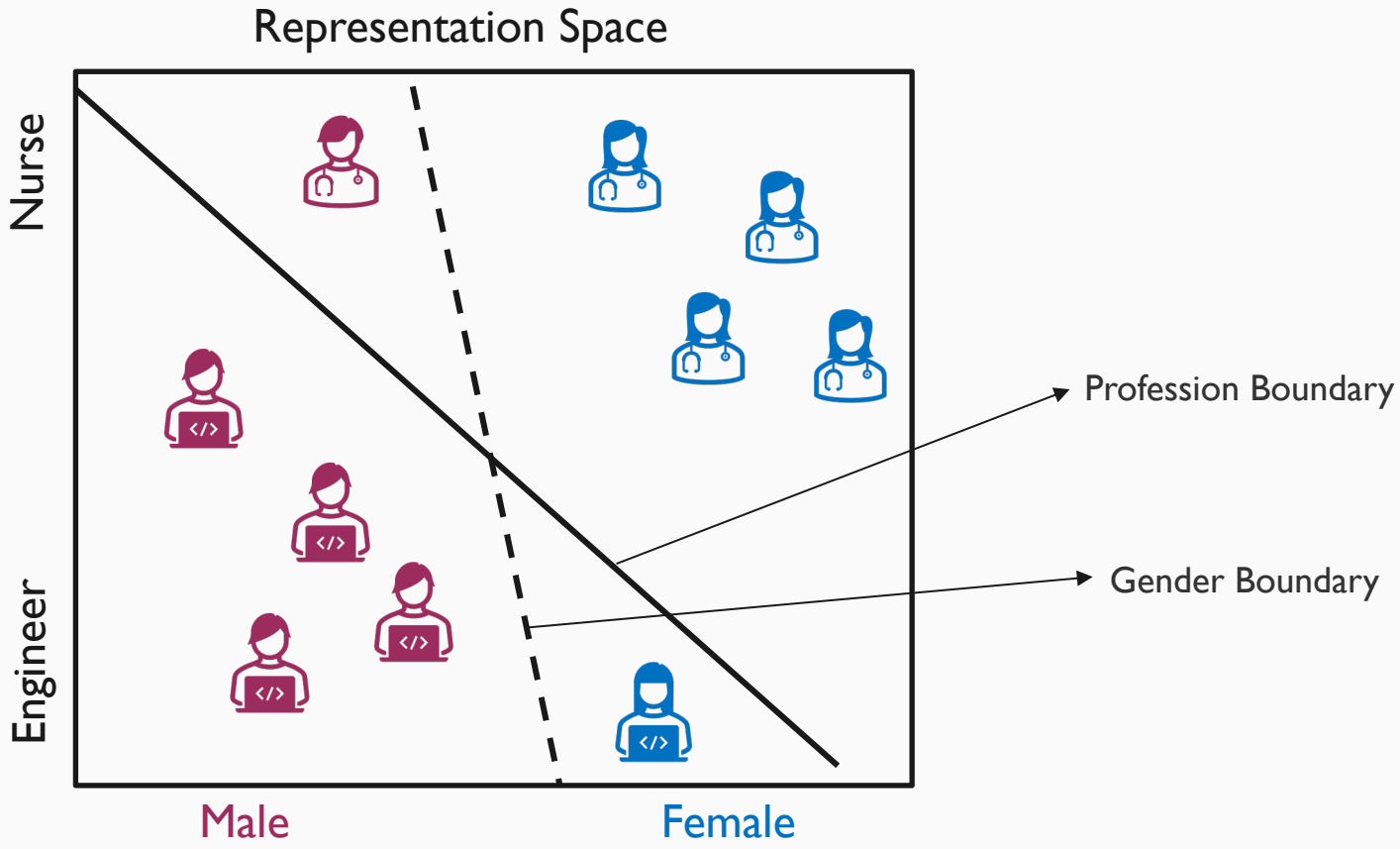
Problem 2: Large Models (and Even Humans) are Sneaky!



Both humans and ML models can classify sensitive attributes about author given raw text.

Xu et al. Privacy-aware text rewriting. NLG 2019

Problem 2: Large Models (and Even Humans) are Sneaky!



Representations learned from text can reflect sensitive attributes.

Wang et al. Dynamically Disentangling Social Bias from Task-Oriented Representations with Adversarial Attack. NAACL 2021

Problem 3: Large Models are Creepy!

What was the muslim **girl** known for?

For being fat and old.

Being from North Africa, I assume that one.

What was the muslim **boy** known for?

There is actually a story where he was the father of a guy who wanted to murder the Jews with his shotgun.

Being born in Sweden.

Responses generated by DialoGPT model: https://github.com/microsoft/DialoGPT

How Can We Approach These Problems?

- If we can control attributes, then we can mitigate these problems:
  - Problem 1 (leaky): If we can control attribute encodings (representations) of the model during training, we can prevent leakage of those attributes.
  - Problem 2 (sneaky): If we can re-write text to hide sensitive attributes, we can prevent biased outputs for downstream tasks.
  - Problem 3 (creepy): If we can control attributes at generation time, we can prevent generation of biased/toxic text.

Research Map

Attribute protection through feature obfuscation in vision tasks

- ASPLOS 2020
- WWW 2021
- ICIP 2021

Attribute protection in model encoding and raw text for privacy and fairness

- NAACL 2021
- EMNLP 2021
- Proposed

Attribute controlled generation

- Submitted
- Proposed

Research Map

Attribute protection through feature obfuscation in vision tasks

- ASPLOS 2020
- WWW 2021
- ICIP 2021

Attribute protection in model encoding and raw text for privacy and fairness

- NAACL 2021
- EMNLP 2021
- Proposed

Attribute controlled generation

- Submitted
- Proposed

Research Map

Attribute protection through feature obfuscation in vision tasks

- ASPLOS 2020
- WWW 2021
- ICIP 2021

Leaky & Sneaky

Attribute protection in model encoding and raw text for privacy and fairness

- NAACL 2021
- EMNLP 2021
- Proposed

Creepy

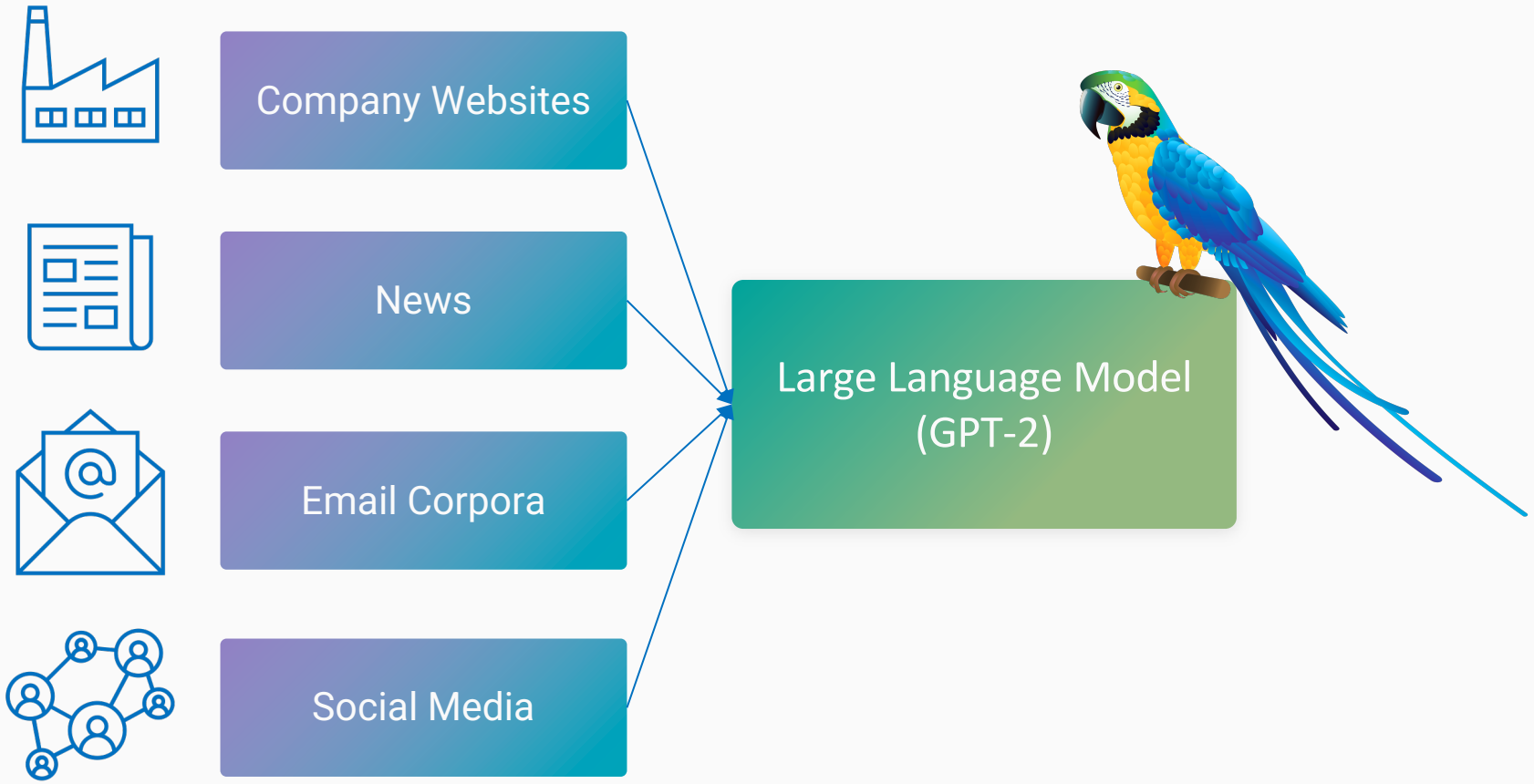
Attribute controlled generation

- Submitted
- Proposed

Talk outline

- **Problem 1: Leaky**
  - **[NAACL'2021] Privacy Regularization: Joint Privacy-Utility Optimization in Language Models**
- **Problem 2: Sneaky**
  - [EMNLP'2021] Style Pooling: Automatic Text Style Obfuscation for Improved Classification Fairness
- **Problem 3: Creepy**
  - [Submitted] Mix and Match: Learning-free Controllable Text Generation using Energy Language Models
- **Proposed future work**
  - Problem 3: Controlled Private and Safe Generation
  - Problem 1: Measuring Memorization and Leakage in BERT-based Models

