Not All LLM-Generated Data Are Equal:
Rethinking Data Weighting in Text Classification

Hsun-Yu Kuo*, Yin-Hsiang Liao*, Yu-Chieh Chao,
Wei-Yun Ma, Pu-Jen Cheng

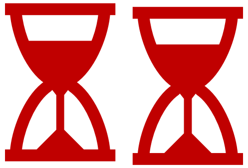
Academia Sinica, National Taiwan University, EPFL

Limited Real-World Data

Real-World P

Labelling

$$$



Large Real-World Dataset

$$D_P = \{(\mathbf{x}_i, y_i)\}_{i=1}^{M}$$

Human labelling is expensive and time consuming

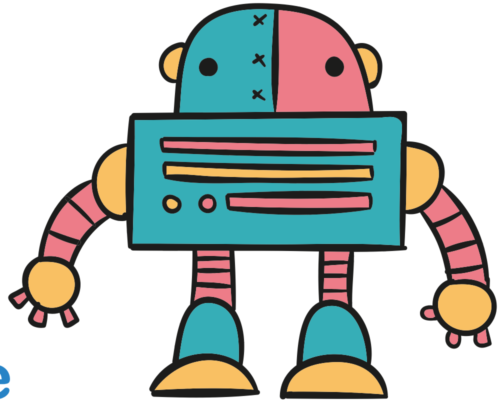
Training on Synthetic Data

Real-World P

Small Real Data

$$D_{P'} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{M}$$

As examples



LLM

Instruction

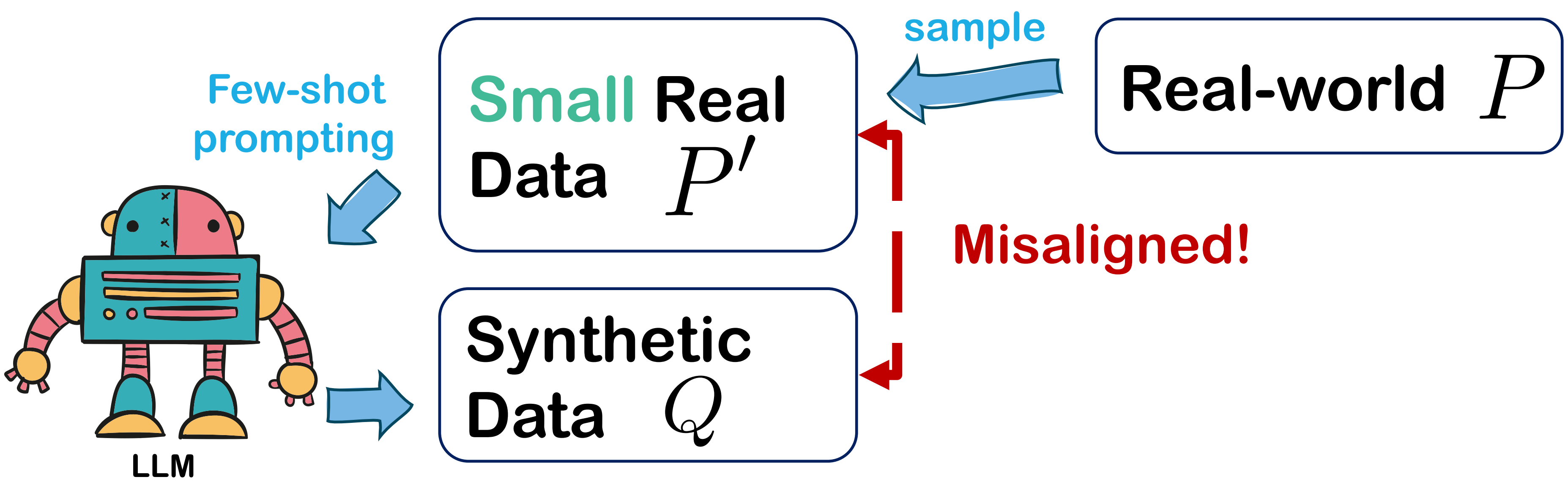
generate

Large Synthetic Data

$$D_Q = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$$

Misalignment from Synthetic Data to Real Data

- Only LLM-generated (synthetic) and small real data (around 200 - 400) are available



