Recovering Private Text in Federated Learning of Language Models






Samyak Gupta* Yangsibo Huang* Zexuan Zhong Tianyu Gao Kai Li Danqi Chen

PRINCETON UNIVERSITY

Paper

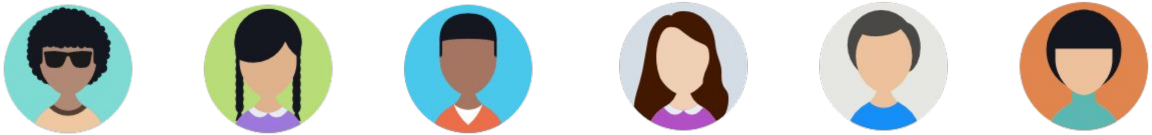
GitHub

Project Page: https://github.com/Princeton-SysML/FILM
Paper: https://neurips.cc/virtual/2022/poster/53215

Federated Learning

(McMahan et al. 16)

Distributed Clients



A Central Server

Federated Learning

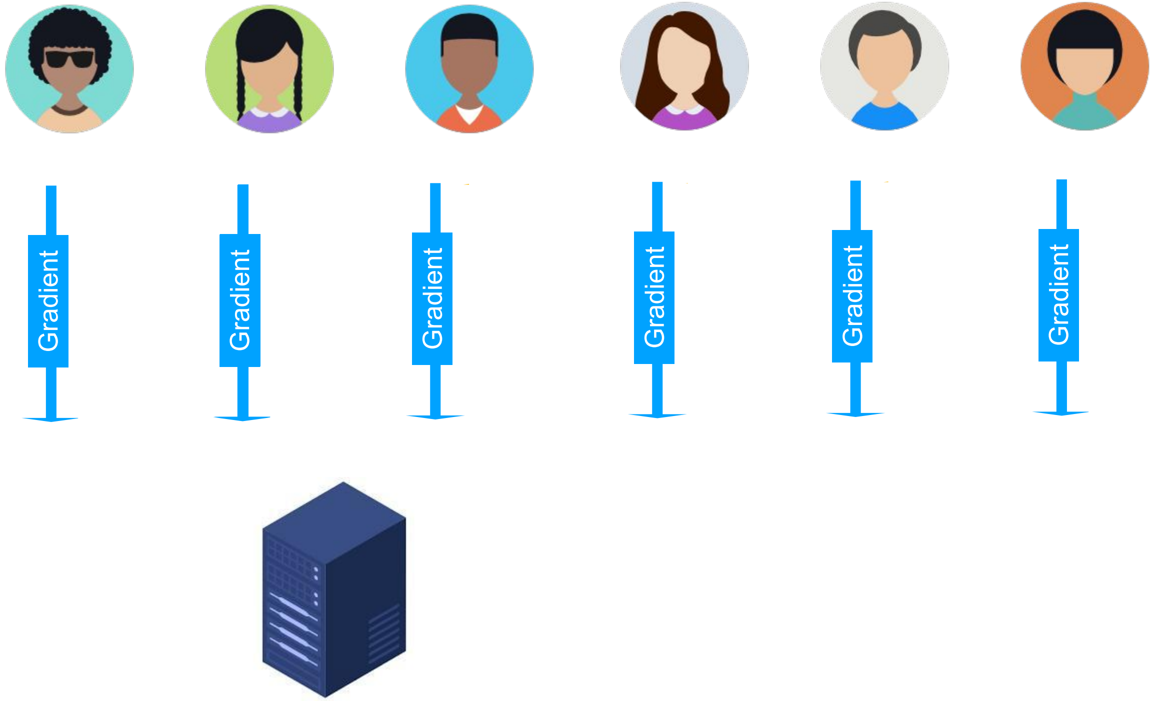
(McMahan et al. 16)

Every training step:

Distributed Clients

1. Users send model update (gradient)

A Central Server



Federated Learning

(McMahan et al. 16)