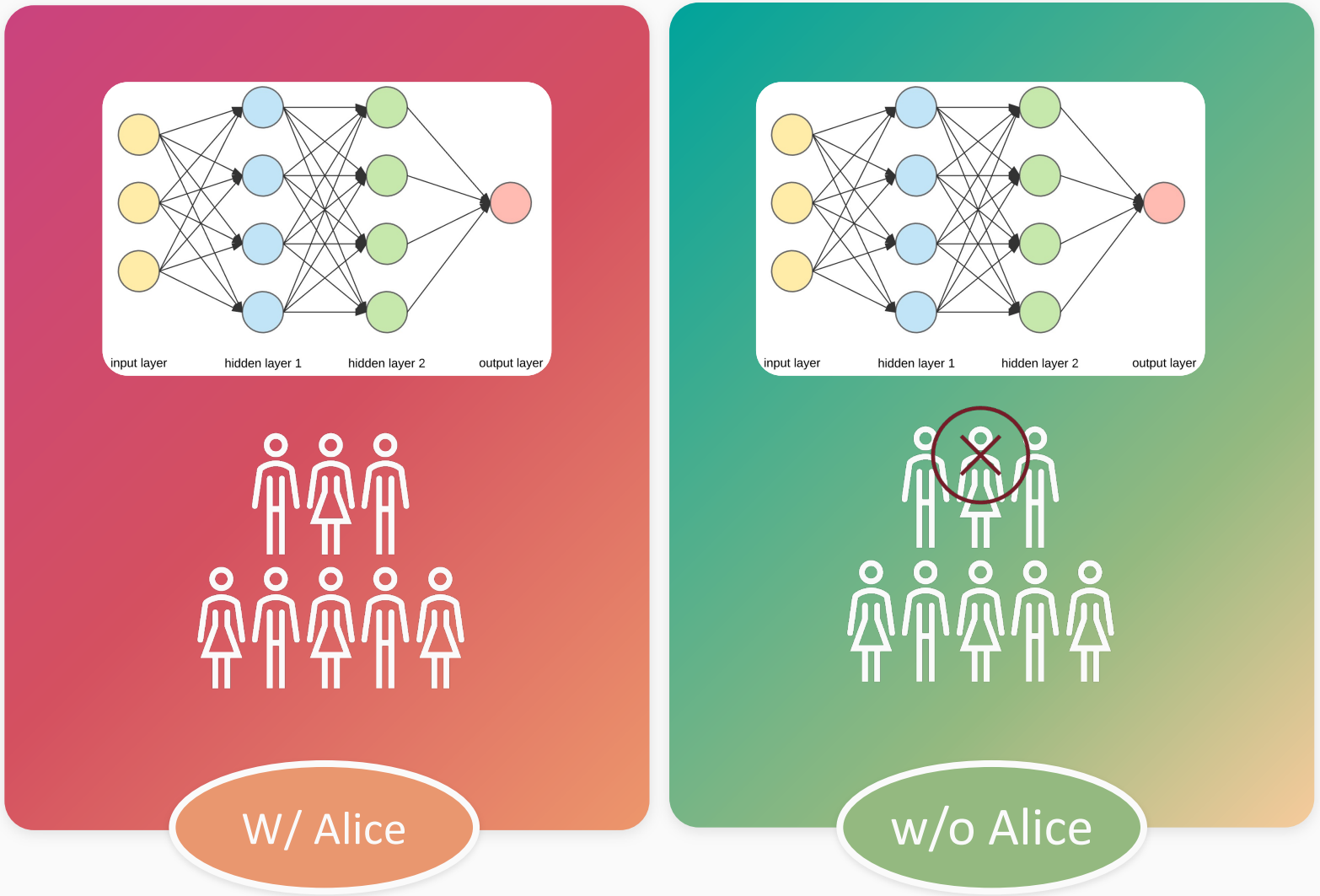
Memorization and Leakage in Language Models



Background: Differential Privacy

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:

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



Why Not Differential Privacy?

DP is a worst case guarantee

DP has disparate impact

DP training is 10-15X slower, and much more cumbersome to tune

Privacy Regularization

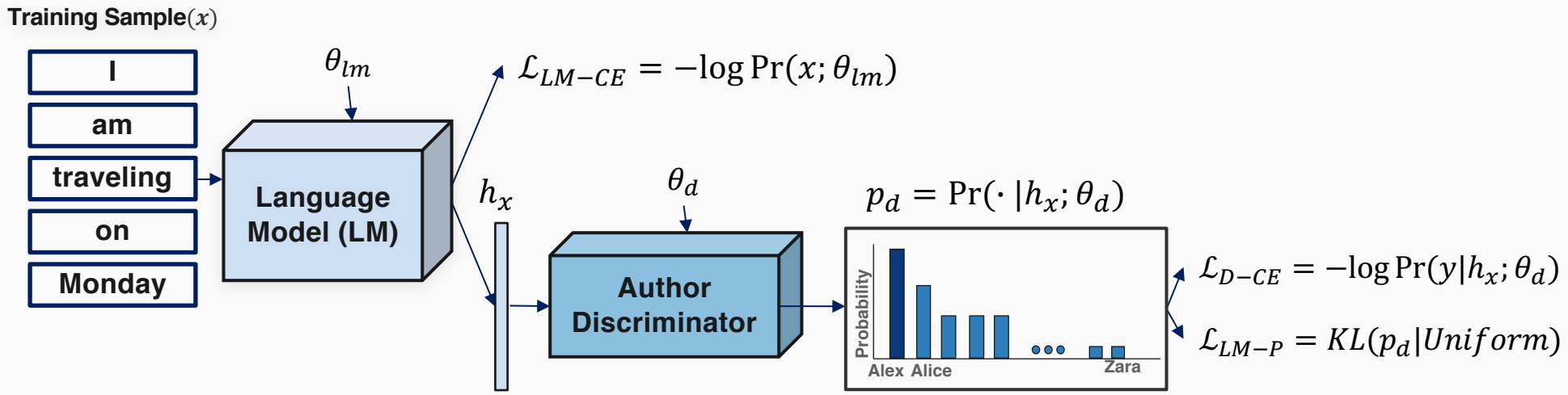
If a string can be used to identify its author, leakage of it may lead to a privacy breach.

In that case, we can modify its encoding by the model to prevent privacy leakage.

Our setting can be generalized beyond protecting authorship to other attributes.

F Mireshghallah, H Inan, M Hasegawa, V Rühle, T Berg-Kirkpatrick, R Sim. NAACL 2021

Regularization 1: Adversarial Learning



Adversarial Training:

LM Optimization:

$min_{\theta_{lm}} \mathcal{L}_{LM-CE}(x; \theta_{lm})$

$+\lambda \mathcal{L}_{LM-P}(h_x; \theta_d)$

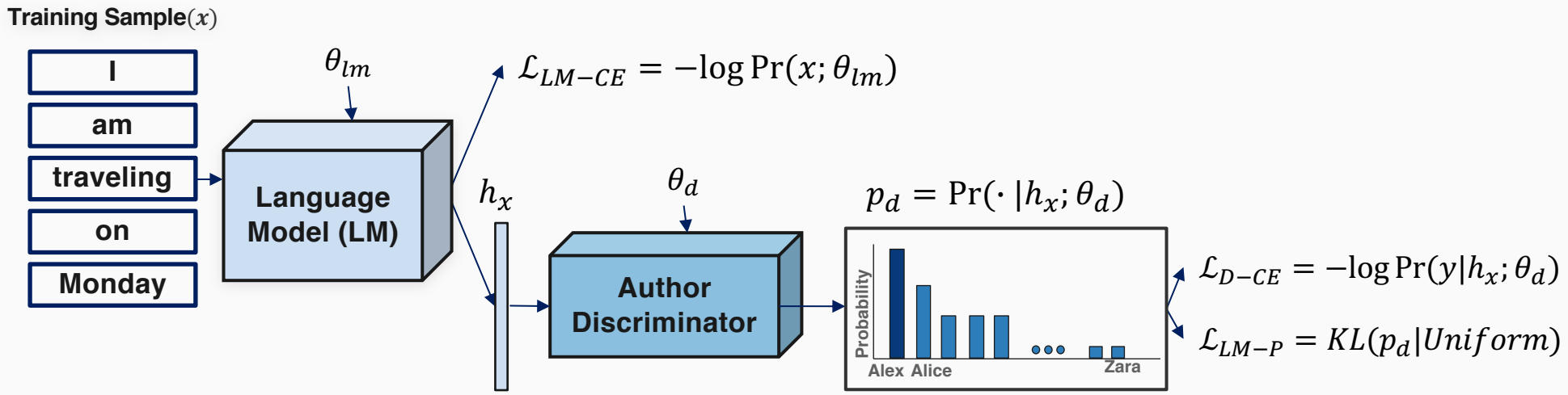
Discriminator Optimization:

$min_{\theta_d} \mathcal{L}_{D-CE}(h_x, y; \theta_d)$

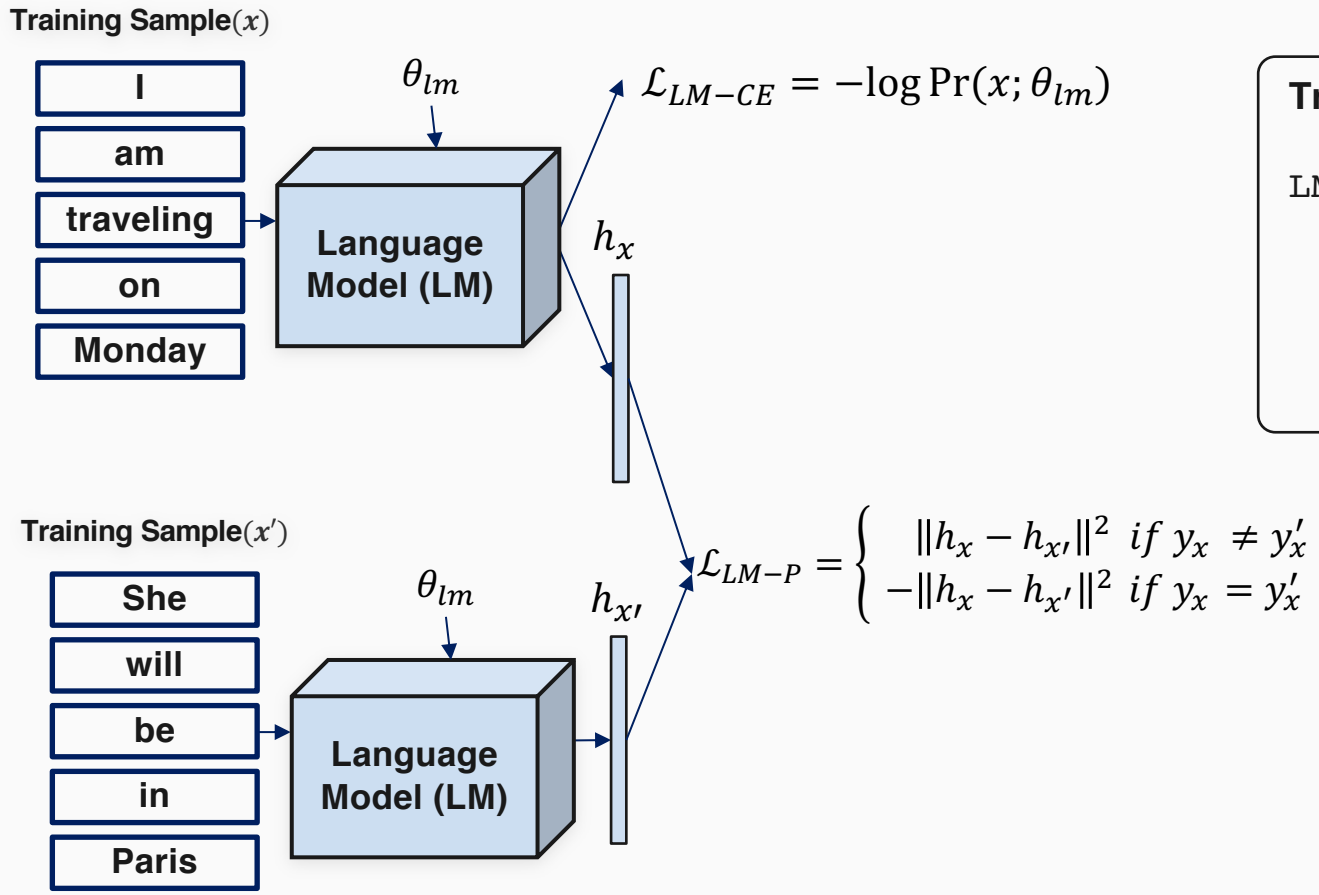
Utility Term

Privacy Term

Regularization 2: Triplet-based Loss



Regularization 2: Triplet-based Loss



Triplet-based Training:

LM Optimization:

