✨Spotlight✨

Differentially Private Synthetic Data via Foundation Model APIs 2: Text

Chulin Xie[1], Zinan Lin[2], Arturs Backurs[2], Sivakanth Gopi[2], Da Yu[3], Huseyin Inan[2], Harsha Nori[2], Haotian Jiang[2], Huishuai Zhang[2], Yin Tat Lee[2], Bo Li[4], Sergey Yekhanin[2]

[1]University of Illinois Urbana-Champaign [2]Microsoft Research [3]Sun Yat-sen University [4]University of Chicago

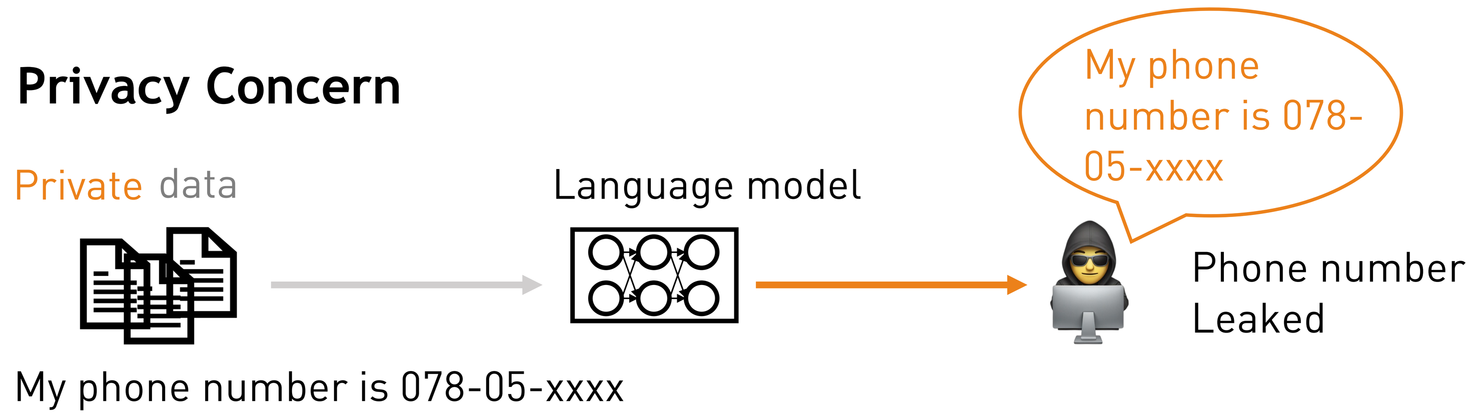
chulinx2@illinois.edu, {zinanlin,arturs.backurs,sivakanth.gopi,huseyin.inan,hanori,haotianjiang,huishuai.zhang,yintatlee,yekhanin}@microsoft.com, yuda3@mail2.sysu.edu.cn, bol@uchicago.edu

