A OUR PRIVACY GUARANTEES

To provide all the information needed to understand our privacy guarantees, we follow the guidelines outlined in (Ponomareva et al., 2023).

1. **DP setting**. We provide central DP guarantee where the service provider is trusted to correctly implement the mechanism.
2. **Instantiating the DP Definition**
   (a) *Data accesses covered*: Our DP guarantees apply only to a single training run. We don't account for hyperparameter tuning in our guarantees.
   (b) *Final mechanism output*: We use DP-Training methods. Only model's predictions (e.g. synthetic data generated by the DP-Trained model) is released, however the mechanism's output is technically the full sequence of privatized gradients, and the guarantee also applies at this level. This also means that all the checkpoints are protected/can be released publicly.
   (c) *Unit of privacy*. We consider example-level DP, where each example is a long sequence of text, e.g. a full yelp review. Maximum length of such unit in tokens is 512, tokens are extracted by SentencePiece algorithm trained on c4. We consider full data protection (full text of a review).
   (d) **Adjacency definition for "neigbouring" datasets"**: We use add-or-remove adjacency definition.
3. **Privacy accounting details**
   (a) *Type of accounting used*: RDP-based accounting.
   (b) *Accounting assumptions* : Poisson sampling was assumed for privacy amplification but shuffling was used in training)
   (c) *The formal DP statement*: We use various levels of ϵ values: 1,3,10. Our $\delta = \frac{1}{training\_data\_size}$
   (d) *Transparency and verifiability*: We are going to open source our code. We use open-sourced t5x framework.

B ADDITIONAL RELATED WORK

While in our paper we concentrate on generating text synthetic data, the task of generating synthetic tabular data has been explored extensively before.

Synthetic Tabular data. For tabular data, early works on synthetic data concentrated on estimating the utility of the synthetic data as *a quality of the statistical answers* over the data (synthetic data for query release). Privacy-preserving aspect for such synthetic data was often achieved via DP methods (Blum et al., 2011; Hardt et al., 2010; Liu et al., 2021) More recently, works that instead evaluated how useful the tabular synthetic data is for some downstream ML model started gaining popularity (Tao et al., 2021). In general, the approaches for generating synthetic tabular data can be categorized into *Marginal-based* and *generative models-based*. Marginal-based models calculate private marginal distribution by taking marginal distribution over attributes and appropriately privatizing it with some DP-mechanism like Gaussian or Laplace. Then attributes for synthetic instances are sampled from these distributions. *Generative models* instead attempt to build a function that approximates the data distribution and sample from this function. GAN-based methods were a common choice for such generative models: e.g., DP-GAN (Xie et al., 2018), Convolutional GAN (Torfi et al., 2020) and PATE-GAN (Yoon et al., 2019). A recent benchmark (Tao et al., 2021) reported that for achieving best downstream ML model performance, marginal-based methods still outperform GAN-based approaches.

C LANGUAGE MODEL PRE-TRAINING.

Model architecture. For syntethic data generation, one can use either decoder only or encoder-decoder model. We used a decoder-only transformer model (Vaswani et al., 2017), which should be

more parameter-efficient, with architecture similar to LaMDA (Thoppilan et al., 2022). The quality of the pretrained model is of paramount importance for generating good fidelity synthetic data. We initially experiemented with 1B model and found that a signficant boost in downstream performance can be achieved by using higher capacity model.

Therefore we use 8B model for final fine-tuning and prompt tuning. A smaller version (1B) model is used to tune hyperparameters like learning rate, clipping norm, batch etc. The hyperparameter values found using 1B model are used for the final 8B model. Statistics for 1B and 8B models are provided in the Table 5.

Table 5: Architecture of transformer models used in experiments.

Parameters	Layers	Attn. heads	d_{model}	d_{ff}	d_k, d_v
1B	8	32	2048	16384	128
8B	16	64	4096	32768	128

Pre-training data. We used public dataset The Pile (Gao et al., 2020) as a basis for our pre-training data. However we run extra post-processing step by deduplicating The Pile against all datasets used in downsteam tasks. This deduplication step is necessary to ensure that private data won't be accidentally learned during model pre-training, which otherwise would invalidate our DP guarantees.

To do deduplication we followed recipe outlined in (Lee et al., 2022). Specifically, we tokenized and constructed suffix array for each involved dataset (The Pile, IMDB, Yelp, AGNews). Then we used suffix arrays to find common sequences of 50 of more tokens which appear in The Pile and any other dataset. Finally we cut all those common sequences from The Pile dataset. After cutting the sequences we de-tokenized dataset back to strings and used it for pre-training. It's important to note that we only deduplicate The Pile against other datasets, we did not run deduplication of The Pile against itself.

Pre-training procedure. We pre-trained our models using T5X codebase (Roberts et al., 2022), however we adopted few tricks which were used to train open sourced GPT-NeoX model (Black et al., 2022). Details are below.

We pre-trained a model for 380k steps using batch size of 1024 example with 1024 tokens sequence length, which result in training for approximately 400B tokens. We used same SentencePiece tokenizer which was used in original T5 model (Raffel et al., 2020). Cross entropy loss on next token prediction (teacher-forcing) was used as a training objective. Additionally we employed weight decay of 0.001 and an auxiliary z-loss $10^{-4}\log^2(Z)$ where $\log(Z)$ is softmax normalizer. Training was done with AdaFactor optimizer with learning rate schedule $\min(0.01, \frac{1}{\sqrt{N}})$ where N is a step counter.

D Analysis of pre-training data contamination of GPT-2 training data.

Multiple prior works on private synthetic data generation (Yue et al., 2022; Putta et al., 2023; Mattern et al., 2022) start with a pre-trained GPT-2 model and then finetune it with differential privacy on some downstream dataset. In this section we demonstrate that pre-training data for GPT-2 model includes examples from downstream tasks which invalidates the privacy guarantees of such prior work.

Let's assume we pre-train model M on some public dataset D_p and finetune with DP-SGD on sensitive dataset D_s. If $D_p \cap D_s = \emptyset$ then we can conclude that the process of obtaining our model M is differentially private w.r.t. D_s. However this is no longer the case if intersection of D_p and D_s is non-empty.

GPT-2 was pre-trained on a WebText dataset (Radford et al., 2019). While the dataset is not released publicly, authors discuss in detail the process of how dataset was constructed by scraping the content of links posted on Reddit website. Additionally, they released a list of top domains [2] from which the dataset was formed. Specifically, it could be seen that 183080 web-pages from IMDB and 36188

[2] https://github.com/openai/gpt-2/blob/master/domains.txt

web-pages from Yelp websites were included GPT-2 pre-training data. This by itself, indicates that parts of IMDB and Yelp dataset were include in GPT-2 pre-training data.

We performed further analysis in the following way. We took a subset of OpenWebText dataset (Peterson et al., 2019) - a public re-implementation of WebText. Specifically we used `c4/webtextlike` from TFDS[3]. We computed an intersection of IMDB dataset and `c4/webtextlike` using approach from (Lee et al., 2022). We found that `c4/webtextlike` contains 136 distinct examples from IMDB training and test set.