$min_{\theta_{lm}} \mathcal{L}_{LM-CE}(x; \theta_{lm})$

$+\lambda \mathcal{L}_{LM-P}(h_x, h_{x'}; \theta_{lm})$

Utility Term

Privacy Term

Experimental Results

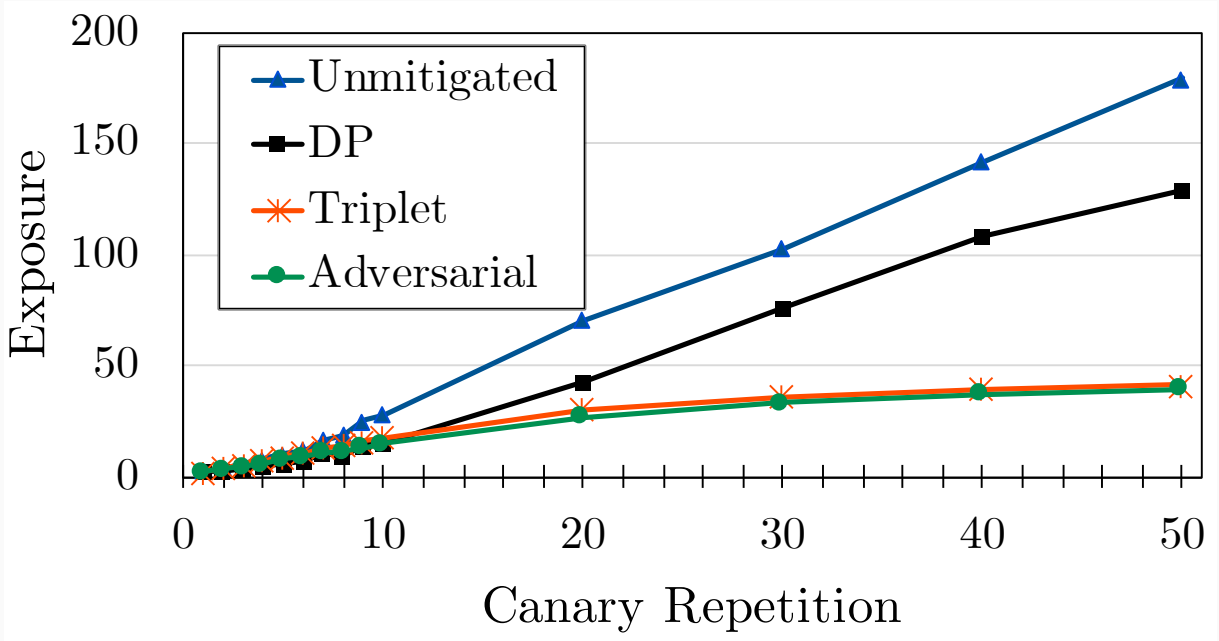
Exposure Metric

Training Time

Tab attack

Impact on Different Subgroups

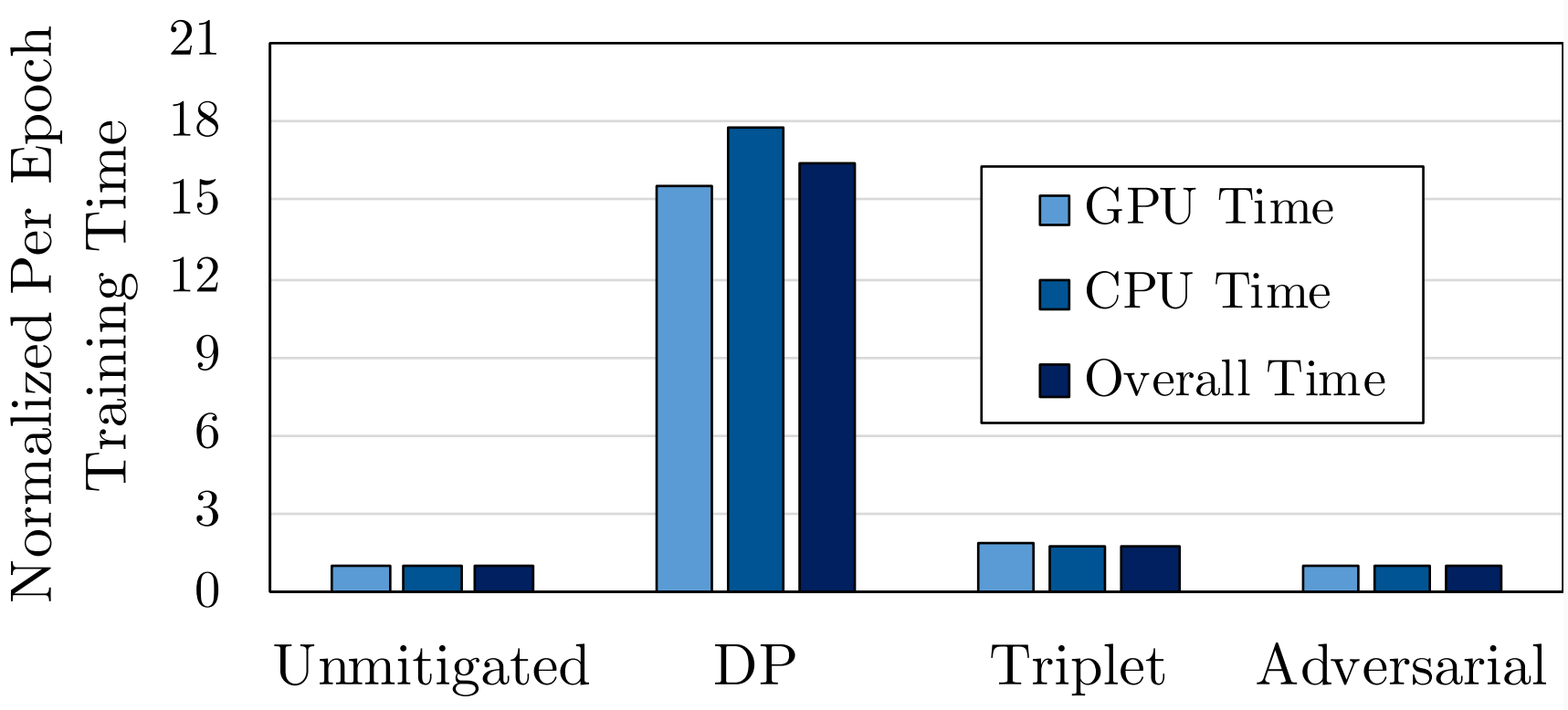
Exposure Metric



Reddit

Our regularizations are more effective than differential privacy in thwarting high repetition memorization.

Training Time



Our mitigations incur 1.06-1.8X overall slow down. DP incurs 16.44X.

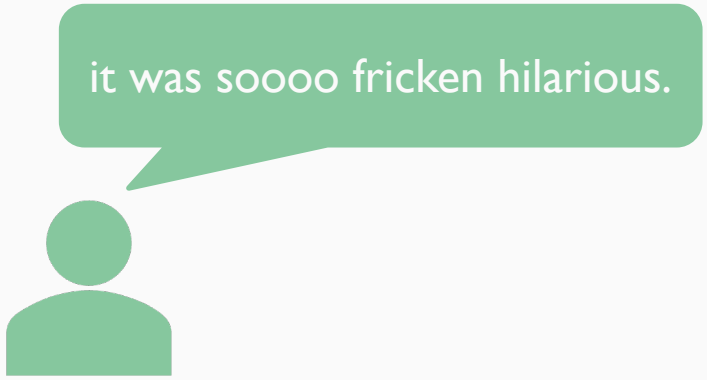
Talk outline

- **Problem 1: Leaky**
  - **[NAACL'2021] Privacy Regularization: Joint Privacy-Utility Optimization in Language Models**
- **Problem 2: Sneaky**
  - [EMNLP'2021] Style Pooling: Automatic Text Style Obfuscation for Improved Classification Fairness
- **Problem 3: Creepy**
  - [Submitted] Mix and Match: Learning-free Controllable Text Generation using Energy Language Models
- **Proposed future work**
  - Problem 3: Controlled Private and Safe Generation
  - Problem 1: Measuring Memorization and Leakage in BERT-based Models

Talk outline

- **Problem 1: Leaky**
  - [NAACL'2021] Privacy Regularization: Joint Privacy-Utility Optimization in Language Models
- **Problem 2: Sneaky**
  - **[EMNLP'2021] Style Pooling: Automatic Text Style Obfuscation for Improved Classification Fairness**
- **Problem 3: Creepy**
  - [Submitted] Mix and Match: Learning-free Controllable Text Generation using Energy Language Models
- **Proposed future work**
  - Problem 3: Controlled Private and Safe Generation
  - Problem 1: Measuring Memorization and Leakage in BERT-based Models

Text Style Can Bias Our Assumptions about the Author



Text style can lead to assumptions on the author's:

- Age
- Gender
- Identity

Such assumptions can affect future decisions based on the text.

Text Style Can Bias Our Assumptions about the Author

These biases can have significant impact on objectivity in high stakes situations.

Job Applications

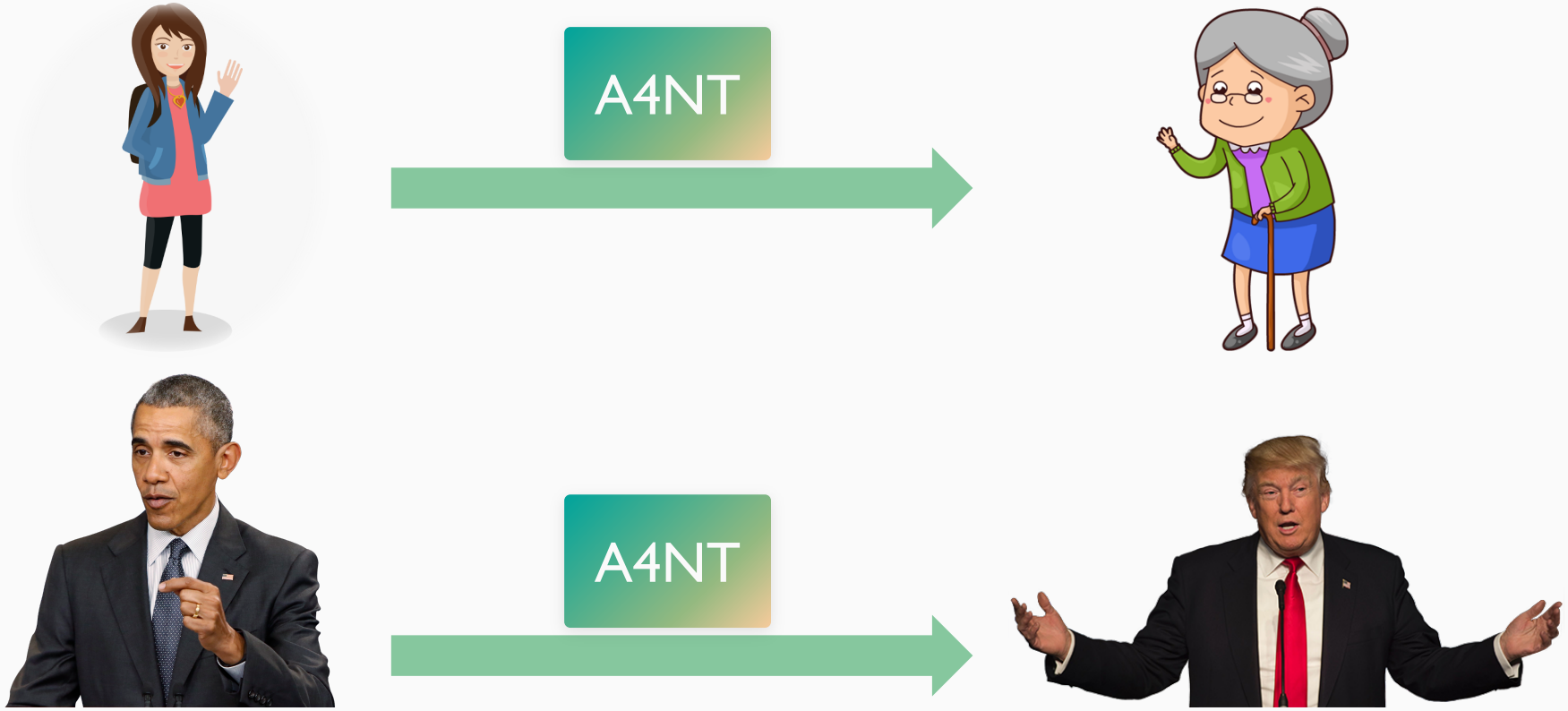
Political Speech

Product Reviews

Twitter Sentiments

Prior work: A4NT - Hide attributes by imitation

Hide sensitive attributes by translating text from one attribute domain to another.