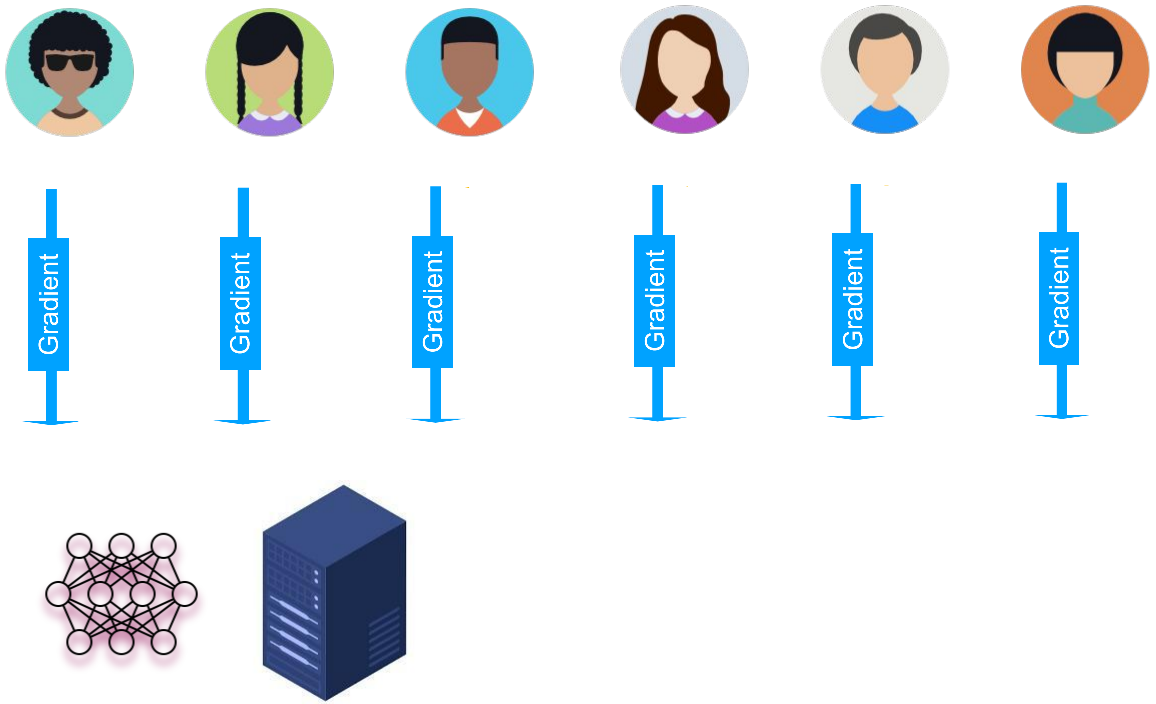
Every training step:

Distributed Clients

1. Users send model update (gradient)

2. Server applies aggregated update on the global model

A Central Server
(& a global model)



Federated Learning

(McMahan et al. 16)

Every training step:

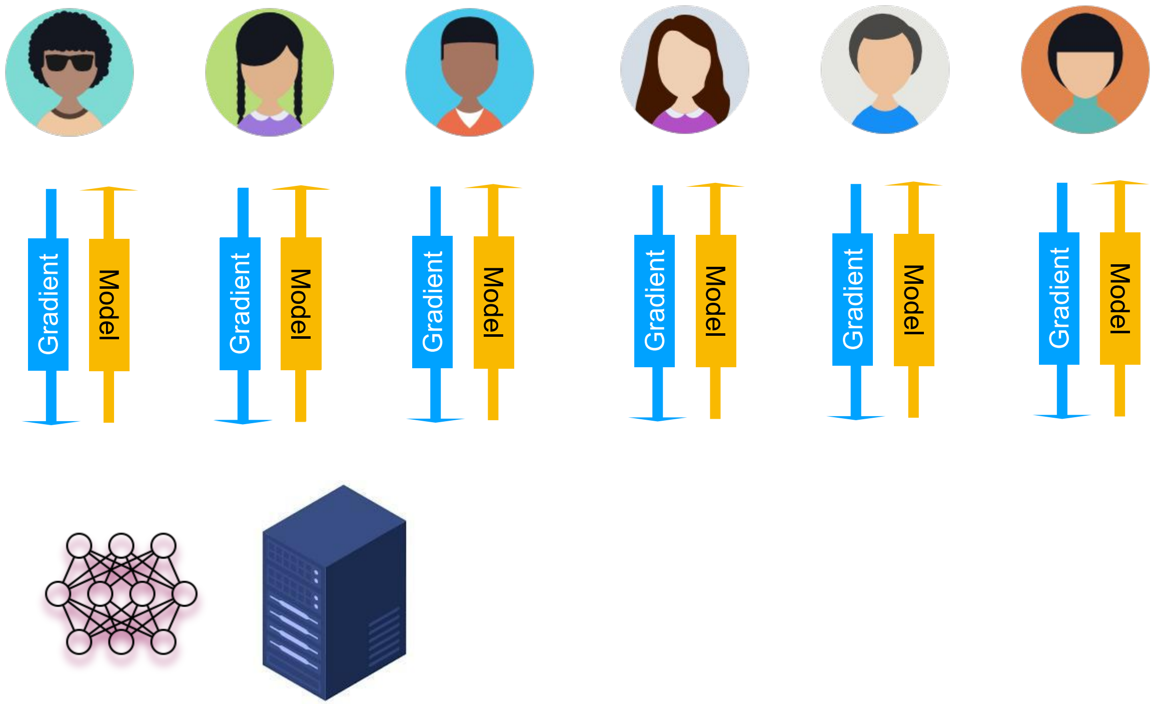
Distributed Clients

1. Users send model update (gradient)

2. Server applies aggregated update on the global model

3. Server broadcasts the new model

A Central Server
(& a global model)



ML-driven Services

Personalized Voice assistant

Personalized Keyboard

Personal Health records

2. Server applies aggregated

But how secure is federated learning?

Federated Learning (Threat Model)

(McMahan et al. 16)

Every training step:

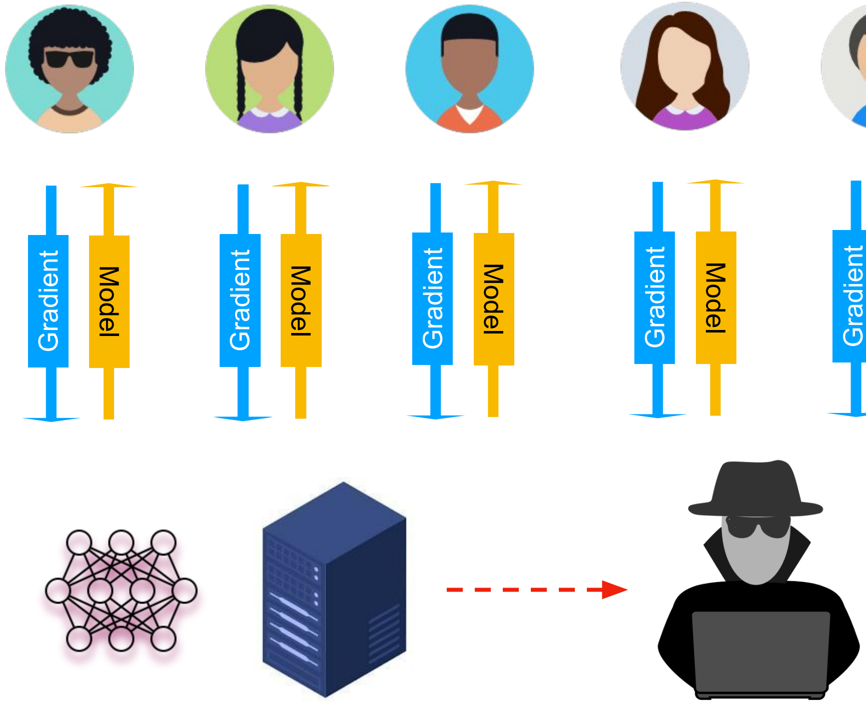
Distributed Clients

1. Users send model update (gradient)

2. Server applies aggregated update on the global model

3. Server broadcasts the new model

A Central Server
(& a global model)



Access to gradients and global model

Goal: Recover a single sentence from a batch

Recovering Image Data from Federated Learning

What about recovery of textual data?