Synthetic Data Leads Unstable Performance

	Method	Financial		Tweet Irony		MRPC	
		Acc	F1	Acc	F1	Acc	F1
Large real-world data	CE-Loss	84.74	82.69	68.75	68.41	80.92	77.73
Small real-world data	CE-Loss (quality checker)	78.05	75.26	62.5	62.38	73.16	68.69
LLM-Generated data	CE-Loss	77.39	74.01	76.91	76.8	72	65.47

Sometimes even worse than small real-world data...

Weighted Loss Function

Synthetic Data Q

Real-world P

Small real-world P'

Weight function

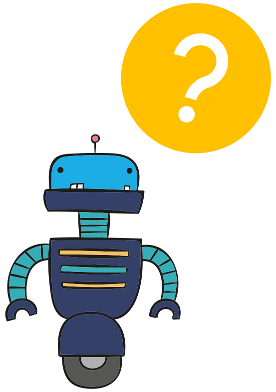
$$\mathcal{L}_{WCE}(\theta, D_Q, w) = -\frac{1}{N}\sum_{i=1}^{N} \overline{w_i} \log \underline{\hat{P}(y_i|x_i;\theta)}$$

Predicted probability of a model

Not all datapoints are equally important!
E.g. false data, hallucination, ...

Question:

Does it exist a weight function that can transform Cross Entropy over Q to P?



Transformation from Q to P

Synthetic Data Q

Real-world P

$$-\frac{1}{N}\sum_{i=1}^{N}\boxed{\frac{P(y_i|\mathbf{x}_i)}{Q(y_i|\mathbf{x}_i)}}\log \hat{P}(y_i|\mathbf{x}_i;\theta) \approx \mathbb{E}_P[-\log \hat{P}(y|\mathbf{x};\theta)]$$

Importance Weight function

Asymptotic Convergence to Cross Entropy over P!
(Under some reasonable assumptions)

Importance Loss (IMP-Loss)

Synthetic Data Q

Real-world P

Small real-world P'

Approximate P by fitting a model using small real data

$$-\frac{1}{N}\sum_{i=1}^{N}\frac{\hat{P}'(y_i|\mathbf{x}_i)}{\hat{Q}(y_i|\mathbf{x}_i)}\log\hat{P}(y_i|\mathbf{x}_i;\theta)$$

Approximate Q by fitting a model using synthetic data

Prioritize Quality and Diversity

Synthetic Data Q

Real-world P

Small real-world P'

⬆ High quality from real-world perspective

Quality Checker

$$-\frac{1}{N}\sum_{i=1}^{N}\frac{\overbrace{\hat{P}'(y_i|\mathbf{x}_i)}^{\text{Quality Checker}}}{\underbrace{\hat{Q}(y_i|\mathbf{x}_i)}_{\text{Diversity Checker}}}\log\hat{P}(y_i|\mathbf{x}_i;\theta)$$

Diversity Checker

⬇ Higher data diversity from generated dataset perspective (Lower Q)

Another Perspective

Synthetic Data Q

Real-world P

Which data makes the current model be closest to P?

$$\theta_{t+1} = f(\theta_t, \{\mathbf{x}', y'\})$$

f is one step optimization algorithm e.g. SGD

Objective:

$$(\mathbf{x}^*, y^*) = \underset{(\mathbf{x}', y') \in D_Q}{\arg\min}\ \mathbb{E}_P \left[-\log \hat{P}\left(y|\mathbf{x}; \overline{\theta_t, \{(\mathbf{x}', y')\}}\right) \right]$$

Which data point causes the model to be closest to P?