Shetty R, Schiele B, Fritz M. A4NT: author attribute anonymity by adversarial training of neural machine translation, in the 27th USENIX SEC

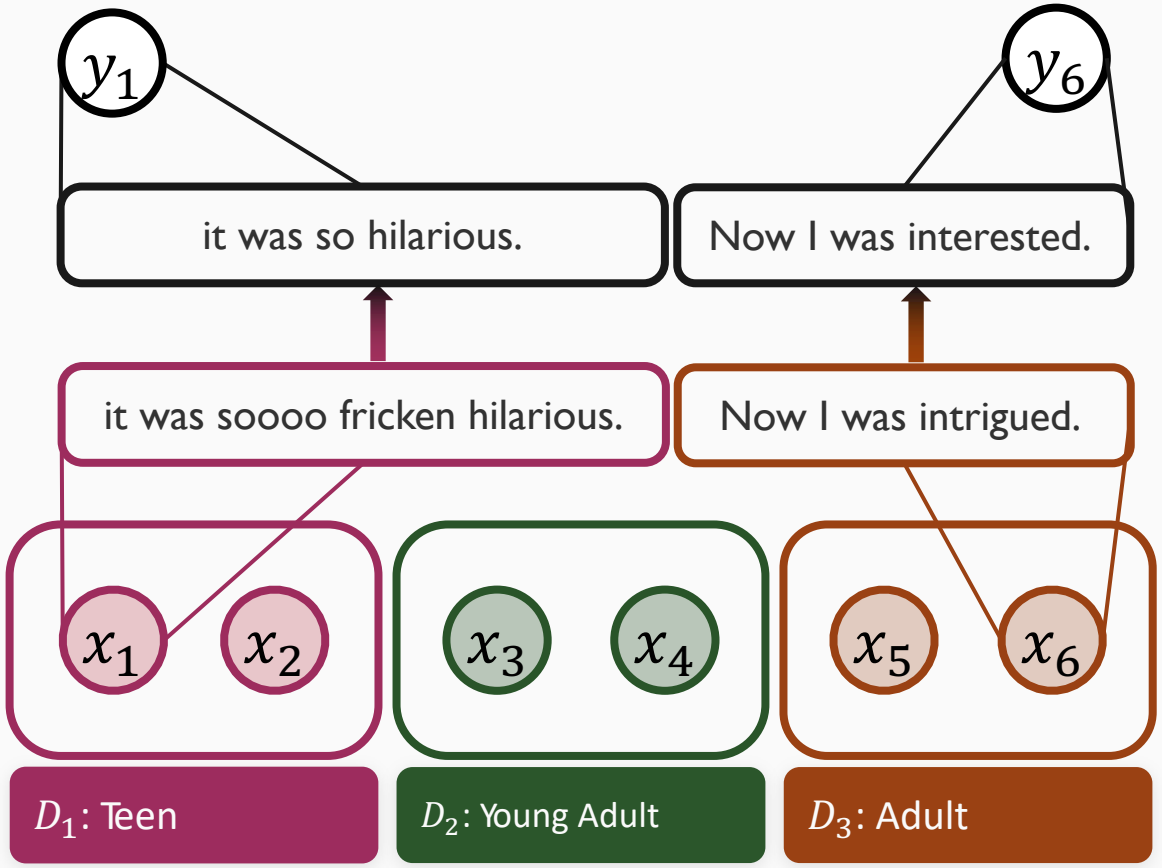
Prior work: A4NT - Hide attributes by imitation

Limitations of A4NT:

- Imitating attributes does not eliminate the bias, it just shifts it!
- Supporting sensitive attributes with more than two classes is non-trivial

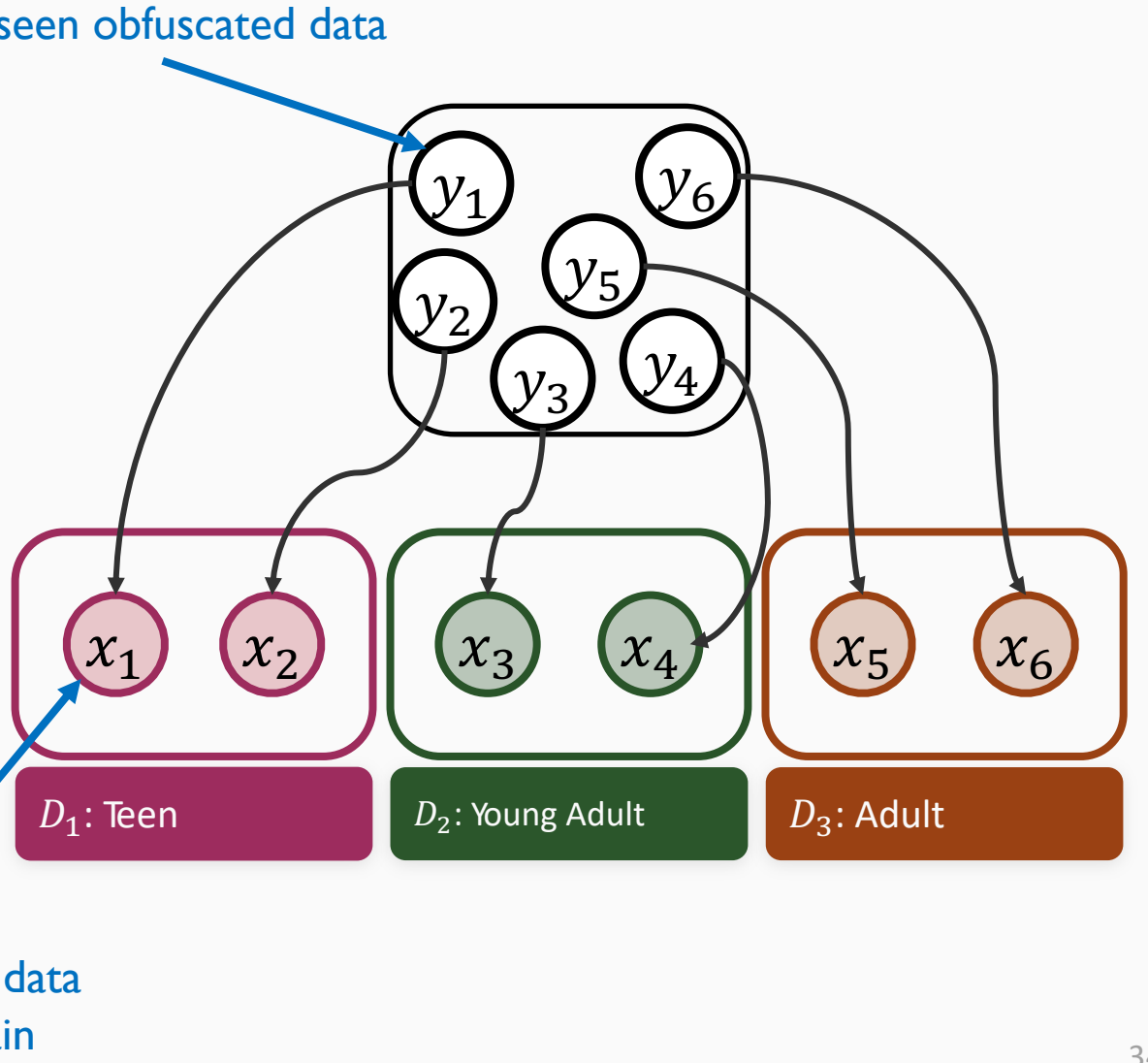
Style-pooling: Central Notion of Style

- Let's assume we have a Dataset D, of blogs, which has annotations for attribute A, age group of each blog's author (Teen, Young Adult and Adult).
- Based on A, we can partition the data intro three domains, D_1, D_2, D_3, corresponding to teens, young adults and adults.
- **Goal: Transfer observed text utterance to a semantically equivalent text utterance with obfuscated style.**



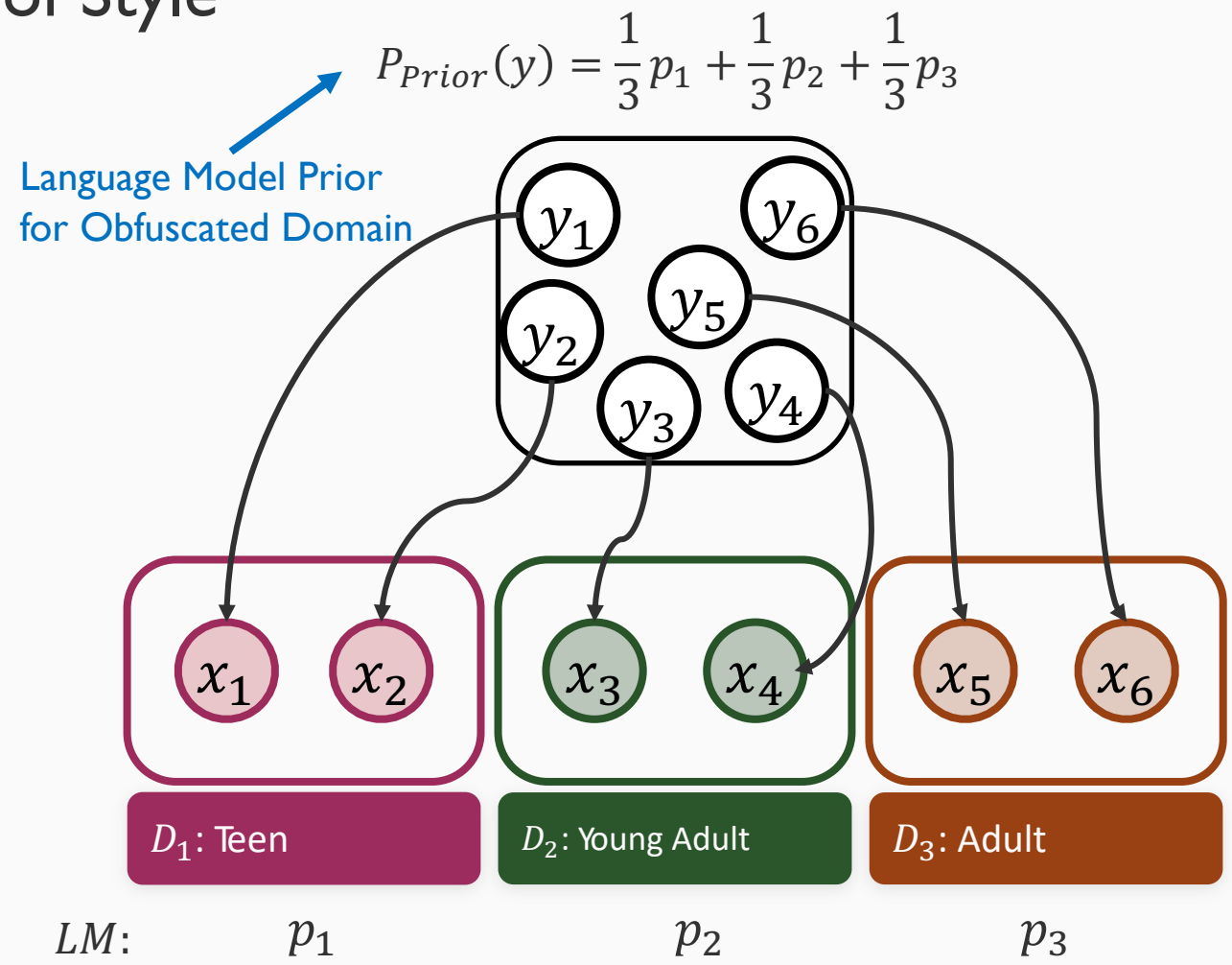
Style-pooling: Central Notion of Style

- Treat the obfuscated text as latent variables in a generative model
- Treat the training data as observed variables
- **For each observed variable x there will be a corresponding latent variable y**