Finally we manually looked at a few of the examples to verify that they are indeed part of IMDB dataset and that they were likely used to construct WebText dataset. Here is a code snippet to obtain one of these examples:

```
import tensorflow_datasets as tfds
ds = tfds.load('imdb_reviews/plain_text:1.0.0',
               split='test',
               shuffle_files=False)
# get 52nd example from IMDB test set
print(next(iter(ds.skip(51)))['text'].numpy().decode('utf-8'))
# Output:
# In the eighties, Savage Steve Holland put out three movies ...

# This example is a second review from here:
# https://www.imdb.com/title/tt0091680/reviews
# The link to the IMDB web-page is mentioned
# in the following reddit post:
# https://www.reddit.com/r/movies/comments/77gna7/comment/dom5mqa
# The Reddit post was written in 2017, two years prior to
# creation of WebText dataset,
# suggesting that it would be included in the dataset.
```

This indicated that indeed at least some portion of the IMDB reviews would have been obtained during WebText dataset construction process and would have been used for GPT-2 pre-training. While it is impossible to say, without an access to the actual pretraining data, what percentage of the downstream tasks data was included for GPT-2, it highlight the importance of our deduplication process for obtaining rigorous privacy guarantees.

E DETAILS OF PROMPT-TUNING.

Training instability with Adafactor optimizer and default prompt initialization. (Lester et al., 2021) recommends Adafactor as a default optimizer. They also suggest to initialize prompt using pre-trained embeddings of tokens from vocabulary. While we confirm that it works well for non-private prompt tuning of LLM, it appeared to be inadequate for DP-prompt tuning.

With this setup we observed significant training instabilities even in the best DP-runs after significant amount of clipping norm and learning rate tuning, see figure 2.

Changing optimizer and prompt tuning initialization.

To achieve good performance with prompt tuning we experimented with two things.. First of all, we tried various optimizers (including Adam and Momentum) instead of Adafactor. Additionally we experimented with random initialization of prompt tensor.

While changing optimizer didn't really improve the downstream performance, we found that Adam and Momentum optimizers lead to more stable training runs. Changing prompt tensor initalization to random uniform with small range was the key for improving prompt tuning performance with DP. Table 6 demonstrates that random initialization results in a lift of up to 30% in downstream CNN performance, and up to 10% for BERT model. We hypothesize that proper initialization is very important to DP-Training and addition of noise makes it hard for the model to "recover" from bad initialization.

[3] https://www.tensorflow.org/datasets/catalog/c4#c4webtextlike

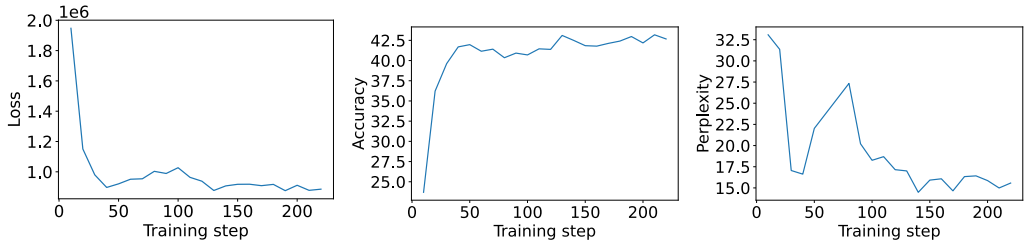

Figure 2: Training loss, training accuracy and validation perplexity for the best prompt tuning training run with Adafactor optimizer, $\epsilon = 1$.

Table 6: Downstream performance of prompt tuning with different optimizers and prompt initialization. In all experiments prompt tuning was done for 10 epochs with 1024 batch size and privacy level $\epsilon = 1$. Column "Vocab Init" correspond to default initialization proposed in prompt tuning paper using embeddings of tokens from vocabulary. Column "Random Init" correspond to random uniform initialization in range $[-0.01, 0.01]$.

Optimizer	BERT		CNN	
	Vocab Init	Random Init	Vocab Init	Random Init
Adafactor	72.1 ± 2.1	88.2 ± 0.2	55.8 ± 0.2	84.9 ± 0.2
Adam	74.6 ± 1.2	85.5 ± 0.7	50.2 ± 0.1	82.4 ± 0.4

We did some limited experiments with the scale of random uniform initialization, however for most of them we didn't run full synthetic data pipeline and only compared LLM perplexity and next token prediction accuracy. These experiments showed that random uniform from range $[-0.01, 0.01]$ was the best, while both decreasing and increasing the range led to worse metrics.

We also compared Adam optimizer with fixed learning rate, Adam optimizer with cosine learning rate decay and Momentum optimizer with cosine learning rate decay. As long as learning rate was tuned, all three performed similarly. Eventually we settled on Adam optimizer with fixed learning rate.

F Details of LoRa

Let's say one of the layers in the network is represented as dense matrix multiplication operation Wh where W is a trainable weight and h is an input to the layer (typically embedding or hidden state vector). LoRa (Hu et al., 2021) proposes to replace weight matrix with a sum $W + LR$, freeze W and tune L and R matrices. In this notation L and R are low-rank matrices such that the result of their multiplication has the same shape as W. For example if W is a matrix with the shape $n \times m$, then L would be $n \times r$ matrix and R would be $r \times m$ matrix, where r is the rank of low-rank adapter.

Similar to (Hu et al., 2021), all layers where LoRa is not applied are considered frozen in our setup. We performed LoRa tuning using Adam optimizer with fixed learning rate. Unlike (Hu et al., 2021) we did not use weight decay because we observed no benefits of weight decay in differentially private LoRa training.

Typically, in a transformer model LoRa can be applied to attention layers and MLP layers. Hu et al. (2021) study LoRa only on attention layers. In this work, we tried to applied LoRa to attention layers only (Attention-LoRa), MLP layers only (MLP-LoRa) and both attention and MLP layers (Full-LoRa). Overall we found out that Full-LoRa is the best choice in most cases, however in some cases MLP-LoRa could be slightly better.

We also studied how rank of LoRa affects downstream performance. Similar to (Hu et al., 2021) we observed that initially increase of LoRa rank lead to increase of accuracy on downstream task, followed by eventual decrease of downstream accuracy with further increase of rank.

Our main results in table 1 are reported using Full-LoRa rank 8 on AGNews, rank 1 for Yelp datasets, and MLP-LoRa rank 32 on IMDB dataset.

Overall we can recommend to use Full-LoRa rank 8 as a good default value, however we can encourage tuning of LoRa parameters when resources allow it and the best possible performance is needed.

See also detailed ablation of LoRa parameters in Appendix J.5.

G DETAILS OF SYNTHETIC DATA SAMPLING.

For data generation/sampling part, LLMs have the following knobs:

1. *temperature*: this is a constant by which the logits are divided prior to softmax and subsequent sampling. Large temperature flatten the tokens distributions, making rarer tokens more likely to be selected; it also increases diversity of the data generated but can negatively affect the quality. Our experiments show that default temperature of 1 (so no modification of the tokens distribution) works the best.
2. *topk*: given a token distribution, topk determines what portion of the distribution to keep before sampling. Topk is similar to low temperature - e.g. setting topk to top 1000 tokens means that next token will be chosen from the most likely tokens. We keep this parameter unmodified (set to ∞).
3. *numdecodes*: is similar to the number of candidates in standard beam search. We use the value of 1 here, resulting in the sequence where each token is the most likely to be returned.