Synthetic Data Q

Real-world P

Small real-world P'

Objective:

$$(\mathbf{x}^*, y^*) = \underset{(\mathbf{x}', y') \in D_Q}{\arg\min} \; \mathbb{E}_P \left[-\log \hat{P}\left(y|\mathbf{x}; \theta_t, \{(\mathbf{x}', y')\}\right) \right]$$

Approximate P by using Small Real Data

$$\underset{(\mathbf{x},y) \in D_Q}{\arg\max} \; \hat{P}(\mathbf{y}_{P'} | \mathbf{X}_{P'}; \theta_t, \{(\mathbf{x}, y)\}) =$$

$$\underset{(\mathbf{x},y) \in D_Q}{\arg\max} \; \frac{\hat{P}(y|\mathbf{x}; \theta_t, D_{P'})}{\hat{P}(y|\mathbf{x}; \theta_t)}$$

Calculation is more tractable!!

Dynamic Importance Loss (DIMP-Loss)

Objective:

$$(\mathbf{x}^*, y^*) = \underset{(\mathbf{x}', y') \in D_Q}{\arg\min} \; \mathbb{E}_P \left[-\log \hat{P}\left(y | \mathbf{x}; \theta_t, \{(\mathbf{x}', y')\}\right) \right]$$

Real-World P

Synthetic Data Q

Small Real Data P'

DIMP-Loss:

$$-\frac{1}{N} \sum_{i=1}^{N} \boxed{\frac{\hat{P}'(y_i | \mathbf{x}_i)}{\hat{P}(y_i | \mathbf{x}_i; \theta_t)}} \log \hat{P}(y_i | \mathbf{x}_i; \theta_t)$$

Dynamic Importance Weight function

Dynamic Importance Loss (DIMP-Loss)

Objective:

$$(\mathbf{x}^*, y^*) = \underset{(\mathbf{x}', y') \in D_Q}{\arg\min} \ \mathbb{E}_P \left[-\log \hat{P}\left(y | \mathbf{x}; \theta_t, \{(\mathbf{x}', y')\}\right) \right]$$

Real-World P

Synthetic Data Q

Small Real Data P'

DIMP-Loss:

Fitting a model using small real data

$$-\frac{1}{N} \sum_{i=1}^{N} \frac{\hat{P}'(y_i | \mathbf{x}_i)}{\hat{P}(y_i | \mathbf{x}_i; \theta_t)} \log \hat{P}(y_i | \mathbf{x}_i; \theta_t)$$

Model itself

Quality and Diversity Checkers

IMP-Loss:

$$-\frac{1}{N}\sum_{i=1}^{N}\frac{\overbrace{\hat{P}'(y_i|\mathbf{x}_i)}^{\text{Quality Checker}}}{\underbrace{\hat{Q}(y_i|\mathbf{x}_i)}_{\text{Diversity Checker}}}\log\hat{P}(y_i|\mathbf{x}_i;\theta)$$

High quality from real-world perspective

Higher data diversity from synthetic dataset's perspective (Lower $\hat{Q}(y_i|\mathbf{x}_i)$)

DIMP-Loss:

$$-\frac{1}{N}\sum_{i=1}^{N}\frac{\overbrace{\hat{P}'(y_i|\mathbf{x}_i)}^{\text{Quality Checker}}}{\underbrace{\hat{P}(y_i|\mathbf{x}_i;\theta_t)}_{\text{Diversity Checker}}}\log\hat{P}(y_i|\mathbf{x}_i;\theta_t)$$

High quality from real-world perspective

Higher data diversity from model's perspective (Lower $\hat{P}(y_i|\mathbf{x}_i;\theta_t)$)