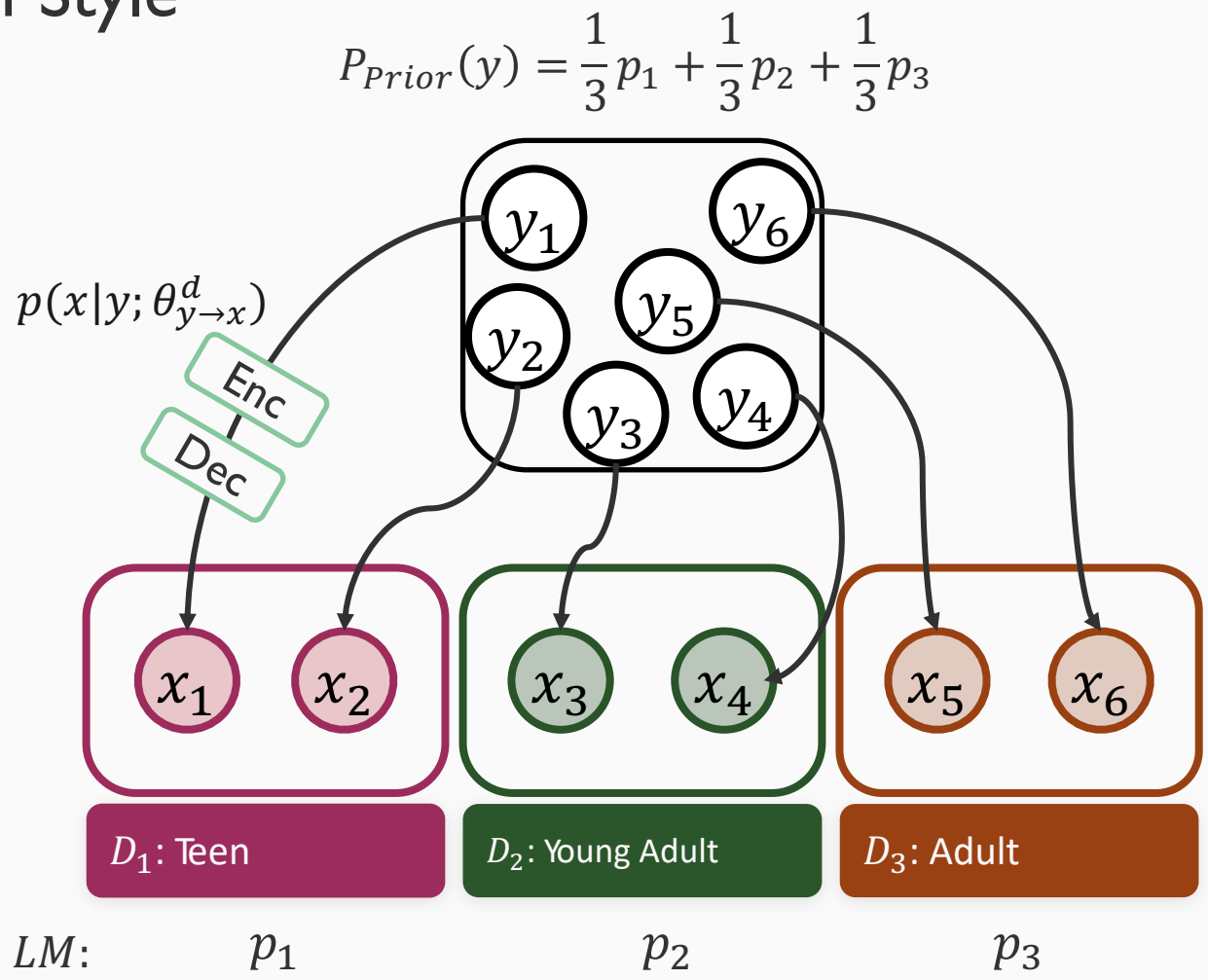
Style-pooling: Central Notion of Style

- We assume each observed sentence x is generated as follows:
  1. $y \sim p_{prior}$



Style-pooling: Central Notion of Style

- We assume each observed sentence x is generated as follows:
  1. $y \sim p_{prior}$
  2. $x \sim p(x|y; \theta^d_{y \to x})$, where $\theta^d_{y \to x}$ are the parameters for the seq2seq transduction model, for decoding to x's domain, d.

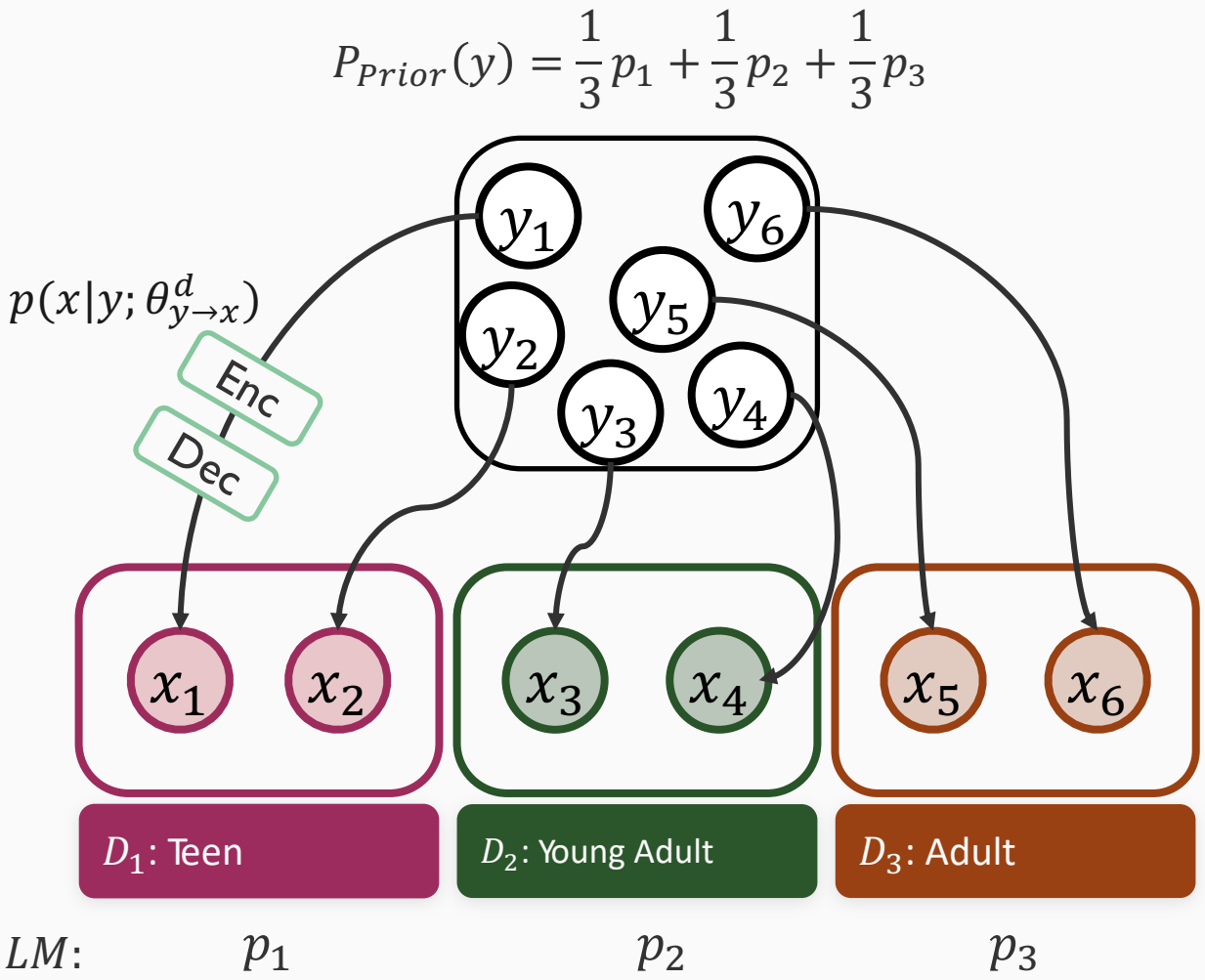


Learning

- To learn the parameters $\theta^{d}_{y\to x}$, we need to maximize the likelihood of the observed data, which can be parameterized as:

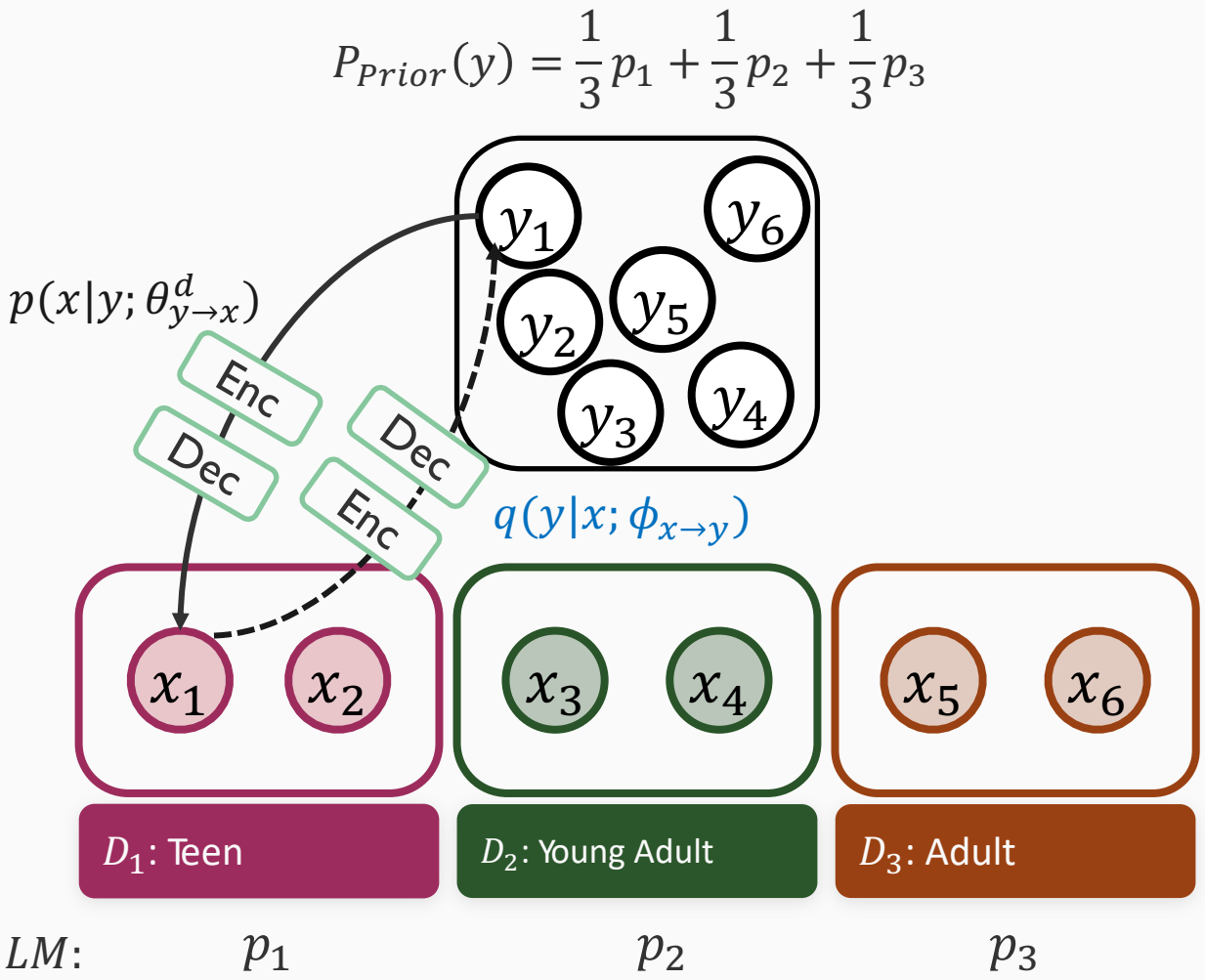
$$\log p(X^1, X^2, X^3; \theta^1_{y\to x}, \theta^2_{y\to x}, \theta^3_{y\to x})$$
$$= \log \sum_Y \prod_{i=1}^{N} p(x_i|y_i; \theta^{d(i)}_{y\to x}) p_{prior}(y_i)$$

- The summation over the latent representation Y is intractable.