While we did experiment with various values of these hyperparameters (see appendix J.3), we find that default settings already provide enough of diversity of generated examples for our datasets. If further diversity needs to be enforced (e.g. for very large datasets), we recommend increasing the temperature or numdecode values.

H DETAILS OF DATASETS FOR DOWNSTREAM TASKS

We conducted experiments on IMDB reviews (Maas et al., 2011), Yelp reviews (Zhang et al., 2015a) and AGNews[4] datasets. On IMDB and Yelp we formulated downstream task as binary sentiment classification. On AGNews the downstream task was classification of titles of news articles into one of 4 topics (World, Sports, Business and Sci/Tech).

All of the considered datasets only provided training and test sets. To obtain validation set we split original training set into two chunks in deterministic way using TFDS split slicing API. We used first 90% of the original training set for training, and last 10% for validation.

Some of the dataset statistics for considered dataset and our train/test/validation splits is provided in table 7.

I ARCHITECTURE AND TRAINING OF DOWNSTREAM MODELS.

We used two types of models for all downstream experiments. First one is a BERT encoder with a dense layer on top of it, second one is a shallow CNN.

BERT model. For BERT-based classifier we used BERT-Base model (Devlin et al., 2018b) pretrained on English data[5] with standard BERT tokenization and preprocessing. We put a dense layer on top of pooled output of the BERT encoder to produce classification score. Output of dense layer was a single floating point number per input sequence which was converted to probability using sigmoid function. We trained entire model (including BERT encoder and dense layer) without freezing any layers.

CNN model. In addition to BERT, we used a shallow CNN model without any extra pre-training. Our model architecture followed the idea from (Johnson & Zhang, 2015) with the main difference that we didn't pre-train embeddings beforehand. To be more specific, our model works as follows. We convert input sequence to lowercase and tokenize it by splitting it on whitespaces and punctuation

[4] https://www.tensorflow.org/datasets/catalog/ag_news_subset

[5] https://tfhub.dev/tensorflow/bert_en_uncased_L-12_H-768_A-12/4

Table 7: Datasets statistics.

	IMDB	Yelp	AGNews
Training examples			
Total	22500	504000	108000
Class 0	11256	252085	26915
Class 1	11244	251915	26985
Class 2	-	-	27081
Class 3	-	-	27019
Validation examples			
Total	2500	56000	12000
Class 0	1244	27915	3085
Class 1	1256	28085	3015
Class 2	-	-	2919
Class 3	-	-	2981
Test examples			
Total	25000	38000	7600
Class 0	12500	19000	1900
Class 1	12500	19000	1900
Class 2	-	-	1900
Class 3	-	-	1900

signs. Then we embed it into 384 dimensional embedding, using vocabulary from most common 30k words from the dataset. Embedding was randomly initialized and trained as a part of downstream training task. Output of embedding layer was passed through 1D convolution with kernel size 3, 256 output filters and RELU activation. This followed by global max pooling along the sequence length. Then we have a fully connected layer with 256 outputs and RELU activation, followed by final logits layer with sigmoid activation.

Training of downstream model. Both BERT and CNN were trained in a similar way. For non-private training we used Adam optimizer and DP-Adam (as implemented in the Keras DPModel library) for private training. Model was regularized with weight decay and we did early stopping (by non-increasing validation accuracy). We choose best hyperparameters by running a sweep over learning rates $\{10^{-7}, \dots, 10^{-1}\}$ and weight decays $\{5 \times 10^{-1}, \dots, 5 \times 10^{-6}, 0\}$. Additionally for DP-training baseline on real data we did a sweep of clipping norm over $\{10^{-2}, \dots, 10^{2}\}$.

J ABLATIONS

We study the influence of various factors on quality of generated synthetic data. We observed that better pre-trained model (in terms of model size and choice of pre-training dataset) leads to better synthetic data in differentially-private case, see details in Appendix J.1.

All standard techniques which are used to improve utility of DP-training (Ponomareva et al., 2023) apply to both full fine-tuning and parameter-efficient tuning of LLMs. Specifically, it's usually better to train longer and with larger batch size (Ponomareva et al., 2023). Also it's important to do a hyperparameter sweep over learning rate and clipping norm, see appendix J.2. We also experimented with various parameters of temperature sampling for data synthesis. We found that sampling with $T = 1$, without truncating sampling distribution (i.e. TopK $= \infty$) and without performing any filtering usually well and is a reasonable default setting, refer to Appendix J.3.

We also looked into variations of loss formulation for LLM training (Appendix J.4). As explained in Section 4.1 we found that Prefix-LM (Raffel et al., 2020) formulation usually leads to better performance in DP setting, compared to Full LM. We observed that normalizing LLM loss by number of non-padding tokens, as suggested in (Ponomareva et al., 2022), makes it easier to tune LLM hyperparameters, especially clipping norm for DP. It also makes it easier for LLM to learn to produce outputs of various length.

J.1 ABLATION: PRE-TRAINING DATASET AND MODEL SIZE

We looked into influence of LLM pre-trainind dataset and model size on performance on downstream task. As shown in table 8 larger size of LLM translated into better downstream performance for both BERT and CNN models.

Table 8: Influence of model size on downstream performance. All models are finetuned or prompt tuned for 10 epochs with 1024 batch size. Prompt tuning is done with loss normalization, fine-tuning is without. DP level is $\epsilon = 1$.

Model size	Prompt tune		Fine tune	
	BERT	CNN	BERT	CNN
1B	83.2 ± 1.1	76.9 ± 0.3	74.7 ± 2.4	61.1 ± 0.6
8B	86.2 ± 0.4	83.2 ± 0.4	85.5 ± 0.2	81.7 ± 0.3

Additionally, we did some limited study comparing pre-training on C4 and The Pile datasets. Initially we expected that C4 (which is essentially a web crawl) better matches text distribution in Yelp and IMDB datasets compared to The Pile (which was intentionally composed to be more diverse). However our experiments showed similar performance on both datasets, so eventually we settled on The Pile, which is easier to download and use.

J.2 ABLATION: HYPERPARMETER TUNING FOR DP-TRAINING OF LLM

Batch size and number of training steps. For our prompt tuning run on IMDB with $\epsilon = 1$ we did a thorough sweep of various batch sizes and number of training steps of LLM prompt tuning. Results are summarized in table 9 and confirm the general observation that longer training with larger batch is typically better when trained with DP.

Table 9: This table show how downstream performance depends on batch size and number of training steps used for LLM prompt tuning. Experiments were done on IMDB task, synthetic data were generated at privacy level $\epsilon = 1$. Noise multiplier for each training run was computed to satisfy privacy budget given chosen number of steps and batch size. Batch size 1024 with 220 steps correspond to 10 epochs of LLM training on IMDB dataset.