Both are Better on LLM-Generated Data

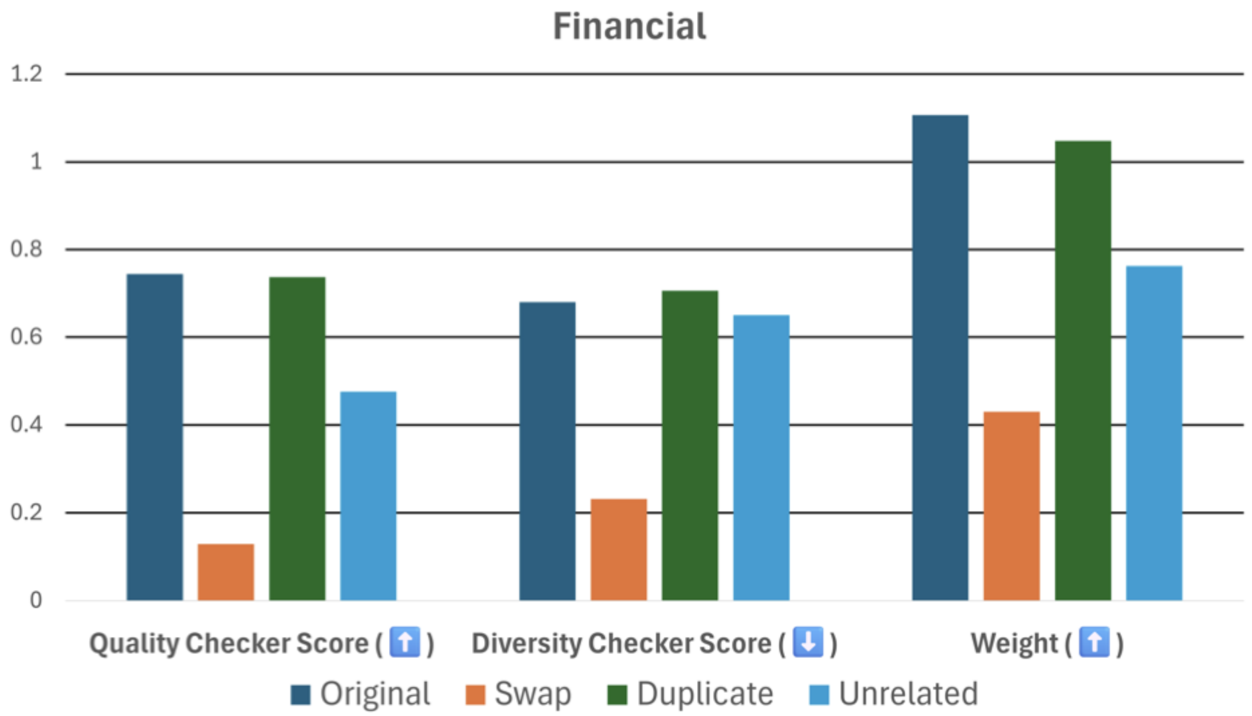
Dataset	Method	Financial		Tweet Irony		MRPC	
		Acc	F1	Acc	F1	Acc	F1
	GPT-3.5 few-shot	79.46	81.6	63.39	69.39	69.28	71.75
Small real world	CE-Loss (quality checker)	78.05	75.26	62.5	62.38	73.16	68.69
	Focal-Loss	78.47	76.2	67.73	62.32	73.10	66.64
	DIMP-Loss (Ours)	**79.87**	**77.05**	**69.01**	**67.05**	**74.84**	**66.80**
GPT-3.5 generated	CE-Loss	77.39	74.01	76.91	76.8	72	65.47
	Focal-Loss	79.29	75.32	74.87	74.82	72.17	62.77
	Hu et al.'s	71.7	61.93	71.42	70.18	67.13	50.08
	SunGen	80.45	76.87	78.96	75.06	71.65	66.08
	IMP-Loss (Ours)	82.09	**79.40**	**81.89**	**81.71**	**75.83**	**70.52**
	DIMP-Loss (Ours)	**82.67**	**79.53**	78.44	78.14	**75.83**	**70.04**
	- w/o diversity checker	81.35	77.94	77.68	77.62	74.72	69.34