(Yin et al. 21)

Key Challenge: Recovery from large batch sizes

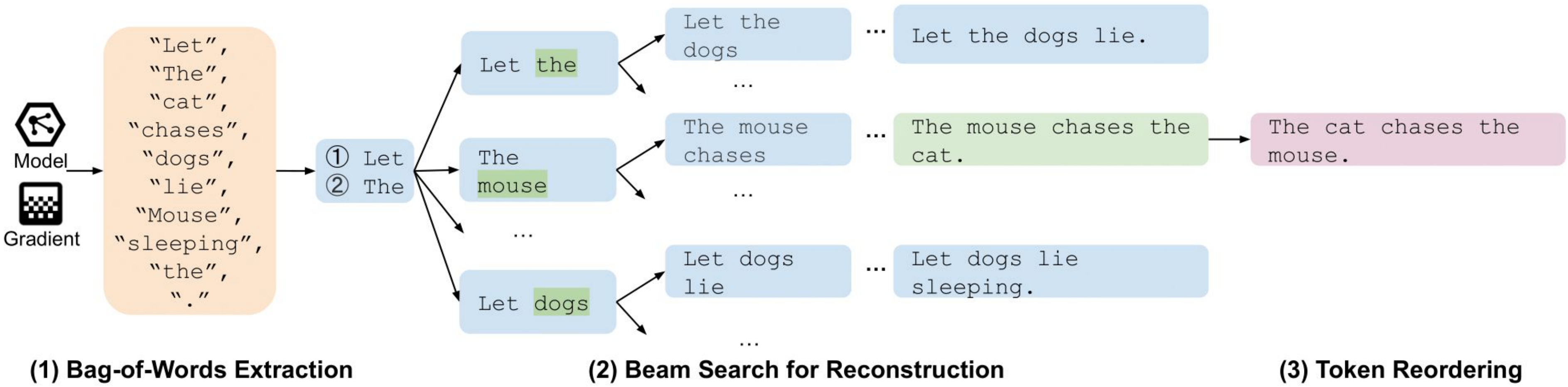
Attack	Can attack a batch of **?** CIFAR images	Can attack a batch of **?** ImageNet images
Zhu et al. 19	8	0
Geiping et al. 20	100	1
Yin et al. 21	100	32

This Paper

We study the privacy of federated learning for neural language models:

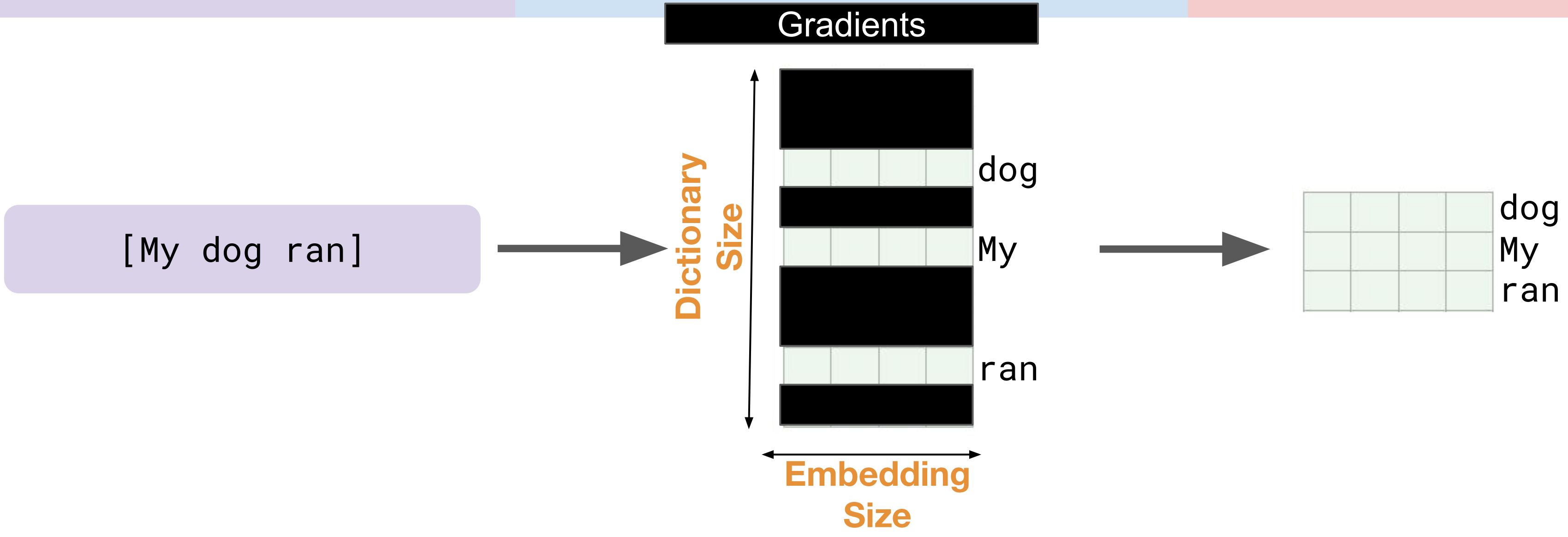
1. We present a novel attack called **FILM** (**F**ederated **I**nversion Attack for **L**anguage **M**odels) which recovers private data during federated learning
2. We demonstrate that FILM can recover private training data from gradients of large batches
3. We evaluate the utility-privacy tradeoffs of various defenses against our attack