Batch size	BERT			CNN		
	220 steps	440 steps	880 steps	220 steps	440 steps	880 steps
1024	85.5 ± 0.7	85.1 ± 0.6	87.8 ± 0.2	82.4 ± 0.4	83.2 ± 0.4	84.3 ± 0.1
2048	85.7 ± 0.2	86.4 ± 0.9	87.9 ± 0.1	83.8 ± 0.2	83.2 ± 0.2	85.0 ± 0.2
4096	86.0 ± 0.6	86.6 ± 0.2	88.1 ± 0.4	82.8 ± 0.4	83.8 ± 0.2	85.4 ± 0.1

Learning rate. In our experiments we found that downstream performance was quite sensitive to learning rate of LLM, see table 10.

Table 10: Sensitivity of downstream task performance to learning rate used for LLM tuning.

(a) Finetuning

Learning rate	BERT	CNN
1e-2	Diverge	
3e-3 (optimal)	85.5 ± 0.2	81.7 ± 0.3
1e-3	69.1 ± 3.0	53.1 ± 0.8

(b) Prompt tuning

Learning rate	BERT	CNN
3e-3	83.4 ± 0.5	80.0 ± 0.4
1e-3 (optimal)	86.2 ± 0.4	83.2 ± 0.4
3e-4	84.7 ± 0.9	80.1 ± 0.6

J.3 ABLATION: SAMPLING PARAMETERS.

After training a generative model, we still need to use that model to create a differentially private dataset. Large language models generate sequences one tokens at a time in an autoregressive manner; given previous tokens the model's final layer emits a probability distribution over all possible next tokens. Instead of sampling directly from this probability distribution, it is common to modify it in three ways:

1. **Temperature**: Shaping the token distribution using temperature t so that the final softmax over the final logits u_i gives the next token probably as

$$P(z_i|z_{<i}) = \frac{exp(u_i/t)}{\sum_i exp(u_i/t)} \tag{2}$$

2. **Top K**: Truncating the token distribution so that only the k most likely tokens are sampled from. All other tokens are given zero probability.
3. **Num Decodes**: Repeating the full sampling process (i.e. decoding) N times and then out of the N candidates return only the sample with the highest likelihood.

To analyze the effect of these parameters on dataset quality we performed a sweep over these parameters and computed the downstream performance of models as discussed in section I. Results are shown in tables 11 and 12.

Results from our ablation study on sampling parameters show that while small gains can be gained from tuning these parameters, such gains are modest compared to using the default parameters of $t = 1.0$, top-k $= \infty$, and num decodes $= 1$. We note that slightly higher temperatures appear to help in both cases (1.4 for IMDB and 1.2 for Yelp) and in both cases should be paired with either tighter top-k or an increase in decodes. However, using additional decodes are computationally costly and likely not worth the additional cost. Using a temperature less than 1.0 never seems to help.

J.4 Ablation: loss of LLM.

LLMs are commonly trained/fine-tuned with next-token prediction teacher forcing. The loss for each token in this setup is a cross-entropy loss for each token. For an instance that is a collection of tokens (e.g. a full yelp review), the loss is therefore is a sum over per-token losses. It is common to normalize this loss by the number of non padding tokens, which roughly translates into the normalization by the batch size and the average number of tokens for instances in the batch. We recommend to follow this normalization scheme because it makes it much easier to find the appropriate clipping norm for DP-Training. When the loss is not normalized, the appropriate clipping norm can be in the thousands. With normalized loss, a standard clipping norm of 1 or 3 usually works out of the box. For example, for Yelp dataset, without the loss normalization the clipping norm was found to be approx 2000, with accuracy of the fine-tuning of approx 0.29. With loss normalization, the clipping norm of 1 resulted in performance of 0.43.

Note that if it's feasible to perform full hyperparameter sweep of clipping norm and learning rate, then benefits of loss normalization diminishes. Nevertheless even in this case loss normalization can provide small advantage, see Table 13.

J.5 Ablation: LoRa parameters.

We conducted detailed sweep of LoRa parameters (rank and in which layers to introduce LoRa) on IMDB and AGNews, see Tables 14 and 15. As could be seen from these tables, Full-LoRa performs better than Attention-Lora and MLP-Lora in most cases on IMDB dataset. However MLP-Lora seem to be better choice overall on AGNews. If practitioner has to pick parameters to introduce LoRa beforehand without tuning, then we would recommend Full-LoRa as a reasonable default.

In terms of rank, best performance is typically achieved for ranks in range $[8, 32]$. From our experiments, it does not make sense to increase rank above 32 because it result in little to no performance gains, however it is more expensive because requires tuning of more parameters.

K Evaluating synthetic data quality: Mauve robustness.

Mauve has multiple parameters that control its behavior. The most influential such parameters include: the degree of dimensionality reduction (PCA explained variance), the number of clusters to use, the number of samples to use, and the model used to initially embed the samples. Pillutla et al. (2021) came to the conclusion that while these affected performance none of these parameters mattered enough to worry about needing to tune for a specific application. They recommended a default setting of buckets $= 500$ and explained variance $= 0.9$, We performed our own ablation studies on these

Table 11: Comparison across sampling parameters for best performing prompt-tuned epsilon=1 model for the **IMBD** task. The parameters Temp $= 1.0$, TopK $= \infty$, Decodes $= 1$ corresponds to the default used in this paper.

Temp	TopK	Decodes	BERT Accuracy	CNN Accuracy
0.6	∞	1	80.7 ± 1.4	78.4 ± 1.2
		2	78.7 ± 1.8	76.9 ± 1.4
		4	79.5 ± 3.0	76.3 ± 1.6
	100	1	81.1 ± 0.7	77.9 ± 1.5
		2	78.1 ± 1.1	77.8 ± 0.5
		4	79.9 ± 1.0	75.9 ± 0.7
	1000	1	79.9 ± 0.7	77.1 ± 1.4
		2	78.7 ± 1.4	75.4 ± 2.4
		4	78.1 ± 2.3	77.8 ± 1.3
0.8	∞	1	84.7 ± 0.5	82.2 ± 0.9
		2	83.5 ± 1.4	82.8 ± 0.4
		4	82.8 ± 0.4	81.9 ± 1.0
	100	1	84.9 ± 0.6	82.2 ± 0.6
		2	85.4 ± 1.1	81.6 ± 0.7
		4	82.1 ± 1.0	82.9 ± 1.0
	1000	1	84.1 ± 1.0	82.2 ± 0.2
		2	82.9 ± 1.0	82.1 ± 0.9
		4	82.1 ± 1.1	83.1 ± 0.3
1.0	∞	1	85.3 ± 0.4	84.2 ± 0.6
		2	86.1 ± 0.2	83.2 ± 0.1
		4	85.4 ± 0.4	83.5 ± 1.7
	100	1	84.5 ± 0.7	84.9 ± 0.4
		2	86.8 ± 0.4	84.5 ± 0.3
		4	84.4 ± 0.9	84.7 ± 0.3
	1000	1	86.3 ± 0.4	83.7 ± 1.0
		2	86.5 ± 1.7	84.3 ± 0.3
		4	85.6 ± 0.4	84.2 ± 0.3
1.2	∞	1	87.1 ± 0.7	77.4 ± 1.2
		2	86.0 ± 1.7	78.5 ± 2.9
		4	85.6 ± 0.9	79.9 ± 1.5
	100	1	85.8 ± 1.3	84.4 ± 0.1
		2	87.2 ± 0.4	84.8 ± 0.6
		4	86.5 ± 1.0	85.2 ± 0.2
	1000	1	86.7 ± 0.2	81.9 ± 0.2
		2	86.9 ± 0.6	81.9 ± 0.7
		4	86.4 ± 0.3	82.3 ± 0.5
1.4	∞	1	84.1 ± 1.6	60.0 ± 3.5
		2	85.5 ± 1.2	66.1 ± 1.1
		4	81.6 ± 2.0	60.6 ± 2.9
	100	1	86.5 ± 0.6	83.3 ± 0.2
		2	86.9 ± 0.3	84.3 ± 0.2
		4	85.8 ± 0.9	84.4 ± 0.2
	1000	1	85.3 ± 1.2	77.2 ± 2.2
		2	86.4 ± 0.2	78.3 ± 0.8
		4	86.6 ± 0.4	79.2 ± 0.4