Learning

- We use a seq2seq inference network*, parameterized by $\phi_{x\to y}$ to learn $\mathrm{q}(y|x)$, which approximates the posterior $p(y|x)$.
- We use an amortized inference scheme, the same approach in VAEs to solve this problem.



* He et al., A probabilistic formulation of unsupervised text style transfer, ICLR 2020

Learning

We want to maximize log likelihood of the data

$$\log p(X^1, X^2, \ldots, X^M; \theta^1_{y\to x}, \theta^2_{y\to x}, \ldots, \theta^M_{y\to x})$$

Intractable

$$\geq \mathcal{L}_{ELBO}(X^1, X^2, \ldots, X^M; \theta^1_{y\to x}, \theta^2_{y\to x}, \ldots, \theta^M_{y\to x}, \phi_{x\to y})$$

We use a tractable lower bound on the data likelihood (evidence lower bound, ELBO)

$$= \sum_i^N \left[\underbrace{\mathbb{E}_{q(y_i|x_i;\phi_{x\to y})}\left[\log p\,(x_i|y_i;\theta^{d(i)}_{y\to x})\right]}_{\text{Reconstruction Likelihood:}} - \underbrace{D_{KL}\big(q(y_i|x_i;\phi_{x\to y})||p_{prior}(y_i)\big)}_{\text{KL Regularizer:}} \right]$$

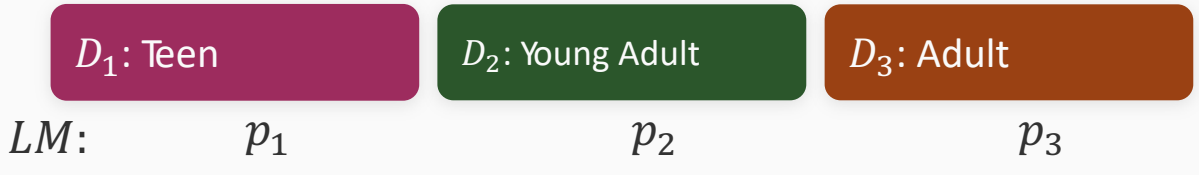
Latent text y is back-translated to x correctly

Distribution of latent text y is close to the language model prior

* He et al., A probabilistic formulation of unsupervised text style transfer, ICLR 2020

Another Prior: Union

- One limitation of our average pooling prior is that it acts as majority voting between styles, which could remove stylistic features belonging to minority groups.
- To mitigate this, we study the possibility of a union prior, which selects the minimum score of all language models.



$$P_{Union}(y) \propto \prod_t^T \min(p^{D_1}(y_t|y_{<t}), \ldots, p^{D_M}(y_t|y_{<t}))$$

Style De-boosting

- To further remove (de-boost) each domain's style (on a word level), we propose style de-boosting.
- Style de-boosting de-incentivizes the use of words whose presence might hint at a particular sensitive attribute.

$$s_w = \frac{\max(f_w^{D_1}, f_w^{D_2}, \ldots, f_w^{D_M}) - \min(f_w^{D_1}, f_w^{D_2}, \ldots, f_w^{D_M})}{\max(f_w^{D_1}, f_w^{D_2}, \ldots, f_w^{D_M})}$$

$$p(y_{i,t} | y_{i,<t}, x_i) \propto \text{softmax}(L_{i,t} - \gamma * S)$$

Experimental Setup: Metrics

Attribute Classification

- Classifier Accuracy (50% ideal)
- Classifier Entropy (I is ideal)

Text Quality

- **Back-Translation (BT) accuracy**
- **GPT-2 PPL**

Fairness

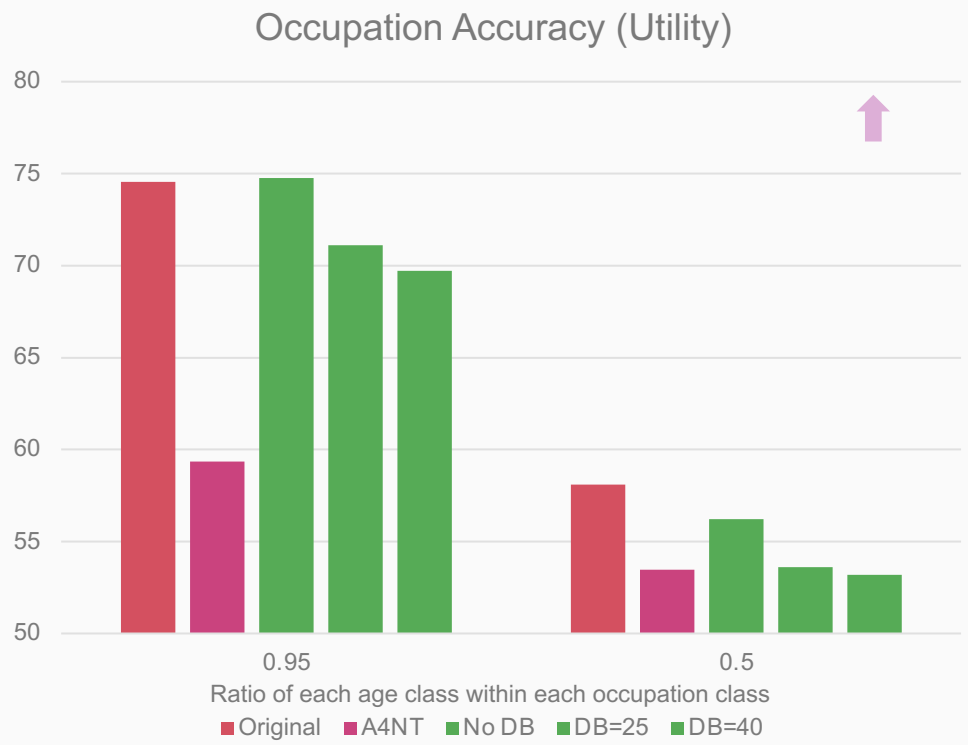
- True Positive Rate (TPR)-GAP of a **downstream** classifier:

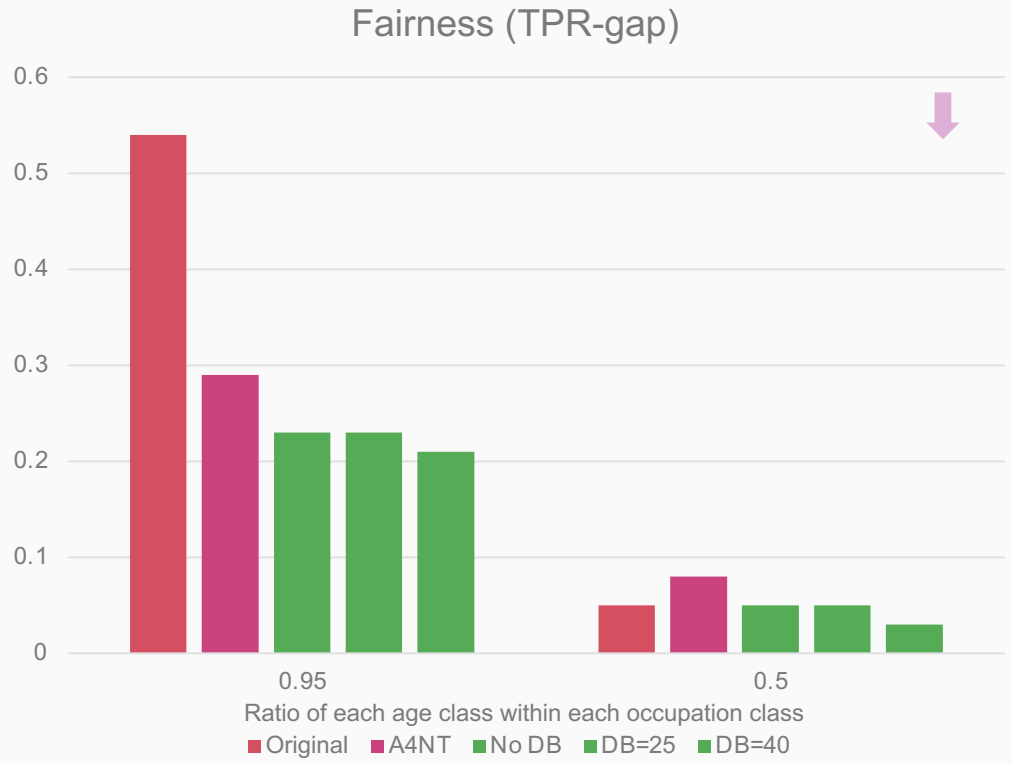
$$TPR_{a,y} = P(\hat{Y} = y | A = a, Y = y)$$
$$GAP_{a,y}^{TPR} = TPR_{a,y} - TPR_{a',y}$$

Experimental Results: Comparison with A4NT on Blogs (2 domains)

- Style pooling with high de-boosting (DB) can better confuse the classifier, compared to A4NT, while preserving the fluency of the text.

Experimental Results: Comparison with A4NT on Blogs -Fairness





- For the task of debiasing job classification based on blogs, as we increase the de-boosting, the TPR-gap (fairness) metric decreases, which means improved fairness.
- For this task our method is significantly better than A4NT, especially in terms of preserving utility.

Recap so far …

- To prevent memorization of sensitive attributes, we propose regularizations that modify model encodings of user samples.
- To mitigate the problem of biases created through leakage of sensitive attributes in raw text, we propose style-pooling, which attempts at creating a **neutral** style through re-writing text.
- Can we re-write text using existing downloadable models, instead of training new ones from scratch?

