NoisyICL: A Little Noise in Model Parameters Calibrates In-context Learning

Anonymous ACL submission

Abstract

In-**C**ontext **L**earning (ICL) is suffering from unsatisfactory performance and under-calibration due to high prior bias and unfaithful confidence. Some previous works fine-tuned language models for better ICL performance with enormous datasets and computing costs. In this paper, we propose NoisyICL, simply perturbing the model parameters by random noises to strive for better performance and calibration. Our experiments on two models and 12 downstream datasets show that NoisyICL can help ICL produce more accurate predictions. Our further analysis indicates that NoisyICL enables the model to provide more fair predictions, and also with more faithful confidence. Therefore, we believe that NoisyICL is an effective calibration of ICL. Our experimental code is uploaded to Github[1].

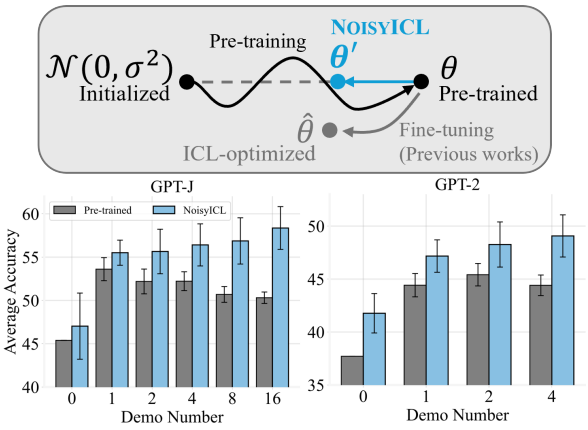

Figure 1: **Upper**: A sketch of NoisyICL: Unlike previous works which fine-tuned LMs towards ICL tasks, we perturb LMs by random noise sampled from the normal distribution $\mathcal{N}(0, \sigma^2)$ with intensity λ, then perform ICL. **Lower**: The average accuracy of ICL with and without NoisyICL w.r.t. the number of demos.

1 Introduction

Language **M**odels (LMs) have demonstrated the ability of **I**n-**C**ontext **L**earning (ICL), where LMs learn tasks from few-shot input-label demonstrations in the form of natural language, without explicit parameter updates (Dong et al., 2022).

Nevertheless, ICL still underperforms the end-to-end models (Mosbach et al., 2023). A recent study shows that vanilla LMs are biased towards the knowledge acquired during pre-training, which is deemed harmful to ICL (Fei et al., 2023). There has been some effort in fine-tuning or calibrating LMs towards ICL tasks (Min et al., 2022; Zhao et al., 2021; Wei et al., 2021; Fei et al., 2023; Lu et al., 2022), which focus on reducing the gap between the knowledge gained from pre-training and ICL tasks, producing significant improvements in the ICL performance. However, fine-tuning these enormous LMs on additional data incurs a considerably high computational cost.

Inspired by Wu et al. (2022), which shows the benefits of noise for fine-tuning end-to-end LMs, we hypothesize that introducing noise to the model parameters of pre-trained LMs can fit LMs towards ICL with a lower computational cost. In this paper, we propose NoisyICL, which simply adds noise to model parameters and then performs ICL on the noised models, as shown in Fig. 1, to investigate the benefit of noise for ICL.

As shown in Fig. 1 and Table 1, our experiments on two models and 12 classification datasets show that adding appropriate noise into model parameters improves the performance of ICL by around 10% with obviously lower computational cost.

To investigate the reason for such performance improvement, our further experiments hypothesize that NoisyICL acts as a model calibration. In detail, we find that: **1.** NoisyICL neutralizes prediction bias among general tokens and labels, and **2.** NoisyICL leads the model to produce more faithful confidence.

[1]Not available during anonymous review.

Table 1: Accuracy and Macro-F1 results ($mean_{std}$, $k = 4$). A better result is in green. λ: The optimal intensity of noise searched on the validation set, **Acc.**: Accuracy, **MF1**: Macro-F1, $\mathbf{ECE_1}$: the Expected Calibrated Error described in §3.3; **w/o**: Not using NOISYICL, **w/**: Using NOISYICL; Datasets abbreviation described in Appendix. A; **Val.**: the results on the validation sets, **Test**: the results on the testing sets.