parameters (see Figure 3). As noted in the paper, while we found MAUVE to be robust to most of these parameters, the model mattered a great deal.

L Evaluating synthetic data quality: Proxy metric experiments.

Proxy metrics are a less expensive method for estimating synthetic data quality. They are particularly useful for helping tune the large number of hyperparameters needed to train and sample from the generator model. This includes tuning the model type (model architecture, pre training process, fine-tuning, prompt-tuning, etc..), the hyper parameters of model training (e.g. epsilon, clipping norm, learning rate, batch size, epochs, etc..), and finally the hyperparameters of the sampling procedure

Table 12: Comparison across sampling parameters for best performing prompt-tuned epsilon=1 model for the **Yelp** task. The parameters Temp $= 1.0$, Top-K $= \infty$, Decodes $= 1$ corresponds to the default used in this paper.

Temp	Top K	Decodes	BERT Accuracy	CNN Accuracy
0.8	∞	1	92.6 ± 0.6	87.9 ± 1.1
		2	89.9 ± 1.1	85.0 ± 1.3
		4	88.3 ± 0.1	81.7 ± 0.9
	100	1	91.7 ± 0.6	86.6 ± 1.6
		2	89.9 ± 0.6	84.2 ± 1.5
		4	87.5 ± 1.6	80.5 ± 0.4
	1000	1	92.5 ± 0.3	87.9 ± 0.6
		2	90.4 ± 1.4	85.4 ± 1.7
		4	86.9 ± 1.1	82.5 ± 1.9
1	∞	1	93.4 ± 0.7	90.9 ± 0.5
		2	93.9 ± 0.5	90.5 ± 0.4
		4	93.7 ± 0.6	90.5 ± 0.3
	100	1	93.2 ± 0.5	90.0 ± 0.5
		2	93.3 ± 0.2	89.4 ± 0.6
		4	92.4 ± 0.2	88.3 ± 0.3
	1000	1	93.9 ± 0.2	90.8 ± 0.2
		2	94.0 ± 0.2	91.1 ± 0.1
		4	93.8 ± 0.2	89.7 ± 0.6
1.2	∞	1	93.6 ± 0.3	90.1 ± 0.2
		2	93.5 ± 0.2	90.3 ± 0.0
		4	94.0 ± 0.1	90.6 ± 0.1
	100	1	93.7 ± 0.2	91.2 ± 0.1
		2	93.7 ± 0.1	91.3 ± 0.3
		4	93.8 ± 0.2	91.3 ± 0.2
	1000	1	94.1 ± 0.1	90.7 ± 0.2
		2	94.0 ± 0.3	90.9 ± 0.1
		4	94.2 ± 0.1	91.0 ± 0.1
1.4	∞	1	93.2 ± 0.2	86.2 ± 1.7
		2	93.1 ± 0.4	87.2 ± 0.8
		4	92.9 ± 0.8	86.7 ± 0.6
	100	1	93.4 ± 0.4	90.8 ± 0.0
		2	93.5 ± 0.4	90.9 ± 0.0
		4	93.5 ± 0.0	91.0 ± 0.0
	1000	1	93.4 ± 0.1	88.9 ± 0.1
		2	93.5 ± 0.0	89.2 ± 0.1
		4	93.4 ± 0.5	89.2 ± 0.3

Table 13: Effect of loss normalization with sufficient hyperparameter tuning.

	BERT		CNN	
	Loss Norm	No Loss Norm	Loss Norm	No Loss Norm
IMDB prompt tuning, $\epsilon = 1$	86.4 ± 0.9	86.0 ± 0.7	83.2 ± 0.2	83.0 ± 0.5

needed to create the final dataset (e.g. temperature, top-k, and decodes). We examined how well various metrics correlate with downstream performance of a final classifier trained on the synthetic data (Figure 4). Since we are interested in selecting the best performing parameters, we want a metric that is most likely to select a high quality model and thus should care primarily about the rank correlation.

M Implementation details and estimates of the required compute

LLM training. We run all our experiments using T5X codebase and used implementation of transformer layers from Flaxformer library. For differentially private training of transformers we used private_text_transformers repository.

We pre-trained 1B model on TPUv3 with 128 cores and 8B model was pretrained on TPUv3 with 1024 cores, both runs took around 8 days. Both finetuning and prompt-tuning of 8B models was done

Table 14: Performance of downstream classifier depending on LoRa parameters, when LLM is trained on IMDB dataset with $\epsilon = 1$.

	LoRa	Rank 1	2	4	8	16	32	48	64
BERT	Full	85.4 ± 0.2	85.9 ± 0.3	89.7 ± 0.1	90.0 ± 0.3	90.0 ± 0.2	89.7 ± 0.2	89.9 ± 0.3	90.2 ± 0.2
	MLP	84.4 ± 0.5	85.9 ± 0.3	88.2 ± 0.4	86.1 ± 0.1	88.2 ± 0.5	87.9 ± 0.2	86.1 ± 0.2	88.1 ± 0.7
	Attn	82.5 ± 0.4	88.8 ± 0.1	87.7 ± 0.3	89.7 ± 0.3	90.1 ± 0.1	89.0 ± 0.3	89.8 ± 0.3	89.6 ± 0.1
CNN	Full	75.5 ± 0.8	84.1 ± 0.7	85.5 ± 0.2	87.6 ± 0.4	87.4 ± 0.3	87.2 ± 0.2	87.3 ± 0.2	87.5 ± 0.3
	MLP	71.9 ± 0.6	81.7 ± 0.4	86.8 ± 0.1	81.6 ± 0.9	85.7 ± 0.2	85.5 ± 0.2	83.6 ± 0.3	85.1 ± 0.3
	Attn	72.2 ± 1.9	86.9 ± 0.3	84.0 ± 0.1	87.4 ± 0.2	87.3 ± 0.1	85.7 ± 0.1	87.3 ± 0.1	86.8 ± 0.4

Table 15: Performance of downstream classifier depending on LoRa parameters, when LLM is trained on AGNews dataset with $\epsilon = 1$.