https://github.com/AI-secure/aug-pe

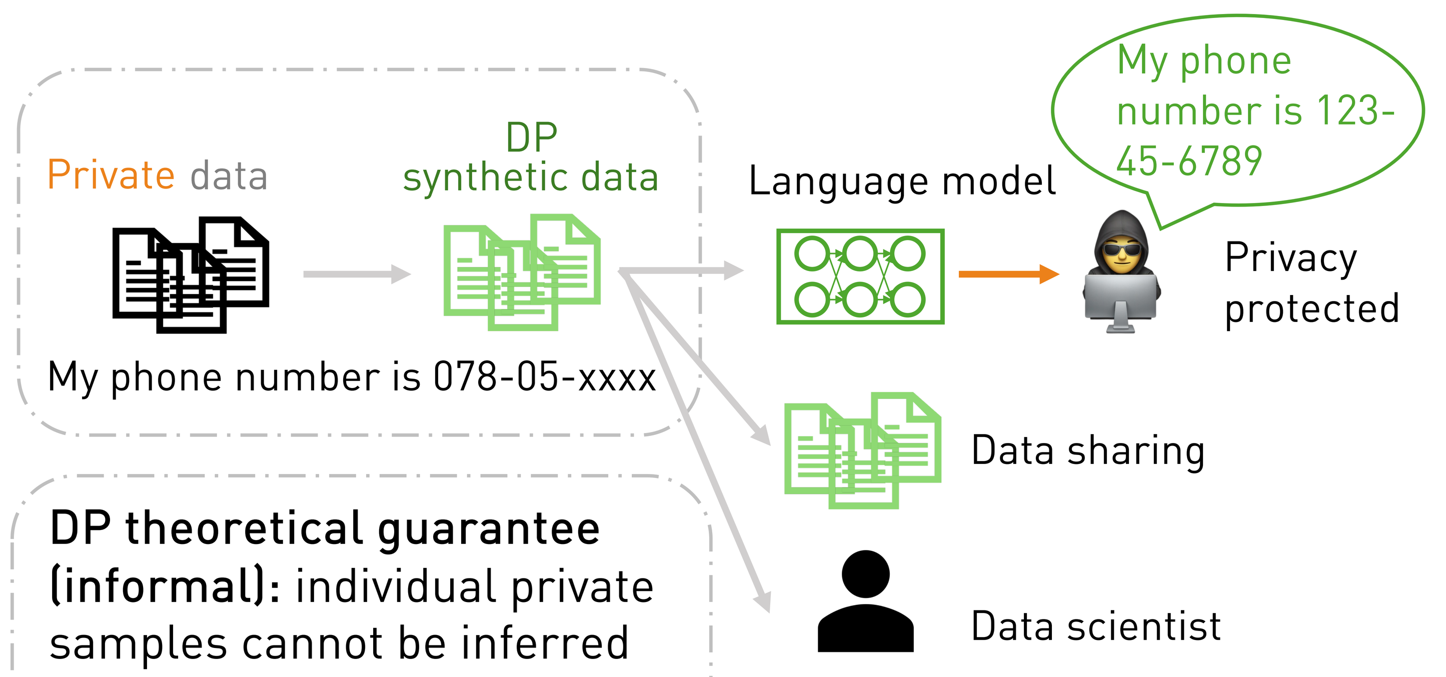
ICML International Conference On Machine Learning

Introduction

Privacy Concern

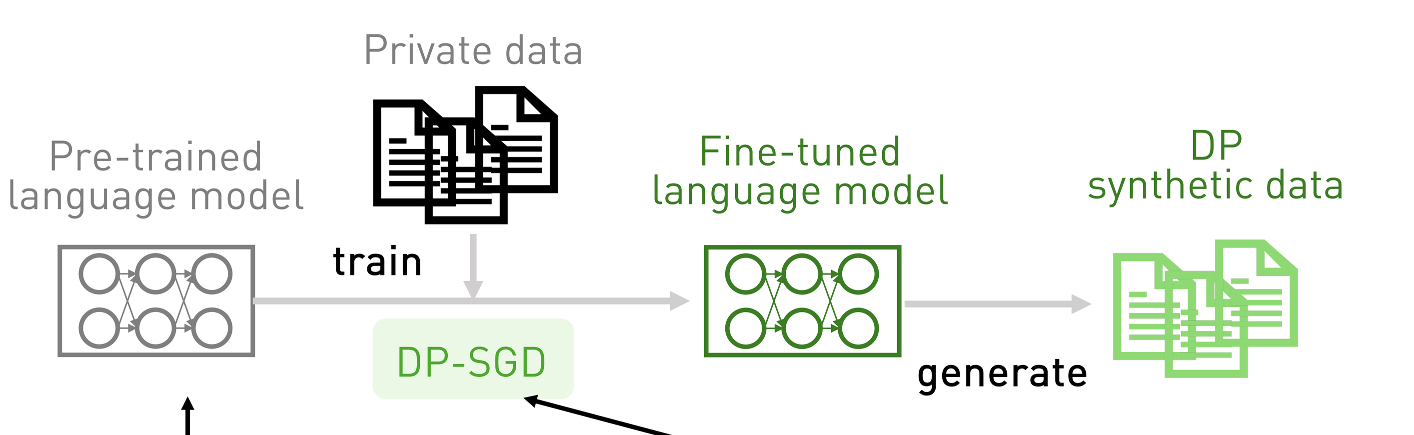


Differentially Private (DP) Synthetic Data



DP theoretical guarantee (informal): individual private samples cannot be inferred with high confidence from the DP synthetic data or post-processing of it