Our Attack (FILM)



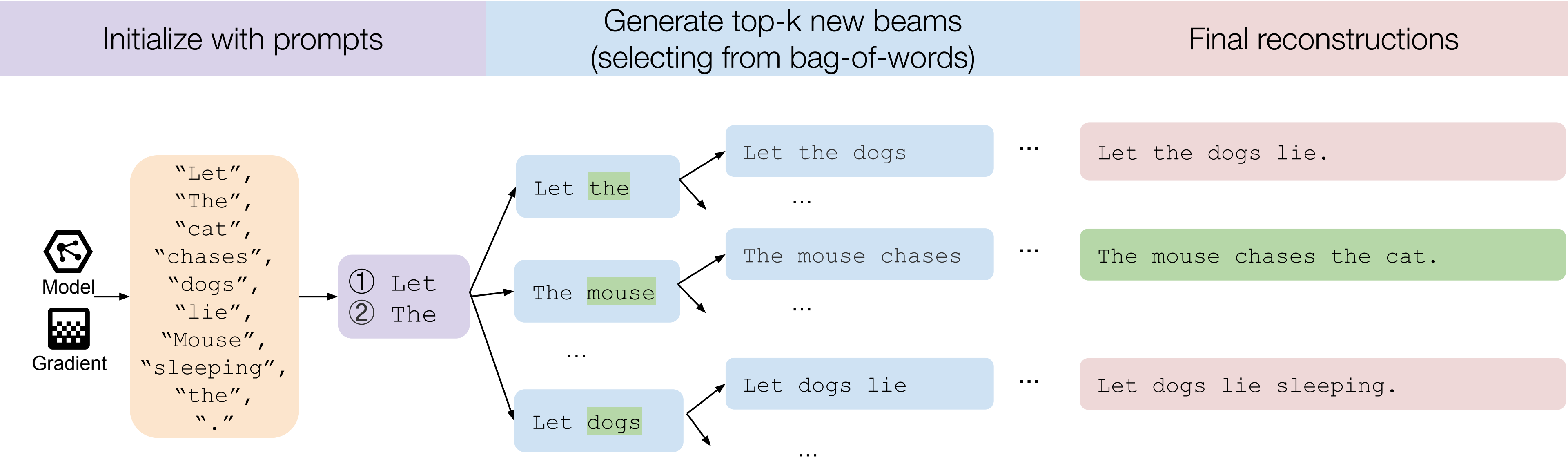
Bag-of-Words Extraction (Melis et al. 19)

Training Sentence	Word Embeddings Matrix	Bag-of-Words Extraction



Recover the set of used words by considering non-zero rows of embedding gradient

Beam Search



Beam scoring function: $$Score(B_i) = \sum_{w \in B} logp(w) - \rho R_n(B_i)$$

Probability of word w Repeated n-gram penalty

Prior-Guided Token Reordering

Example Sentence: The chases the mouse cat fieldGar. Let the dog lay

Strip Redundant Segments	Phrase-wise Reordering	Token-wise Reordering
The chases the mouse cat fieldGar. Let the dog lay	The chases the mouse cat fieldGar.	The cat fieldGar chases the mouse.
	The cat fieldGar chases the mouse.	The cat Garfield chases the mouse.