	LoRa	Rank 1	2	4	8	16	32	48	64
BERT	Full	88.1 ± 0.1	88.0 ± 0.1	88.2 ± 0.3	88.5 ± 0.1	87.9 ± 0.1	88.3 ± 0.2	88.8 ± 0.1	88.9 ± 0.1
	MLP	87.9 ± 0.1	88.0 ± 0.2	88.2 ± 0.0	88.4 ± 0.1	88.8 ± 0.3	89.4 ± 0.1	88.3 ± 0.2	88.3 ± 0.1
	Attn	87.7 ± 0.2	88.4 ± 0.1	88.1 ± 0.2	88.0 ± 0.3	88.2 ± 0.1	88.7 ± 0.1	-	-
CNN	Full	85.3 ± 0.1	85.0 ± 0.1	85.2 ± 0.1	85.4 ± 0.1	84.6 ± 0.2	85.3 ± 0.1	85.6 ± 0.1	85.7 ± 0.2
	MLP	84.9 ± 0.2	85.0 ± 0.1	85.2 ± 0.1	85.7 ± 0.1	85.6 ± 0.2	85.8 ± 0.0	85.2 ± 0.1	85.2 ± 0.1
	Attn	85.2 ± 0.1	85.3 ± 0.1	84.6 ± 0.1	85.0 ± 0.1	85.1 ± 0.1	85.4 ± 0.1	-	-

on TPUv3 with 128 cores. Diferentially private finetuning of 8B model required between 20 hours (for shortest run on IMDB) and up to 80 hours for some of the longer runs. On the other hand prompt tuning required 1.5 hour for short run and up to 20 hours for longest runs.

Downstream model. Downstream classifier was implemented in Tensorflow using Keras library for training and TFDS to load datasets.

Each downstream model was run on TPUv2 with 8 cores. To obtain each downstream accuracy number we run a sweep of around 28 different hyperparameter settings. Entire sweep took around 4 hours for CNN model and up to 80 hours combined for BERT model and synthetic dataset of 0.5M examples. Each sweep was repeated 3 times to compute error bars.

N EXAMPLES OF GENERATED AND REAL DATA

Table 16 shows examples of generated synthetic data.

Table 16: Examples of real and synthetic datasets.

Eps	Dataset	Class	Example
∞ (real)	Yelp	Negative	Mediocre burgers - if you are in the area and want a fast food burger, Fatburger is a better bet than Wendy's. But it is nothing to go out of your way for.
		Negative	Not at all impressed...our server was not very happy to be there...food was very sub-par and it was way to crowded. Not the good kind I crowded where you feel like \""wow this is great it must be busy because the food is so great..\"" But the type of crowded where you feel a fight may break out. Also, if the chocolate fountain is their golden gem...why is it ok for people to dip the strawberry,lick off the chocolate and re dip it right then and there...absolutely disgusting! I Waited almost 10 minutes to dip...saw that and was immediately turned away...just saying...Never again!
		Positive	I know, I know a review for Subway, come on. But I have to say that the service at this subway is top notch. The staff is very friendly and always goes out of there way to say hello. It is highly appreciated. The food is pretty standard Subway fare. Not too much to say there, but I will be be back again on my lunch breaks.
∞ (synthetic)	Yelp	Negative	I waited for a table to be seated but it was not until 45 minutes into my 45 minute lunch. The service and attention to detail is outstanding. I got served by one gentleman which was fine until another needed help bringing an food order. He was not willing to ring me back until the staff had technically completed more than just a couple steps in the process. This was one of the most memorable bad encounters I have had with service anywhere. It was so beyond poor that I couldn't even settle on a 1 star review for the sole purpose of the bad service.This is managements reptile effect. I may even review this restaurant on the yelp website as part of the bad service. ? r ? n ? r ? nWhat i do remember about the food was that I got the (medium well)lamb chops which my family enjoyed. I believe this was the only 4.5 star review; all others are 5 stars. ? r ? n ? r ? nThank you Taste of Thai for being so exceptional. I will definitely eat here again.\" 2083 S. Colorado Ave. ? r ? nLansing, MI 492019 ? r ? nEmail: revgiebun@yahoo.com ? r ? n ? r ? n\"[eod] [pile] ; RUN: llc,?,%s -mtriple=sparclite-serializer-macho -o /dev/null ; The powerpc 64 bit ABI hard float ABI contains 32-bit values types ; (e.g. 6 (struct half*) or 7 (struct bfin64_2float3216*)). The Cortex-A8 ; processor doesn't contain facilities to load bulky 32-bit types (it only ; has equivalents for 8-bit and 16-bit types and it joined processing ; these groups into the same registers) ; RUN: ld.ll,?,%s,?,%p/../Inputs/structures.s ; RUN: llvm-objdump -h -j ../Inputs/structures.exe.macho.x86_64 -b binary