SOTA Method: DP Finetune Generator



Problem 1: Not able to use SOTA API-based foundation models, e.g., GPT-4, Claude

Problem 2: DP-SGD training is computationally intensive and hard to implement

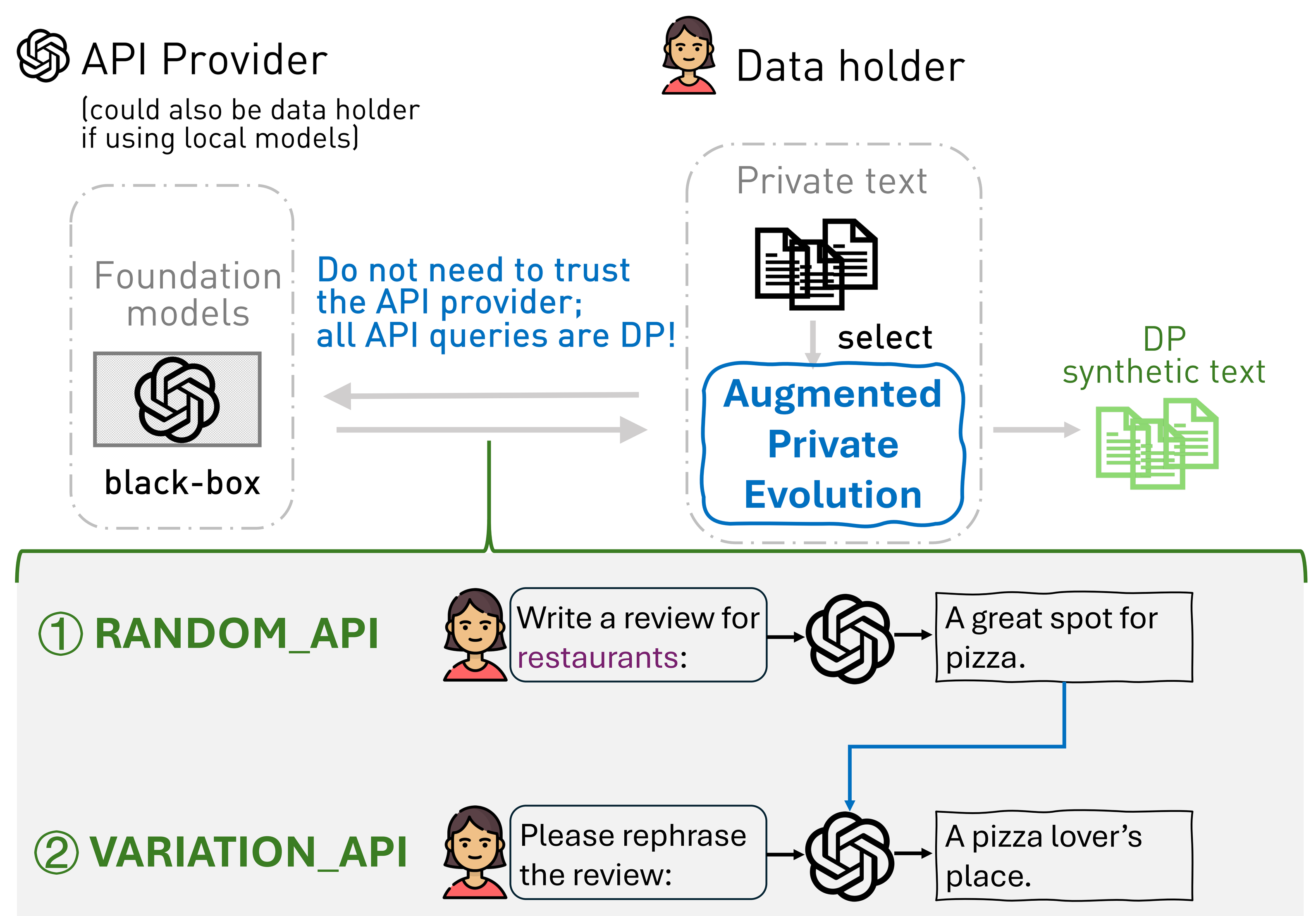
This Work: Augmented Private Evolution (Aug-PE)