Token Reordering Score:

$$\mathcal{S}_\theta(\mathbf{x}) = \underbrace{\exp\left\{-\frac{1}{n}\log \mathbb{P}_\theta(\mathbf{x})\right\}}_{\text{Perplexity}} + \beta \underbrace{\|\nabla_\theta \mathcal{L}_\theta(\mathbf{x})\|}_{\text{Gradient Norm}}$$

Experimental Setting

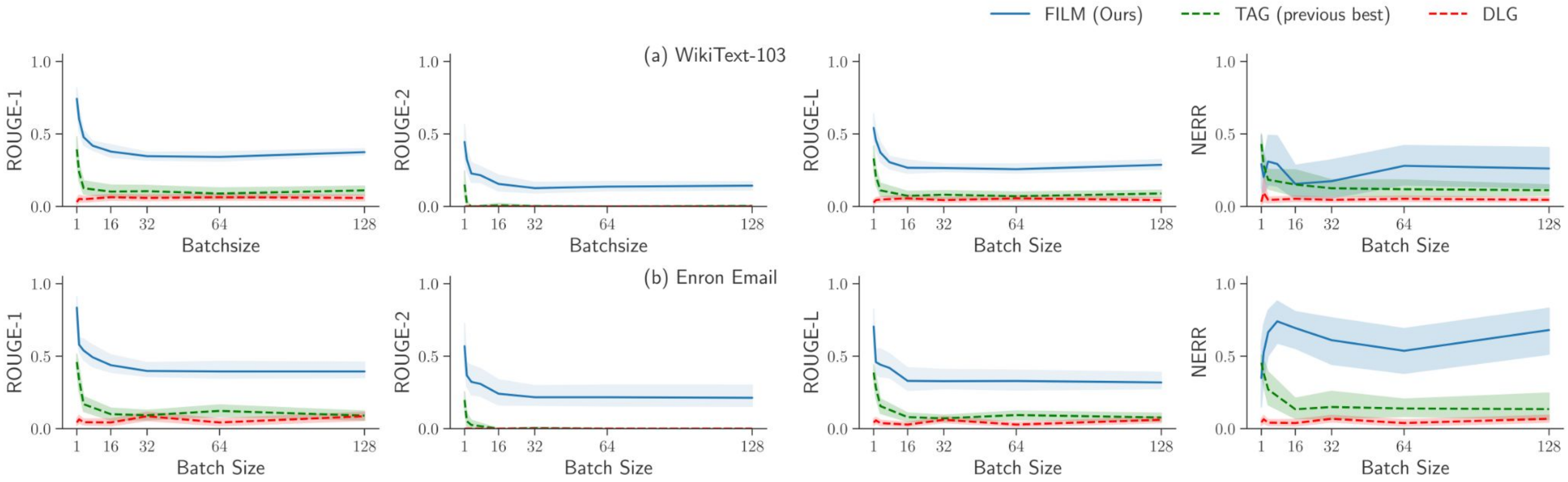
Metrics

ROUGE scores R-1/R-2/R-L, which measure the F-score of Unigrams/Bigrams and length of longest shared subsequence, respectively

Named Entity Recovery Ratio (NERR) which measures recovery of named entities which usually contain sensitive information (e.g., names, addresses, dates, or events

Dataset	Wikitext-103 (subset of 200k samples)
Dataset (real-world)	Enron emails (30k samples)
Model Architecture	GPT2

Quantitative Results (Batchsize)



FILM is stronger than prior approaches on all batch sizes

FILM recovers more named entities on a real-world dataset

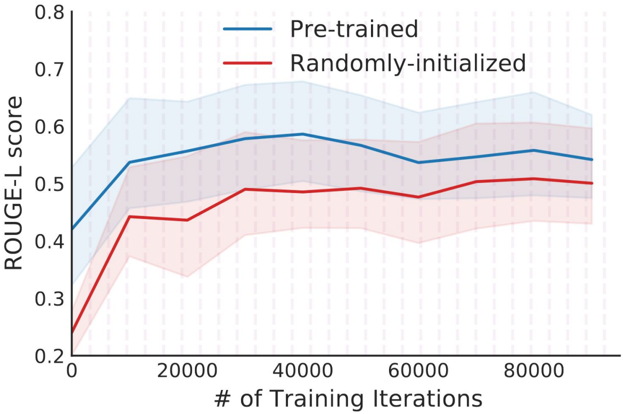
Qualitative Results (WikiText-103)