		Negative	Restaurant is not worth the drive other places near Ames and Des Moines have better options. Service was weird and owner was a jerk for not doing anything with it. Saw some actual fish caught and skinned by people. They would take their picture just put it,on the wall. Submitted by: TomRindal A Restaurant Review of Guadalabar [yelp],?,?,?,?,? ! Food could be a lot tastier if we paid more for the whole dining experience. They only charged me for one third of my dish and always put a kick in the case. However, the fish and ric e?! Huh? Fish A..a..er... jiabitangellois so so bad..and it's served on top of the rice which kills all of the yuminess. There are still so many great places to eat in this town, just need a little more energy to do it MrTripA4NO beer!!!! Good place to bring children... a dinky, cr appy fake Spongebob movie saturday morning in a 10 year old's restaurant.... been to Des Moines a few times and tried to find something tastier then Guadalabar, but at least they had smoothies.. and know what you are getting before you overpay.. A great nd cheap breakfast! $2 sandw ich for \"bonus\"... A calligraphy class flyer!! wow.. our waiter was the dishwasher; the restaurant was dirty & smelled like wet concrete. Definitely would not recommend.... this place has no charm... Low Yelp rating = the customer service and manager are living up the secret. May be its just another Ogden this time next year. Next time demote after 3rd year.,? ! ?,?,? cheesy ? ! ? ! Had three dinners at Guadalabar. Last one was last thurs, Wow! The pig in a blanket is outstanding! If lamb n onions don't click your menu versus noodle risotto is a clear,must. My partner got the cheese appetizer. Decent! The fish sans name was tasty, and definitely
		Positive	Super cool food and interesting menu. Anything or everything can be made vegan and they are more than willing to cater to the vegan individual or families. I went in with a group of sisters who aren't into tofu and cepes on a regular basis. They were thrilled! I ordered the Mixed Marinara Veratofu which its made with squash, gnocchi, mushrooms, onions and is all natural. The Greens are also amazing, with quinoa and not so many red charcters like burgers. Look forward to this place anytime. ? n The best part is that they are great to the environment ? n Great service (despite the fact it took a long, long time for the food to arrive), and low prices. Great food and it is away from downtown SF ? n We are fans! ? n ? n ? n Florence 2232 S Broadway #102 ? n Studio City, CA 91306 (562) 285-9947 http://theamazingvegetemenu.com ? n ? n ? n DTLA #100 ? n Serving: Late Night Until 2 AM ?
1	Yelp	Negative	Best bahn in Belgium 50 years ago and 50 years later. First, the breakfast is awful and thin. Reviews on yelp keep complaining. Most certainly were never sold in Poland. The taste is not bread at all, but as crumbly and flat pancakes filled with sugar.
		Negative	If you are planning a party night out, don't go here. They back stab you, yell at you to leave the store, and are rude and rude. I told my mom that if we went to Frisch's for party supplies she would stay away a good 10 miles away. Thank goodness for Walmart.
		Positive	I am actually surprised they weren't busy like Wendi tweeted me. What a beautiful restaurant and atmosphere! Please take my advice and make sure you stroll around golf course for some free and attractive restaurants, shops and more.
		Positive	OMG, WOW, WHAT IS THIS?? Go for the Thai Iced Tea. My usual order is iced capp and has tea but this is better. Reliable, inexpensive, and delicious! I have never had this before and this was a revelation!
3	Yelp	Negative	Good choice, bad service. I live in Aliso Viejo and the service at Windrider Station has always been great but our experience this evening was profoundly disappointing. The services mood, warmth, and grace is what set Windrider Station services apart from the rest, so I was surprised to see it be rendered to an adept officer.
		Negative	Been to the place a few times for Detroit wings. The taco meat was uneatable.. .carelessness in processing.. seriously.. wasn't a tasty taco at all. .the frog tails were actually fried pieces of fish.. huge turn off. The unsavory tails..the frog wasnt half done.. all the meats were cold.. silly tacos for the price..
		Positive	We got a private room for three. When we arrived, the bartender was very pleasant and attentive. They had an early dinner special ($16) that was not available, so we had only two choices (entrees): Paradiso in Chilean Seafood & Lots of Focaccia ($10.99) or Steelhead Salmon ($17.99). The entrees came pretty quickly.
		Positive	These guys can help you with same day prices on everything you need. Best service - best prices! I come back here every time I need something.
10	Yelp	Negative	my boyfriend and i went there a couple of times - it was pleasant and very modest and they had a GREAT happy hour! but on the dinner side, it's not really worth the price. the atmosphere was nice, but the food was so-so and the amount of service was adversely affected - two servers there at a time only for dinner service
		Negative	I must strongly recommend that anyone planning to visit Vegas anytime soon avoid this location. The staff (the two servers and their food delivery person) were all rude, insulting and insensitive. I have been to many places and have had very bad experiences and have never returned to any of them.
		Positive	Ok so I went for their lunch buffet which cost like 34 clams. This is the second buffet buffet I've been to here in the last couple months and I always noticed that theirs are always more than 80 dollars per person, which is a a little high, but it's more than worth it.
		Positive	This place is definitely cool. They have a really cool beer selection and a couple salsa dancers! nIboy was the DJ at the event, and he's the real deal. His set was fast paced and exciting! If you get a chance to throw a party there, don't hesitate!
∞ (real)	IMDB	Negative	This was an absolutely terrible movie. Don't be lured in by Christopher Walken or Michael Ironside. Both are great actors, but this must simply be their worst role in history. Even their great acting could not redeem this movie's ridiculous storyline. This movie is an early nineties US propaganda piece. The most pathetic scenes were those when the Columbian rebels were making their cases for revolutions. Maria Conchita Alonso appeared phony, and her pseudo-love affair with Walken was nothing but a pathetic emotional plug in a movie that was devoid of any real meaning. I am disappointed that there are movies like this, ruining actor's like Christopher Walken's good name. I could barely sit through it.
		Positive	This is the kind of film for a snowy Sunday afternoon when the rest of the world can go ahead with its own business as you descend into a big arm-chair and mellow for a couple of hours. Wonderful performances from Cher and Nicolas Cage (as always) gently row the plot along. There are no rapids to cross, no dangerous waters, just a warm and witty paddle through New York life at its best. A family film in every sense and one that deserves the praise it received.
		Negative	The film is based on a genuine 1950s novel.Journalist Colin McInnes wrote a set of three "London novels": "Absolute Beginners", "City of Spades" and "Mr Love and Justice". I have read all three. The first two are excellent. The last, perhaps an experiment that did not come off. But McInnes's work is highly acclaimed; and rightly so. This musical is the novelist's ultimate nightmare - to see the fruits of one's mind being turned into a glitzy, badly-acted, soporific one-dimensional apology of a film that says it captures the spirit of 1950s London, and does nothing of the sort.Thank goodness Colin McInnes wasn't alive to witness it.
∞ (synthetic)	IMDB	Positive	"This movie is filled with twists and turns from the first moment it enters the big screen, until the final moments of the film. It gives you enough information to keep you from extrapolating into going anywhere other than the truth. And the only way to know the truth is to simply pay money to see (probably the more fun thing to do, money being the "other" presence in this equation) the DVD (if you just have Netflix, just buy it it's cheaper). ? br />? br />This is one of those movies like Meme ...
		Negative	"I thought they rewrote a much better movie that made some changes here and there. Even though the plot was a bit slow and idiotic at times, it was acted well by all of its leads, including Morgan Freeman. It had little to do with the original story, other than the fact that Morgan Freeman played Maurice, the dumb Irish dude from The Usual Suspects."
		Positive	"Nice deserted beach nude scene. ? br />? br />Sourced from a DVD I picked up a few months ago, this is the 1972 Oscar winning Fernando Trueba documentary of the legendary Sea Stars along Catalan coastline beaches. The 1970s, when Trueba was a young artist who traveled to beaches all over Europe, was a great soul renaissance of the French Riviera, 1970-83. The 1970s Paris Inter City was the main hub of the beach culture, and Spanish Swimsuit was the fraternal new generation's version of Euro Tr ...