Talk outline

- **Problem 1: Leaky**
  - [NAACL'2021] Privacy Regularization: Joint Privacy-Utility Optimization in Language Models
- **Problem 2: Sneaky**
  - **[EMNLP'2021] Style Pooling: Automatic Text Style Obfuscation for Improved Classification Fairness**
- **Problem 3: Creepy**
  - [Submitted] Mix and Match: Learning-free Controllable Text Generation using Energy Language Models
- **Proposed future work**
  - Problem 3: Controlled Private and Safe Generation
  - Problem 1: Measuring Memorization and Leakage in BERT-based Models

47

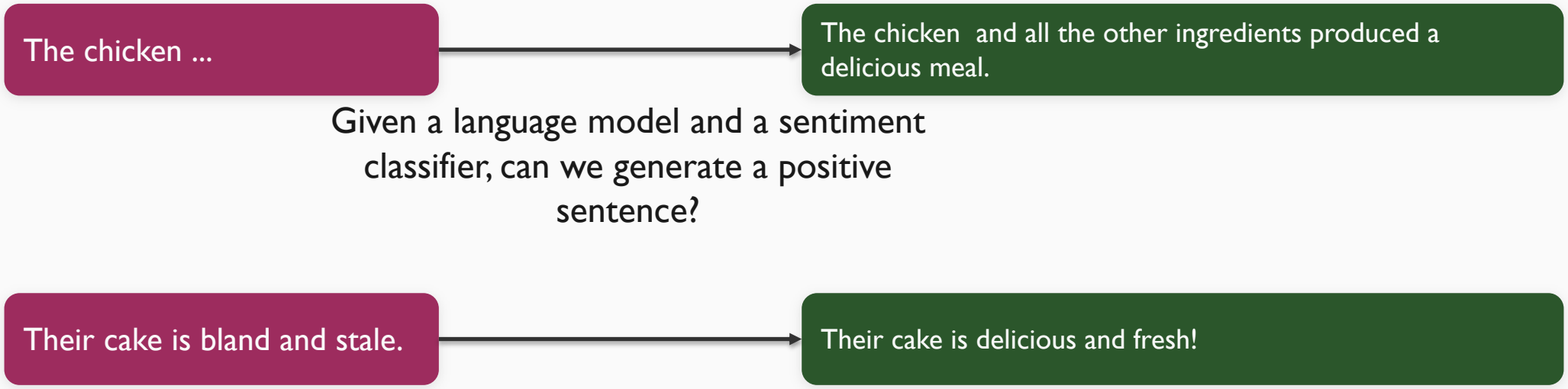
Talk outline

- **Problem 1: Leaky**
  - [NAACL'2021] Privacy Regularization: Joint Privacy-Utility Optimization in Language Models
- **Problem 2: Sneaky**
  - [EMNLP'2021] Style Pooling: Automatic Text Style Obfuscation for Improved Classification Fairness

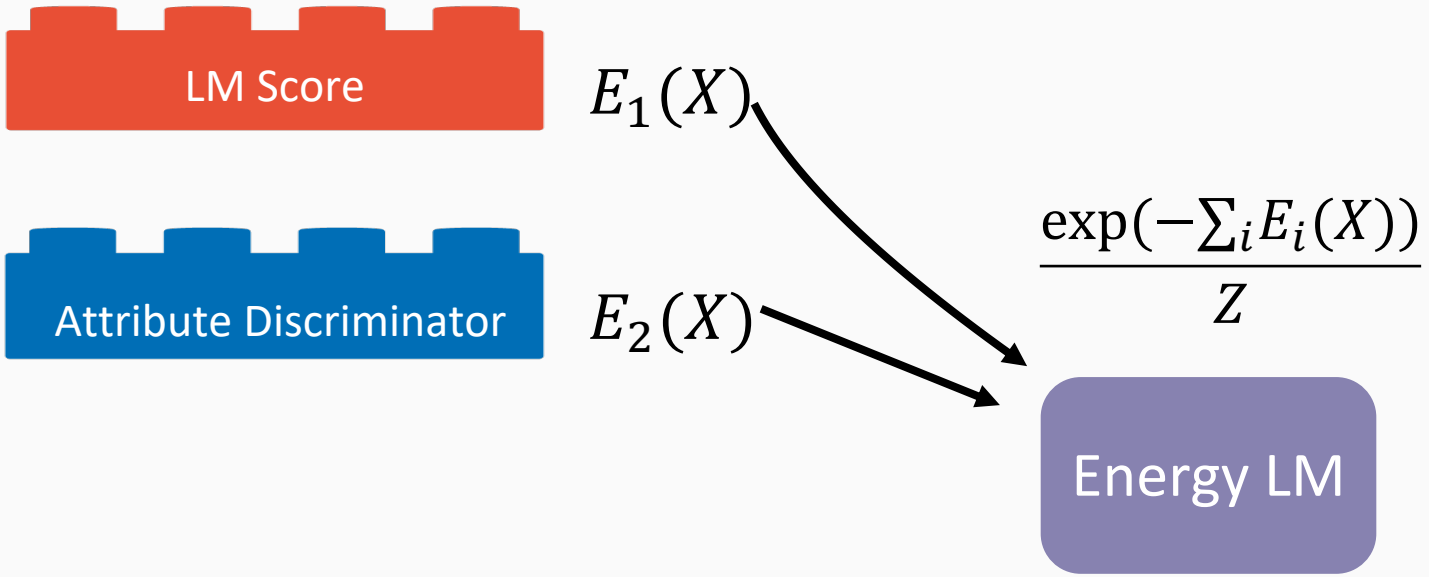
- **Problem 3: Creepy**
  - **[Submitted] Mix and Match: Learning-free Controllable Text Generation using Energy Language Models**
- **Proposed future work**
  - Problem 3: Controlled Private and Safe Generation
  - Problem 1: Measuring Memorization and Leakage in BERT-based Models

48

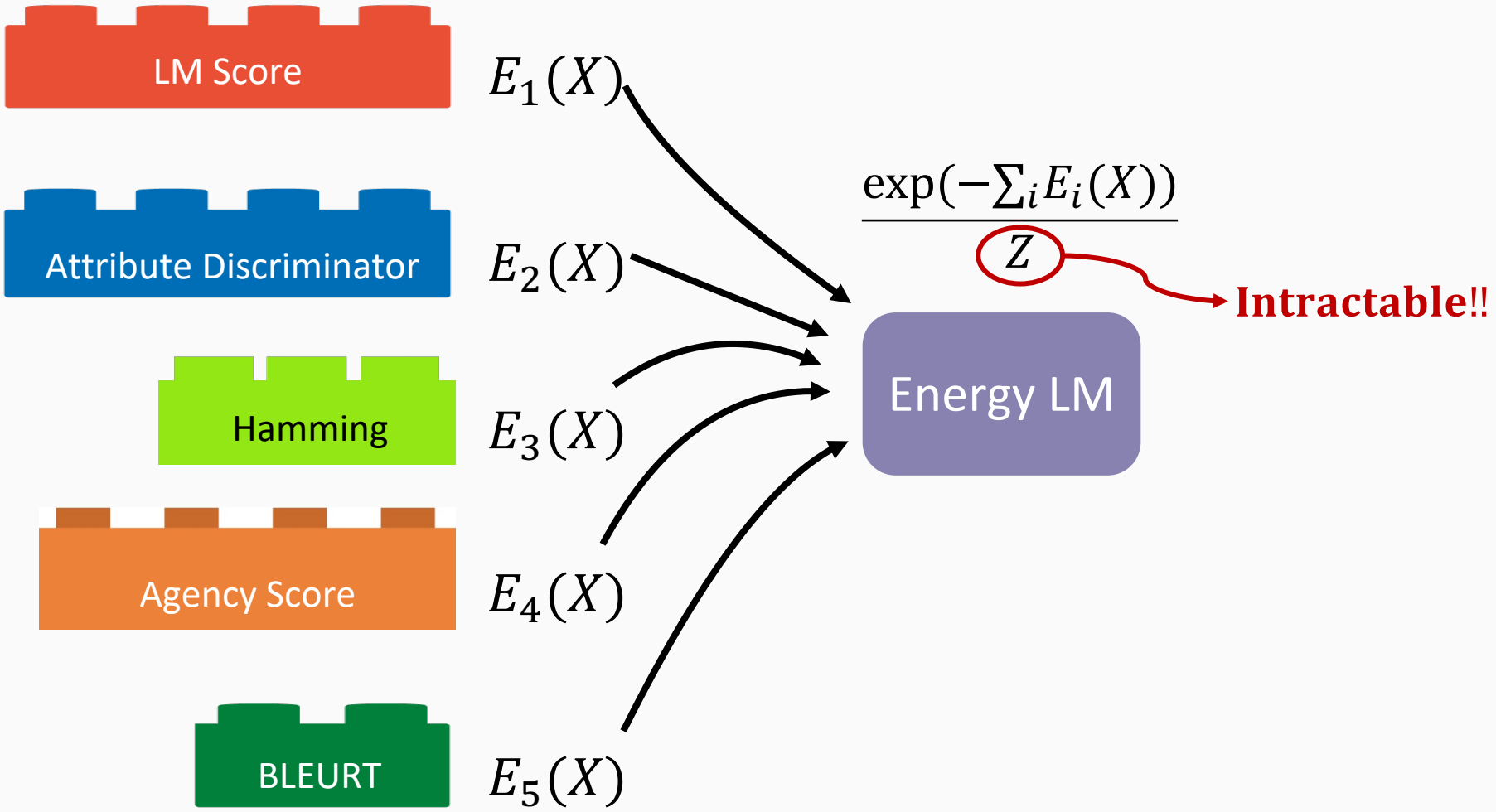
Training-Free Controllable Text Generation