Better than Models from Small Real-world Data

Dataset	Method	Financial		Tweet Irony		MRPC	
		Acc	F1	Acc	F1	Acc	F1
	GPT-3.5 few-shot	79.46	81.6	63.39	69.39	69.28	71.75
Small real world	CE-Loss (quality checker)	78.05	75.26	62.5	62.38	73.16	68.69
	Focal-Loss	78.47	76.2	67.73	62.32	73.10	66.64
	DIMP-Loss (Ours)	**79.87**	**77.05**	**69.01**	**67.05**	**74.84**	**66.80**
GPT-3.5 generated	CE-Loss	77.39	74.01	76.91	76.8	72	65.47
	Focal-Loss	79.29	75.32	74.87	74.82	72.17	62.77
	Hu et al.'s	71.7	61.93	71.42	70.18	67.13	50.08
	SunGen	80.45	76.87	78.96	75.06	71.65	66.08
	IMP-Loss (Ours)	82.09	**79.40**	**81.89**	**81.71**	**75.83**	**70.52**
	DIMP-Loss (Ours)	**82.67**	**79.53**	78.44	78.14	**75.83**	**70.04**
	- w/o diversity checker	81.35	77.94	77.68	77.62	74.72	69.34

- Both are consistently better than models only trained on small real-world data.

Checkers' Reaction on Noisy Data



Average weights of IMP-Loss for datapoints in Financial dataset

$$-\frac{1}{N}\sum_{i=1}^{N}\frac{\overbrace{\hat{P}'(y_i|\mathbf{x}_i)}^{\text{Quality Checker}}}{\underbrace{\hat{Q}(y_i|\mathbf{x}_i)}_{\text{Diversity Checker}}}\log \hat{P}(y_i|\mathbf{x}_i;\theta)$$

- Original: Original data
- Swap: Swapped ground true label (Low Quality)
- Duplicate: Duplicate datapoint twice in dataset (Low Diversity)
- Unrelated: Other benchmark (Low Quality)

Robust Performance on Noise Data

Dataset	Method	Financial		Tweet Irony		MRPC	
		Acc	F1	Acc	F1	Acc	F1
	GPT-3.5 few-shot	79.46	81.6	63.39	69.39	69.28	71.75
Small real world	CE-Loss (quality checker)	78.05	75.26	62.5	62.38	73.16	68.69
Noisy Data	CE-Loss	78.38	73.44	60.46	60.14	74.03	67.5
	Focal-Loss	78.55	74.97	62.11	61.12	74.72	69.59
	IMP-Loss (Ours)	81.6	78.24	**64.8**	**64.51**	76	70.46
	DIMP-Loss (Ours)	**82.59**	**80.28**	64.16	**64.09**	**76.58**	**71.32**

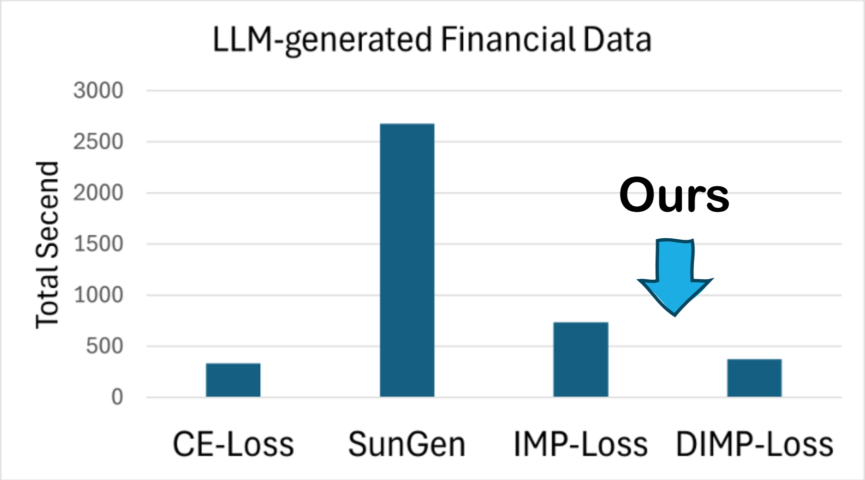
Noise Data: Randomly add duplicate, swapped label, and irrelevant data into large real-world dataset