- Only needs API access → Applicable to both API-based or open-sourced foundation models
- Does not need any model training
- Could even outperform DP finetune generator in terms of privacy-utility trade-off in some cases
- Extension of Private Evolution [ICLR 2024] from image to text, with new algorithmic techniques to increase the diversity and quality of text generation

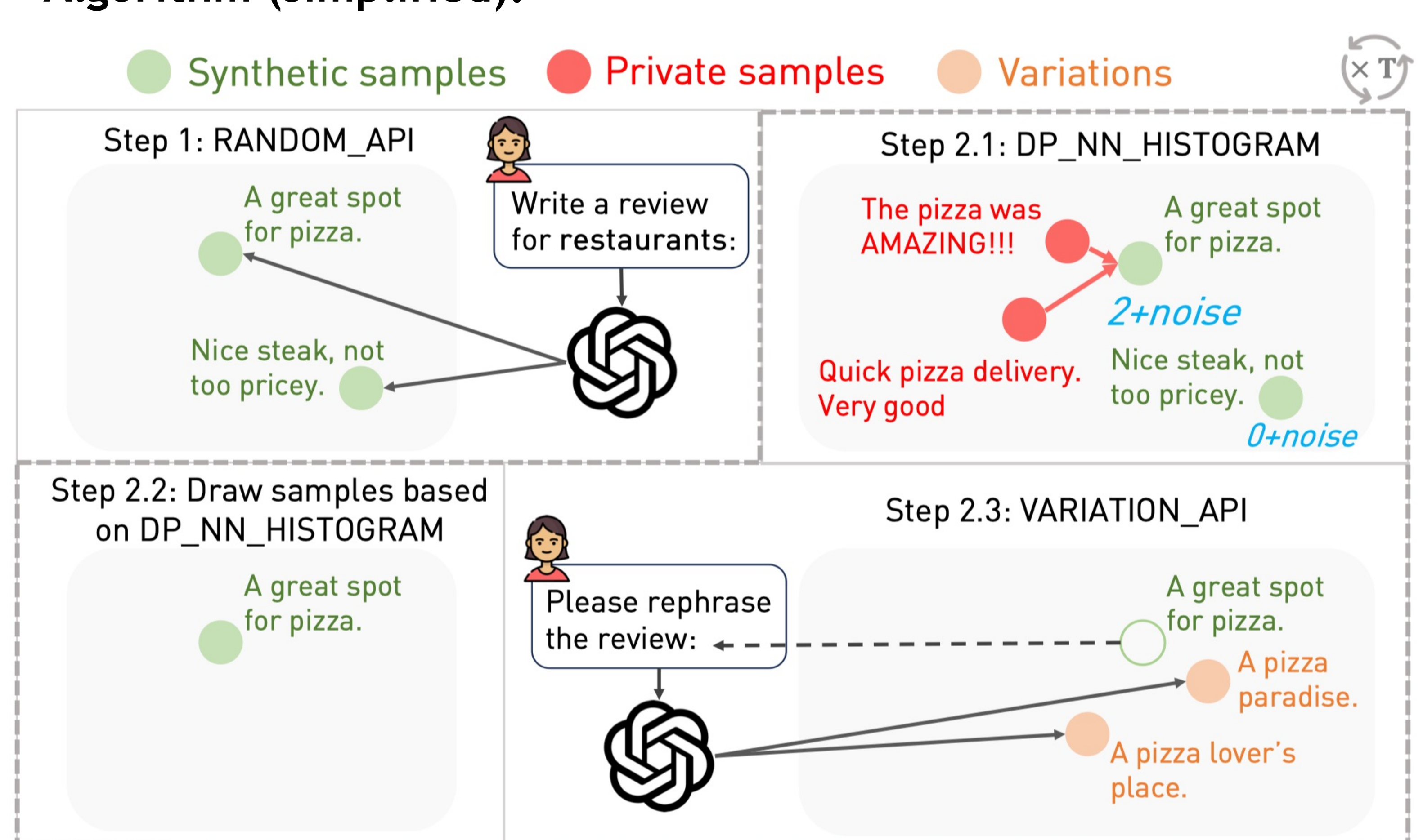
* [ICLR 2024] Differentially Private Synthetic Data via Foundation Model APIs 1: Image
Zinan Lin, Sivakanth Gopi, Janardhan Kulkarni, Harsha Nori, Sergey Yekhanin