	Dataset			PS	HS	SE'14R	SE'14L	RTE	MRPC	Ethos	FP	SST2	TEE	TES	TEH	Mean
GPT-J	λ			0.2	0.2	0.1	0.1	0.012	0.2	0.004	0.08	0.004	0.01	0.08	0.002	—
	Acc. (%)	Val.	w/o	$37.46_{1.36}$	$71.97_{1.19}$	$36.48_{0.84}$	$34.82_{0.21}$	$49.88_{1.03}$	$43.18_{1.48}$	$53.44_{0.56}$	$62.36_{0.17}$	$81.48_{0.97}$	$45.68_{0.48}$	$49.32_{1.07}$	$50.04_{1.68}$	51.34
			w/	$59.82_{5.82}$	$82.95_{1.36}$	$51.54_{7.44}$	$45.14_{4.06}$	$51.00_{1.35}$	$59.82_{6.15}$	$56.00_{0.89}$	$62.42_{0.43}$	$82.52_{1.03}$	$46.35_{0.53}$	$50.90_{1.32}$	$50.35_{1.06}$	61.23
		Test	w/o	$33.09_{1.76}$	$81.30_{0.79}$	$33.67_{1.31}$	$34.26_{1.33}$	$46.52_{1.11}$	$50.97_{1.43}$	$57.74_{1.38}$	$61.87_{0.16}$	$79.15_{1.17}$	$44.72_{0.62}$	$50.20_{0.85}$	$53.16_{1.07}$	52.22
			w/	$55.96_{3.43}$	$87.16_{5.29}$	$46.79_{4.47}$	$41.77_{3.80}$	$47.64_{1.27}$	$53.43_{5.23}$	$57.07_{1.53}$	$61.90_{0.30}$	$78.82_{0.75}$	$45.17_{0.57}$	$48.11_{1.38}$	$53.12_{1.10}$	56.41
	MF1 (%)	Val.	w/o	$24.11_{0.87}$	$26.75_{1.38}$	$32.53_{1.46}$	$28.62_{0.68}$	$47.37_{1.14}$	$43.18_{1.48}$	$53.24_{0.69}$	$25.91_{0.26}$	$81.24_{1.04}$	$25.65_{0.54}$	$32.87_{1.81}$	$49.01_{1.81}$	39.20
			w/	$20.80_{1.13}$	$24.22_{0.63}$	$46.88_{6.70}$	$43.25_{4.12}$	$48.61_{1.38}$	$48.38_{2.69}$	$55.89_{0.86}$	$27.81_{0.74}$	$82.33_{1.06}$	$26.31_{0.80}$	$38.19_{5.33}$	$49.37_{0.92}$	42.67
		Test	w/o	$21.86_{0.86}$	$26.61_{1.61}$	$30.71_{1.47}$	$33.17_{1.52}$	$44.94_{1.30}$	$49.56_{1.48}$	$57.24_{1.45}$	$26.59_{0.52}$	$77.44_{1.36}$	$23.57_{0.86}$	$34.23_{1.25}$	$53.16_{1.07}$	39.92
			w/	$22.87_{1.57}$	$24.22_{0.89}$	$43.88_{3.90}$	$42.17_{4.22}$	$45.93_{1.31}$	$46.51_{1.56}$	$56.46_{1.54}$	$27.24_{1.22}$	$77.11_{0.84}$	$23.78_{0.84}$	$32.68_{2.17}$	$53.11_{1.10}$	41.33
	ECE_1 (%, ↓)	Val.	w/o	$28.70_{1.78}$	$14.39_{1.13}$	$31.20_{0.69}$	$38.56_{0.36}$	$29.35_{0.95}$	$27.79_{1.50}$	$23.20_{0.62}$	$23.90_{0.17}$	$11.45_{0.93}$	$42.79_{0.36}$	$17.57_{0.81}$	$28.45_{1.23}$	26.44
			w/	$12.77_{2.94}$	$9.87_{2.56}$	$13.96_{7.26}$	$21.34_{6.75}$	$28.11_{0.95}$	$8.79_{3.41}$	$20.33_{0.99}$	$21.43_{3.83}$	$12.86_{1.06}$	$42.03_{0.88}$	$10.28_{5.37}$	$28.18_{1.37}$	19.16
		Test	w/o	$30.33_{1.42}$	$9.41_{0.78}$	$34.45_{1.39}$	$36.78_{1.56}$	$32.42_{1.10}$	$28.29_{1.47}$	$18.97_{1.45}$	$23.70_{0.49}$	$10.75_{1.02}$	$44.09_{0.79}$	$16.56_{0.56}$	$24.50_{0.99}$	25.85
			w/	$9.65_{2.83}$	$5.97_{3.45}$	$17.53_{5.87}$	$23.85_{5.43}$	$31.49_{1.45}$	$20.24_{6.56}$	$19.28_{1.91}$	$22.77_{3.17}$	$10.62_{0.84}$	$43.81_{0.76}$	$14.43_{2.40}$	$24.67_{0.92}$	20.36
GPT-2	λ			0.014	0.1	0.08	0.06	0.002	0.08	0.2	0.04	0.014	0.04	0.04	0.2	—