→Our methods still help even the training set is messy

Computational Time

Method	Build QC	Build DC	Precalculate weights	Training	Total
CE-Loss	-	-	-	333.242s	333.242s
SunGen	-	-	-	2680s	2680s
IMP-Loss	8.824s	333.516s	57.695s	333.328s	733.363s
DIMP-Loss	8.824s	-	29.274s	333.426s	371.524s

- **IMP-Loss requires approximately twice the computational time compared to using CE-Loss**
- **DIMP-Loss just requires slightly higher than CE-Loss! → It is Time Efficient!**
- **Both need less time than a typical meta-learning approach, i.e., SunGen**



Total Time

Without Diversity Checkers

Dataset	Method	Financial		Tweet Irony		MRPC	
		Acc	F1	Acc	F1	Acc	F1
GPT-3.5 generated	CE-Loss	77.39	74.01	76.91	76.8	72	65.47
	Focal-Loss	79.29	75.32	74.87	74.82	72.17	62.77
	Hu et al.'s	71.7	61.93	71.42	70.18	67.13	50.08
	SunGen	80.45	76.87	78.96	75.06	71.65	66.08
	IMP-Loss (Ours)	82.09	**79.40**	**81.89**	**81.71**	**75.83**	**70.52**
	DIMP-Loss (Ours)	**82.67**	**79.53**	78.44	78.14	**75.83**	**70.04**
	- w/o diversity checker	81.35	77.94	77.68	77.62	74.72	69.34

- **Diversity checkers** are important!

$$-\frac{1}{N}\sum_{i=1}^{N}\frac{\overbrace{P'(y_i|\mathbf{x}_i)}^{\text{Quality Checker}}}{\underbrace{\hat{P}(y_i|\mathbf{x}_i;\theta_t)}_{\text{Diversity Checker}}}\log \hat{P}(y_i|\mathbf{x}_i;\theta_t)$$