1	IMDB	Negative	So... Off the top. Ill start this off saying that I thought Ethan Hawke and Dakota Fanning were excellent, and this movie + I were great. However, this movie was just too sad and depressing + I kept falling asleep when the characters visited their pasts to either s̈aveör ëscape. The only good scene was that fanning sung in the shower with her father on his death bed, she actually gave him her first big laugh... incredible. All and all, this film was a horrible waste of time. I'm almost embarrassed for rating it this bad, because I enjoyed it when I saw it at the movies, but now that I had the time to re-watch it I decided to deduct a point or two.
		Negative	An underwhelming, blandly written, dry, dull romance. A subtle attack on the Bible and marriage as the supreme goal of woman. Toni Collette's Stephie She is the perfect stereotype of the Southern housewife, flawed but go-getting, a role that was traditionally assigned to the woman and embodied by stereotypical southern belles. Alexandra and Danny represent the emerging of the consumerist age, and don't consider themselves traditional, but not dirty minded enough to change the old ways, even as they clash with Collette's Stephie, an emancipated, tough, independent, rebellious young woman living the new age. Directors Jim Gianopulos and Roman Polanski are wacky, and impressionistic.
		Positive	This movie is what they say - it is well done, as of perfection. This is a true story of a man on the run for stealing his best friend's business and the mental battle of escaping by himself across many states. The shooting scenes are good as a bootlegger must shoot his way through a posse, but unfortunately the ending and some events which I'll keep for private reasons lead to a trickle down effect between a friend killing the lead female singer and the actor who leads the posse. The rest of the movie though is good. If you read the book you'll notice that the names are changed because a book wouldn't allow too many people to get away with their crimes. The film is the blueprint for success.
		Positive	This movie seriously features one of the actress' very best performances. Jenny Klein plays a mysterious writer that appears to write as a zombie (she appears to feed the moat of her garden and dead animals to her zombie voice) then right after she finds out that her husband has cheated on her in a bizarre revenge scenario, she has to confess to the world that she is a writer as well as have some mysterious SCREAM revelation. This bravura performance allows the viewer to meets the father of childhood and perform real art in a particularly violent but weirdly charming way. Don't miss it.Recommended to lovers of H.P. Lovecraft and T̈he Babysitters Club.
3	IMDB	Negative	Training day ripoff, producer Related Entertainment The Movie Kitoshia Izumi plays J̈acob Marloweä doctor who loves making macho statements like Ï'm not sexually attracted to womenänd ẅomen are pigs. With a woman dressed in a pig outfit that looks like something peroxided from a beauty parlor, lumbo-lumbagoed into place. N̈ia Catastropheïs about a Police officer (Marlowe) who decides to become a doctor to save the lives of the citizens of Beaver Falls (Pennsylvania). But the lives he saves will only result in the eventual death of all the headliners he's supposed to be bringing back to life. Izumi's insufferable character is almost laughable. The attempt to portray him as sensitive and caring, never works, either. Things get worse when there's a drug dealing merchant on the bus who pretends to be a passenger. But most of the movie is a parade of wretched acting.
		Negative	I was bored watching this film. The characters weren't sympathetic and the writing was horribly melodramatic. The ending was so petty. It smacked of the cliche' l̈ook how the cat got his revengeẗheme. I'll wait until someone makes a slicker, funnier film out of this material.
		Positive	I would say all of the 3 or 4 days I was watching Golden Boy, I was very interested. Even a small, almost anonymous person with no name, a poor little bird, a small and weak sport - french- speaking effet, from a city without a name. That was his table. Don't watch his card but stop, look and marvel. The way he fights, his shots, how he makes an attack, how he avoids the devastating blow. He fights consistsantly, competently, always attacking, never believing his opponent is able overwhelm him.
		Positive	This is the perfect older brother summary. I know very few children whose real siblings are more mature then them. Mina Wang is a gifted doctor. Here she is an advisor, takes care of a far off plantation of the nature. This girl is the center of all things worng. She comes up with a plan to get revenge on Tom, father of her boyfriend Peter. Tom has just made a big mistake that could really affect Milo as a child.
10	IMDB	Negative	First, why, on the LORD'S day, was this movie commercial to Bears players head-upar water, a DH baseball player, and a big, hairy bachelor? Funniest irespondent thing I've ever seen in my life. Second, don't do it again, JW. Didn't cut it with Bears baseball, so don't do it with movies.
		Negative	i known i'll please a majority of people while a few others will vilify me for this review, but one deduction i might make is that i have seen movies much more professional, i.e. directed by m̈asters, made with more professional actors of course, etc
		Positive	I love Westerns and shows that take a somewhat exaggerated and satirical view of life. This is such a show! If you like Westerns, this one is a must-see. Of course The Cisco Kid is a fine example of that favorite genre. Mr. Melish adds the character of Mochica to the Chicken Ranch, not many westerns do wilder and funnier things. Well worth watching!!!!!!!
		Positive	A story based on the life of one of Sad Hill Folks'. Being a character of how his life has shaped him, this film was directed/narrated by the man himself. By novelising the story I was able to view a character worthy of a film, and a full 45 minute Theme song! Totally beautiful.

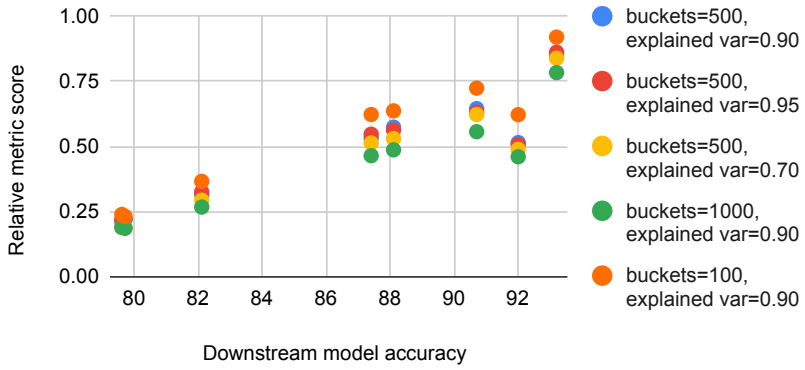

Figure 3: Example of varying MAUVE parameters on estimating IMDB downstream performance on datasets differing in training epsilons from Table 1. Results are shown for the Sentence-T5-8B model.

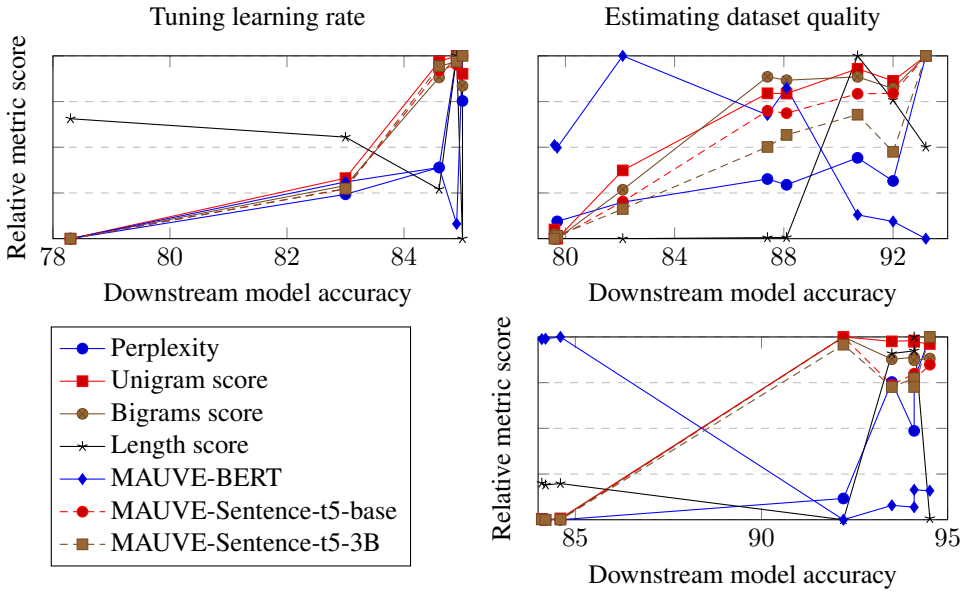

Figure 4: Proxy metrics for estimating dataset quality. Each point represents a metric's estimate of a synthesized dataset plotted against its true downstream classifier performance. X-axis shows the metric values re-scaled. All metrics are only useful to compare datasets, and thus their absoluteabsolute value is uninformative. Top Left: Different learning rates of an IMDB prompt-tuning model. Top Right: Estimating IMDB dataset quality for results in Table 1. Bottom Right: Estimating Yelp dataset quality for results in Table 1.