Augmented Private Evolution

Workflow:



Algorithm (simplified):

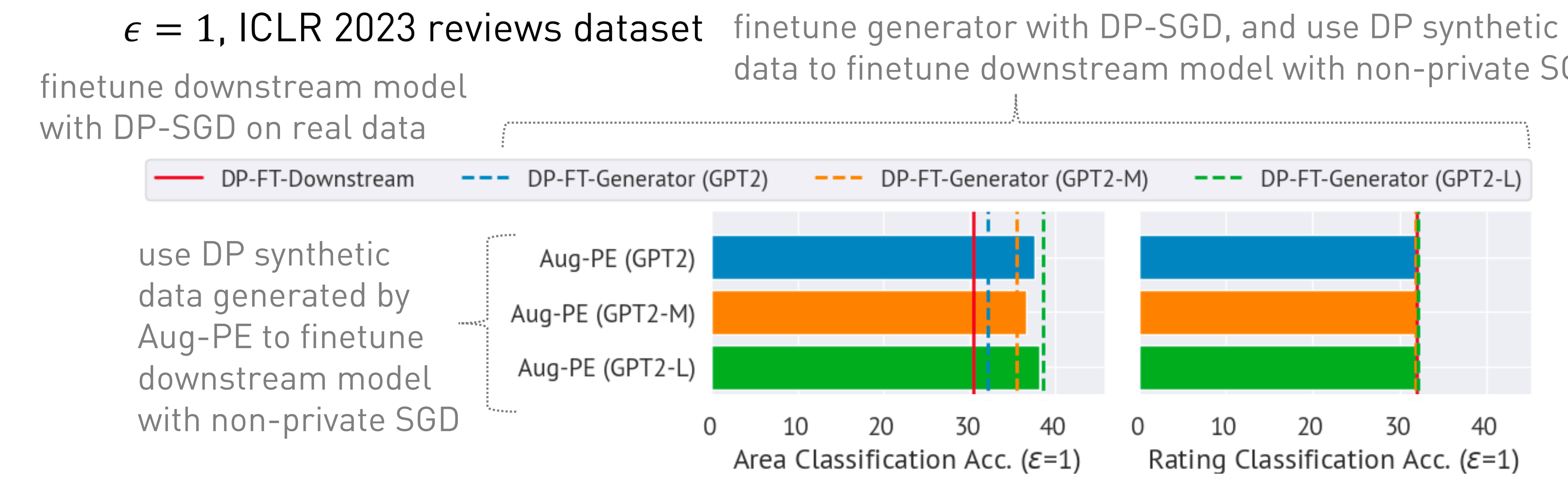


- Step 1 (RANDOM_API): prompt LLM to generate random samples.
- Step 2: go through steps 2.1-2.3 iteratively to refine the synthetic samples towards the private samples.
  - Step 2.1 (DP Nearest Neighbor Histogram): each private sample votes for their closest synthetic sample in the embedding space induced by embedding model. Then, add Gaussian noise to the votes to ensure DP.
  - Step 2.2: resample the generated texts according to the histogram.
  - Step 2.3 (VARIATION_API): prompt LLM to generate new similar samples, which will be used in the initial synthetic samples for the next iteration.