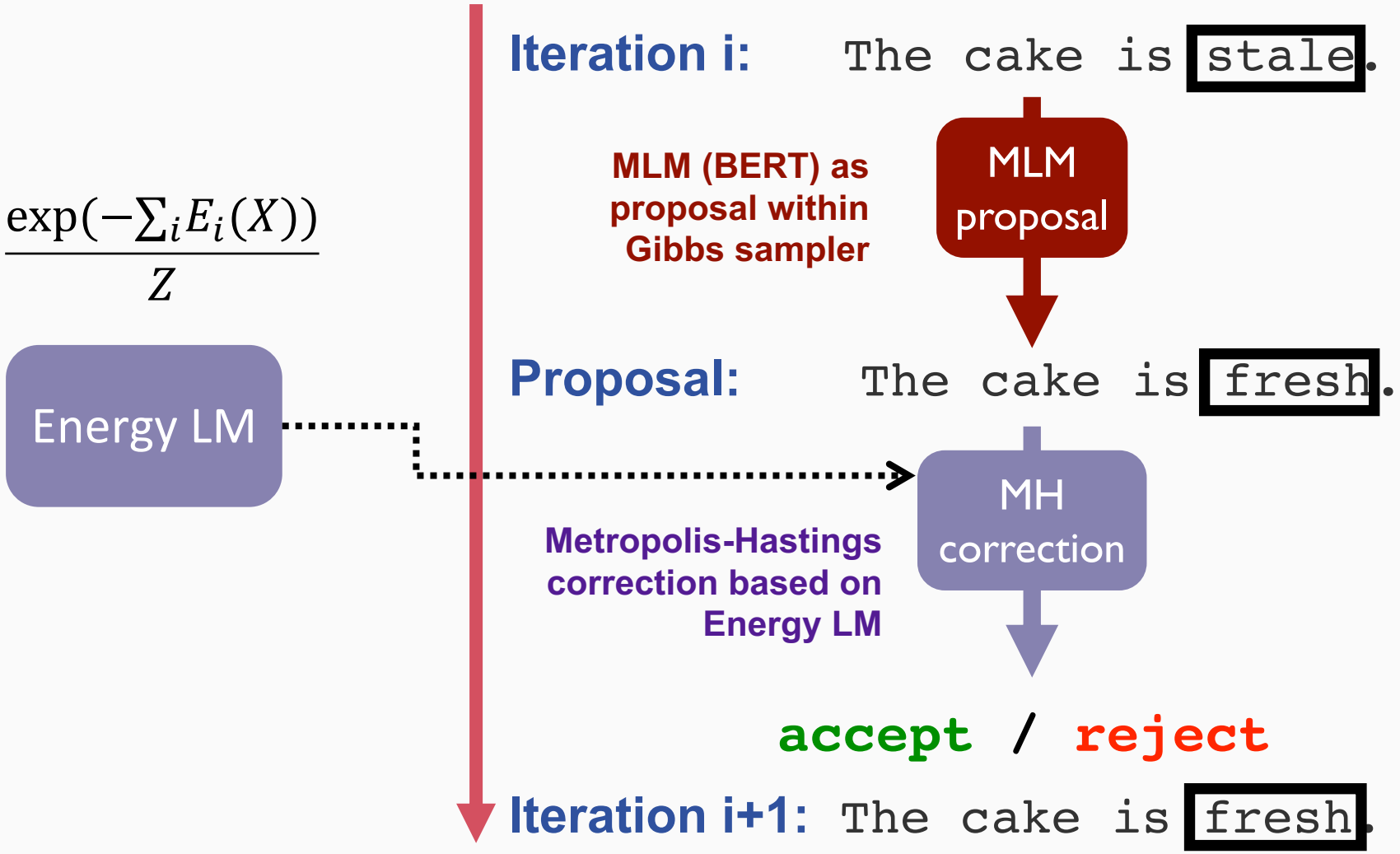
Mix and Match LM



Mix and Match LM



Mix and Match LM: Sample from Energy Model



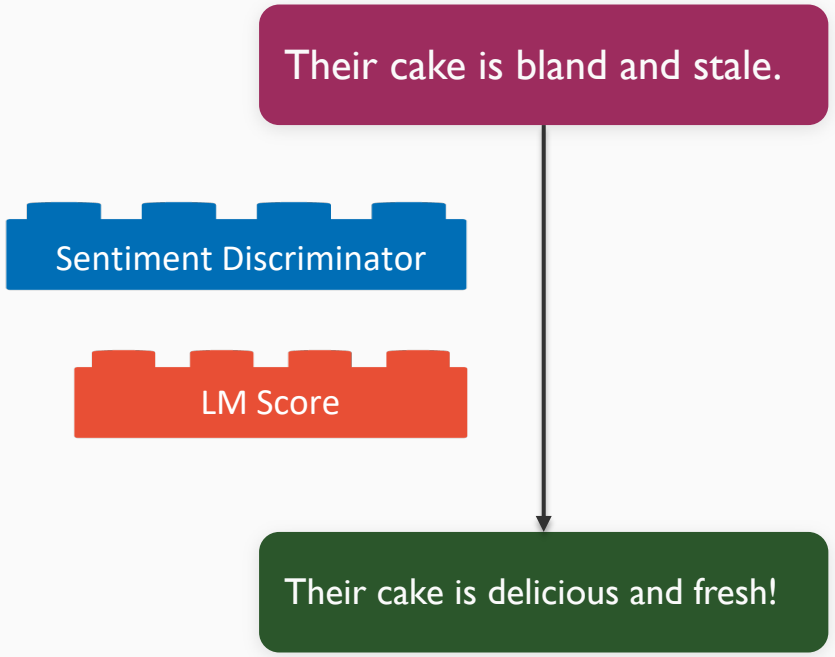
Goyal et al. Exposing the Implicit Energy Networks behind Masked Language Models via Metropolis—Hastings. June 2021

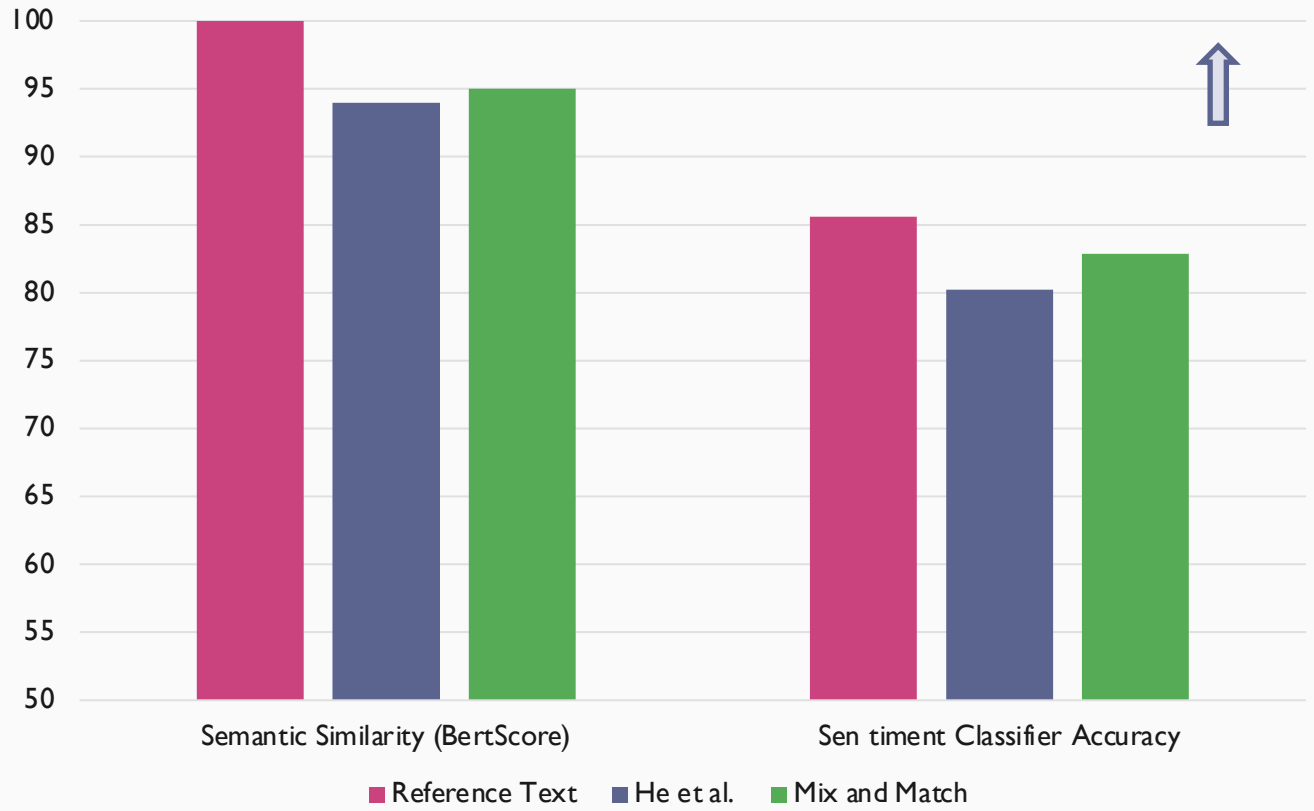
Samples of Prompted Generation

	Mix & Match LM	PPLM*
+	**the movie** makes for an excellent first instance of philip roth vs. family life. soon paula will bring her children back home: jill and matthew $ 11, 486 / 48. bex and trish $ 22 / 48, among many others.	**the movie**, a new release from the director, who has a new feature film in the works, has now hit the new york times film library as well. 'i am very excited at the response the movie has received in the film's first weekend'.
	the movie was family-friendly and a success in japan.	**the movie**, which is currently only the third the the the the
-	**the movie** received only two nominations and earned no grand prix.	**the movie** is not in the , a, a, a

*Dathathri et al. Plug and Play Language Models: A Simple Approach to Controlled Text Generation. ICLR 2020

Sentiment Transfer Results





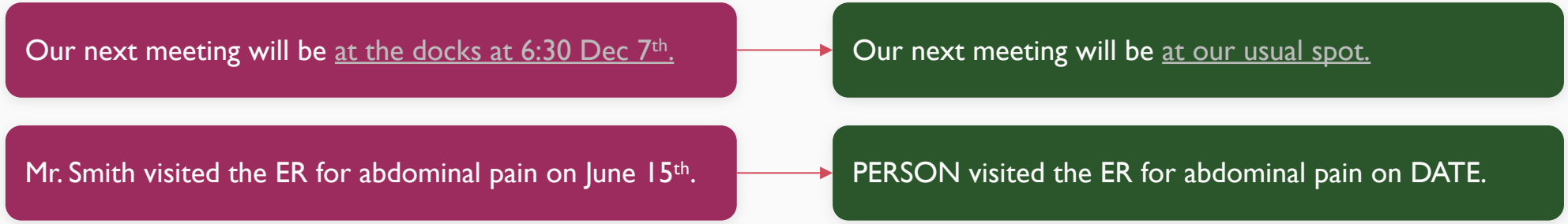
Talk outline

- **Problem 1: Leaky**
  - [NAACL'2021] Privacy Regularization: Joint Privacy-Utility Optimization in Language Models
- **Problem 2: Sneaky**
  - [EMNLP'2021] Style Pooling: Automatic Text Style Obfuscation for Improved Classification Fairness
- **Problem 3: Creepy**
  - [Submitted] Mix and Match: Learning-free Controllable Text Generation using Energy Language Models
- **Future work**
  - **Problem 3: Controlled Private and Safe Generation**
  - Problem 1: Measuring Memorization and Leakage in BERT-based Models

55

Controlled Private and Safe Generation

- Explore other applications of mix and match:
  - Privacy-aware generation and revision



  - Control the toxicity of generation and removing hate speech and toxic language.
  - Control bias in generated sentences.
    - Agency bias

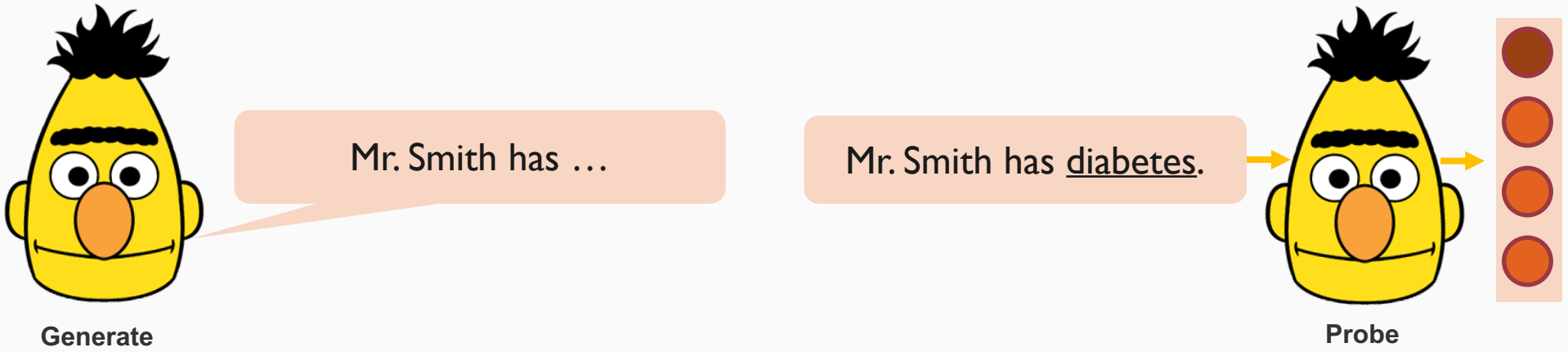
Talk outline

- **Problem 1: Leaky**
  - [NAACL'2021] Privacy Regularization: Joint Privacy-Utility Optimization in Language Models
- **Problem 2: Sneaky**
  - [EMNLP'2021] Style Pooling: Automatic Text Style Obfuscation for Improved Classification Fairness
- **Problem 3: Creepy**
  - [Submitted] Mix and Match: Learning-free Controllable Text Generation using Energy Language Models
- **Future work**
  - Problem 3: Controlled private and safe generation generation
  - **Problem 1: Measure memorization and leakage in BERT-based models**

57

Measuring Memorization and Leakage in BERT-based models

- Recent works (Lehman et al. 2021, Vakili et al. 2021) have shown that extracting training data from ClinicalBERT is not trivial.



- Question: Does this mean that BERT is memorizing less because of the MLM training? Or do we need different extraction methods?

Lehman et al. Does BERT Pretrained on Clinical Notes Reveal Sensitive Data? NAACL 2021.

Measuring Memorization and Leakage in BERT-based models

- We propose an attack where:
  - We use the Metropolis-Hastings sampler to sample from ClinicalBERT.
  - We use the energy of the samples to sift through generations and find memorized ones.
  - We propose the use of mix and match to incentivize generations that have sensitive information, such as diseases and patient names.
  - We propose new metrics for measuring memorization in generations.

Timeline

- Dec 2021: Thesis Proposal
- June 2022: Finish the BERT sample extraction project
- August 2022: Finish Controlled generation project
- Sept-Dec 2022: Prepare statements for job market and work on thesis
- June 2023: Thesis Defense

Thank you!

fatemeh@ucsd.edu