Attack & Batch Size b	**Original Sentence**	**Best Recovered Sentence**
	WikiText-103	
Zhu et al. (2019), $b = 1$	As Elizabeth Hoiem explains, "The most English of all Englishmen, then, is both king and slave, in many ways indistinguishable from Stephen Black.	thelesshovahrued theAsWords reporting the Youngerselagebalance, mathemat mathemat mathemat reper arrangpmwikiIndia Bowen perspectoulos subur, maximal
Deng et al. (2021), $b = 1$	Both teams recorded seven penalties, but Michigan recorded more penalty yards.	Both recorded teams seven but Michigan to recorded penalties 40 more penalty outstanding
FILM, $b = 1$	The short@-@tail stingray forages for food both during the day and at night.	The short@-@tail stingray forages for food both during the day and at night.
FILM, $b = 16$	A tropical wave organized into a distinct area of disturbed weather just south of the Mexican port of Manzanillo, Colima, on August 22 and gradually moved to the northwest.	Early on September 22, an area of disturbed weather organized into a tropical wave, which moved to the northwest of the area, and then moved into the north and south@-@to the northeast.
FILM, $b = 128$	A remastered version of the game will be released on PlayStation 4, Xbox One and PC alongside Call of Duty: Infinite Warfare on November 4, 2016.	At the time of writing, the game has been released on PlayStation 4, Xbox One, PlayStation 3, and PC, with the PC version being released in North America on November 18th, 2014.

Qualitative Results (Enron Emails)

Attack & Batch Size b	Original Sentence	Best Recovered Sentence
Enron Email		
Zhu et al. (2019), $b = 1$	Don and rogers have decided for cost management purposes to leave it consolidated at this point.	dj "... Free, expShopcriptynt)beccagressive Highlands andinos Andrea Rebell impacts
Deng et al. (2021), $b = 1$	We should not transfer any funds from tenaska iv to ena.	We should transfer any funds not from tenaska iv to iv en happens
FILM, $b = 1$	Volume mgmt is trying to clear up these issues.	Volume mgmt is trying to clear up these issues.
FILM, $b = 16$	Yesterday, enron ousted chief financial officer andrew fastow amid a securities and exchange commission inquiry into partnerships he ran that cost the largest energy trader $35 million.	Yesterday, enron ousted its chief financial officer, andrew fastow, amid a securities and exchange commission inquiry into partnerships he ran that cost the company $35 million in stock and other financial assets.
FILM, $b = 128$	Yesterday, enron ousted chief financial officer andrew fastow amid a securities and exchange commission inquiry into partnerships he ran that cost the largest energy trader $35 million.	Yesterday, enron ousted chief financial officer andrew fastow amid a securities and exchange commission inquiry into partnerships he ran that he said cost the company more than $1 billion in stock and other assets.

Results (Training Iterations)



FILM is stronger for later iterations of training

Original: The short@-@tail stingray forages for food both at night and during the day.	
t	**The recovered sentence**
0	The, short for short-tail stingray at night during night day during day at day night at
10,000	The stingray forages for food at night both at both during food food shortages at short food
20,000	The short@-@tail stingray forages for food during the day and for the night.
40,000	The short@-@tail stingray forages for food both at night and during the day.

Defenses (Against Bag-of-Words Recovery)

Gradient Pruning			
Prune ratio	Perplexity	Precision	Recall
0	11.46	1.00	1.00
0.9	11.57	1.00	1.00
0.99	12.77	1.00	1.00
0.999	15.34	1.00	0.98
0.9999	19.21	1.00	0.90
DPSGD			
ϵ of DPSGD	Perplexity	Precision	Recall
1	16.31	0.00	0.00
5	14.32	0.29	0.01
10	12.86	0.88	0.17
15	11.98	0.97	0.49
inf.	11.46	1.00	1.00

Gradient pruning only provides partial defense in exchange for large utility tradeoffs

DPSGD offers varying levels of protection, but perfect protection requires significant utility tradeoffs

Defenses (Frozen Embeddings)

	From Scratch		From Pretrained	
	Unfrozen	**Frozen**	**Unfrozen**	**Frozen**
Wikitext-103	27.31	118.69	11.40	11.48
Enron Email	15.16	69.17	7.09	7.30

Harsh utility costs when training from scratch

Minimal utility costs when fine-tuning

Paper

Princeton SysML

GitHub

Princeton SysML