Experiments

Aug-PE matches/beats SOTA on text quality vs. privacy

$\epsilon = 1$, ICLR 2023 reviews dataset



Aug-PE is compatible with advanced LLMs for DP synthetic text generation (challenging/infeasible for DP finetuning)

	OpenReview				PubMed			
	$\epsilon=\infty$		$\epsilon=1$		$\epsilon=\infty$		$\epsilon=1$	
LLM	Area	Rating	Area	Rating	$\text{BERT}_{\text{Mini}}$	$\text{BERT}_{\text{Small}}$	$\text{BERT}_{\text{Mini}}$	$\text{BERT}_{\text{Small}}$
GPT-2	42.4	32.1	37.6	32.0	24.5	26.7	24.3	26.5
GPT-2-Medium	41.0	32.3	36.6	32.1	25.5	27.7	24.9	27.0
GPT-2-Large	42.1	32.1	38.1	32.0	25.7	27.9	25.1	27.2
Opt-6.7b	43.6	32.2	30.5	32.1	26.5	28.6	25.8	27.9
Vicuna-7b-v1.5	42.9	35.7	35.2	35.4	24.6	26.9	23.1	24.9
Falcon-7b-instruct	38.6	32.6	39.0	33.3	22.3	24.4	22.4	24.5
Llama-2-7b-chat-hf	45.5	38.5	36.4	37.0	25.8	28.4	24.8	27.5
Mixtral-8x7B-v0.1	**45.9**	41.8	**43.6**	42.3	24.9	27.6	24.5	27.1
GPT-3.5	45.4	**43.5**	41.9	**43.1**	**30.4**	**32.7**	**30.1**	**32.4**

Hard to DP fine-tune due to computation requirement of DP-SGD

infeasible for DP finetuning as weights/ architectures are unavailable

Aug-PE uses private data to guide synthetic data selection

use GPT-3.5 as data generator

Setting	Yelp		OpenReview		PubMed	
	Rating	Category	Area	Rating	$\text{BERT}_{\text{Mini}}$	$\text{BERT}_{\text{Small}}$
Random API	62.3	73.7	34.4	42.0	29.7	31.9
Random API + Variation API	62.3	73.7	36.4	42.0	29.6	31.9
AUG-PE ($t=1$)	64.4	74.1	39.3	42.5	30.0	32.2
AUG-PE ($t=T$)	**67.9**	**74.7**	**45.4**	**43.5**	**30.4**	**32.7**

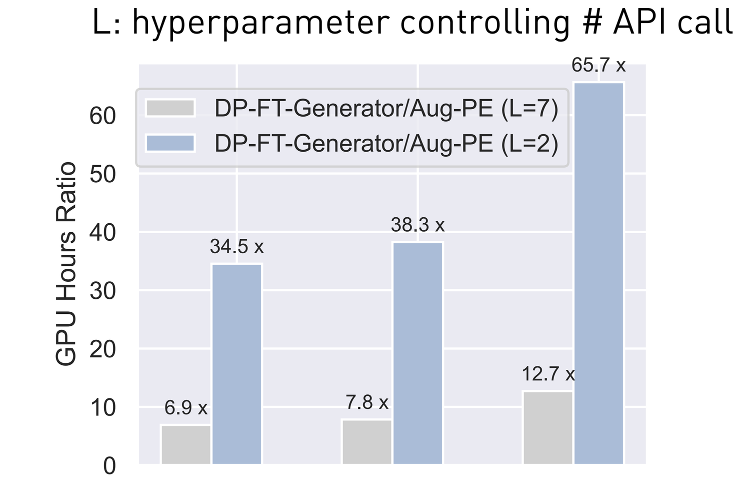
Aug-PE outperforms PE for text generation

- apply the same API designs and models to PE [ICLR 2024] to support text generation
- use GPT-2 as data generator
- the results show that new algorithmic techniques introduced in Aug-PE are effective

Method	Yelp		OpenReview		PubMed	
	Rating	Category	Area	Rating	$\text{BERT}_{\text{Mini}}$	$\text{BERT}_{\text{Small}}$
PE ← AUG-PE ($k=6, L=1$)	44.9	71.8	35.3	32.0	20.1	22.3
AUG-PE ($k=0, L=7$)	**67.5**	**74.8**	**42.4**	**32.1**	**24.5**	**26.7**

Aug-PE can be computationally cheaper

L: hyperparameter controlling # API calls



Aug-PE can capture text length distribution