Superior and Robust Accuracy across Epochs

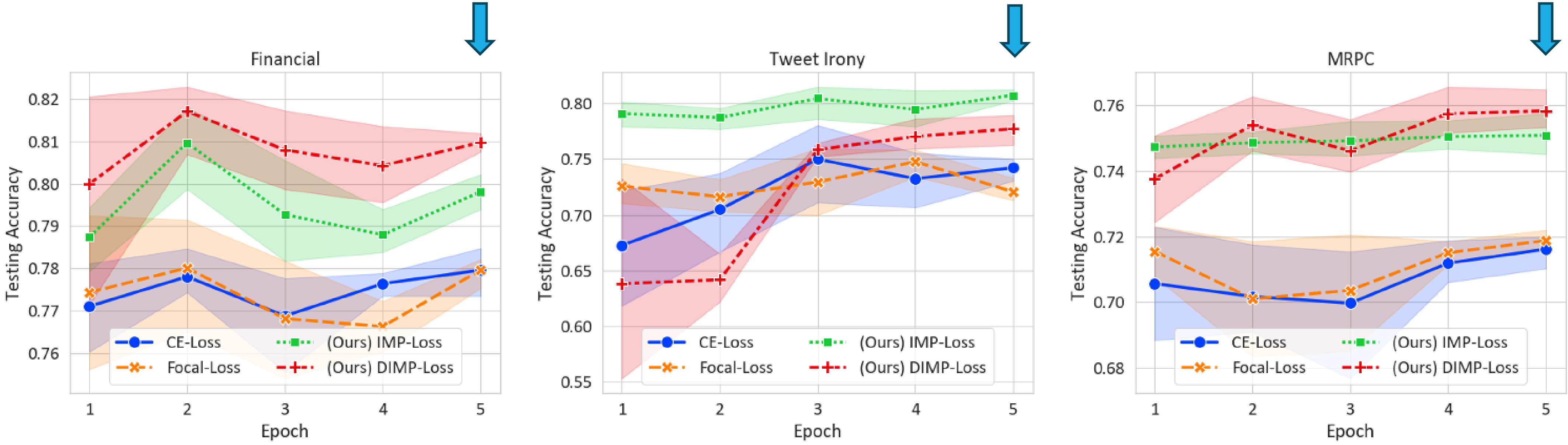
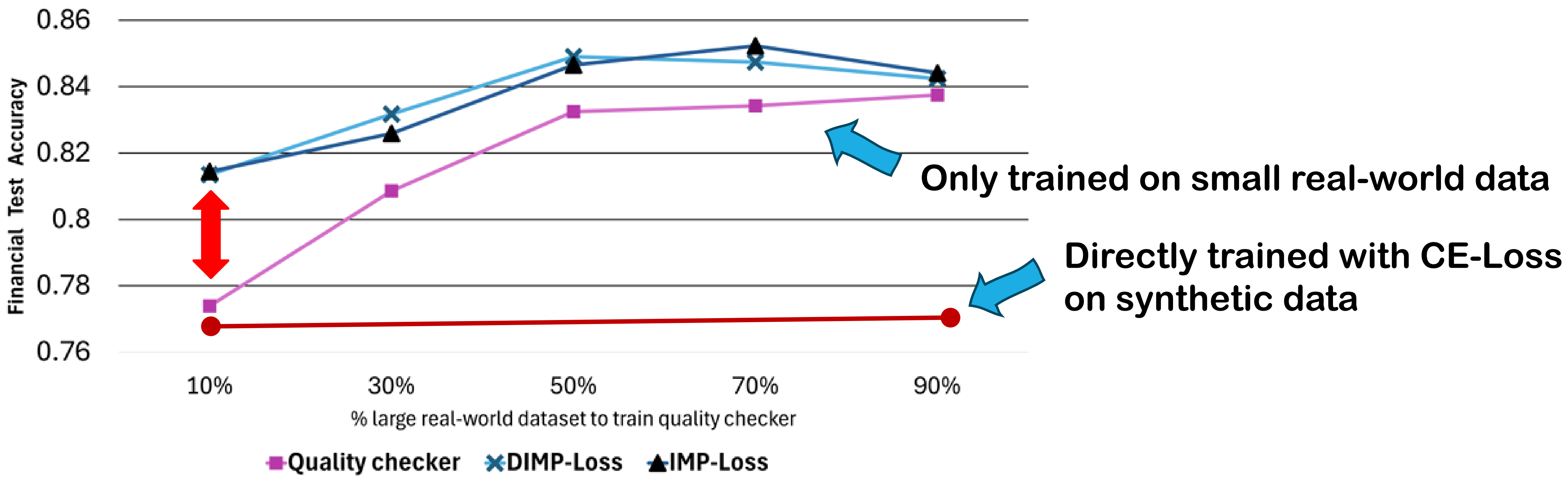


Figure 1: Training dynamics shows the testing accuracy over five epochs for benchmarks. This chart displays the minimum, maximum, and average accuracy observed across four runs with different random seeds, comparing our proposed methods with the standard CE-Loss and Focal-Loss.

Quality Checker is Data Size Efficient



Small data can still have large improvement!
→ data size efficient

Smaller Quality Checker on Larger Text-classifier

Model Size	Method	Financial	Tweet irony	MRPC
Base	Quality checker	78.05	62.5	73.16
Large	CE-Loss	80.45	78.83	74.2
	IMP-Loss (base DC)	80.94	74.23	75.36
	IMP-Loss (large DC)	81.93	78.83	76.41
	DIMP-Loss	**83.25**	**81.25**	**77.04**

Table 2: Accuracy of methods on benchmarks when training a larger model with smaller Quality Checkers. ”base DC” and ”large DC” denote smaller and larger Diversity Checkers, respectively. Bold entries highlight the top value of metrics within each dataset.

Small size Quality Checker still helps for DIMP-Loss

Conclusion

- We considered the issue of distribution misalignment and proposed IMP-Loss, DIMP-Loss, aiming to make model closer to the real-world distribution.
- Both loss robustly outperform on various data (LLM-generated, real-world, and noisy data)
- Training on DIMP-Loss is efficient of checker size, data requirement, and computational time!

Let's Prioritize high Quality and Diverse datapoints!