	Acc. (%)	Val.	w/o	$45.44_{1.50}$	$39.26_{0.95}$	$44.65_{0.51}$	$42.83_{0.93}$	$51.41_{0.76}$	$69.06_{0.19}$	$43.20_{0.41}$	$53.03_{1.35}$	$59.63_{0.28}$	$29.67_{0.77}$	$38.55_{1.43}$	$41.82_{0.51}$	46.54
			w/	$49.17_{1.43}$	$64.06_{3.41}$	$46.05_{0.66}$	$46.99_{1.77}$	$52.11_{1.50}$	$59.65_{3.26}$	$51.64_{1.67}$	$58.26_{2.02}$	$61.50_{1.23}$	$31.17_{0.94}$	$42.44_{1.86}$	$51.48_{4.38}$	51.21
		Test	w/o	$39.62_{1.33}$	$43.62_{1.17}$	$40.03_{1.26}$	$39.41_{1.29}$	$50.89_{1.04}$	$63.12_{0.38}$	$45.93_{0.51}$	$52.27_{0.81}$	$59.47_{1.41}$	$27.43_{0.59}$	$29.48_{1.29}$	$41.66_{0.57}$	44.41
			w/	$40.95_{1.42}$	$67.25_{4.00}$	$45.28_{3.50}$	$36.90_{0.18}$	$51.01_{0.91}$	$61.20_{0.68}$	$50.90_{2.63}$	$56.19_{1.18}$	$61.54_{1.51}$	$33.38_{1.16}$	$34.01_{4.15}$	$50.28_{2.44}$	49.07
	MF1 (%)	Val.	w/o	$25.62_{1.81}$	$17.88_{0.37}$	$37.69_{0.08}$	$37.71_{0.88}$	$50.82_{0.67}$	$41.45_{0.49}$	$34.36_{0.41}$	$35.02_{2.00}$	$57.74_{0.42}$	$20.33_{0.49}$	$33.74_{1.86}$	$30.98_{0.50}$	35.28
			w	$26.05_{1.59}$	$24.20_{0.85}$	$32.40_{0.25}$	$30.78_{1.69}$	$51.57_{1.59}$	$48.90_{1.11}$	$49.38_{0.17}$	$35.35_{1.00}$	$61.28_{1.18}$	$22.14_{0.91}$	$34.46_{1.13}$	$46.20_{3.91}$	38.56
		Test	w/o	$24.70_{1.17}$	$18.24_{0.52}$	$35.78_{1.49}$	$38.79_{1.25}$	$49.37_{1.31}$	$39.90_{0.94}$	$35.56_{0.75}$	$35.78_{1.11}$	$59.46_{1.42}$	$18.89_{0.62}$	$29.36_{1.29}$	$34.43_{0.79}$	35.02
			w/	$24.49_{1.20}$	$23.74_{1.10}$	$34.03_{1.20}$	$27.41_{1.97}$	$49.56_{0.99}$	$41.46_{1.26}$	$47.25_{3.97}$	$36.67_{2.14}$	$61.15_{1.50}$	$24.06_{1.39}$	$31.48_{3.32}$	$49.05_{1.35}$	37.52
	ECE_1 (%, ↓)	Val.	w/o	$5.94_{0.81}$	$36.18_{0.96}$	$14.79_{0.92}$	$15.51_{1.25}$	$30.35_{1.12}$	$19.39_{0.40}$	$47.50_{0.37}$	$8.33_{1.12}$	$2.10_{0.88}$	$30.85_{1.06}$	$14.17_{1.04}$	$51.18_{0.62}$	23.02
			w/	$5.74_{1.40}$	$11.64_{2.21}$	$16.36_{4.41}$	$17.55_{1.50}$	$29.51_{1.54}$	$21.23_{1.51}$	$24.01_{0.97}$	$8.20_{1.78}$	$2.99_{1.10}$	$30.92_{1.99}$	$13.52_{1.18}$	$23.93_{5.48}$	17.13
		Test	w/o	$9.29_{1.37}$	$28.14_{1.21}$	$17.90_{1.04}$	$15.04_{1.29}$	$31.42_{1.13}$	$23.02_{0.54}$	$45.21_{0.78}$	$8.00_{0.92}$	$1.89_{0.55}$	$32.76_{0.80}$	$22.30_{1.33}$	$48.69_{0.51}$	23.64
			w/	$8.16_{1.12}$	$9.01_{2.57}$	$18.69_{1.81}$	$25.35_{2.98}$	$31.07_{1.22}$	$26.06_{1.64}$	$25.52_{4.12}$	$6.88_{1.71}$	$3.12_{0.80}$	$29.73_{2.28}$	$19.15_{4.69}$	$21.63_{2.11}$	18.70

Our contribution can be summarized as:

- We propose NOISYICL, which simply adds noise into LMs and then executes ICL (§2). Our experiment shows that NOISYICL obtains a better ICL performance (§3.2).
- We show that adding noise effectively calibrates LMs to reduce prediction bias and unfaithful confidence in ICL (§3.3).

2 NOISYICL

Here we introduce the basic form of ICL and our perturbation method named NOISYICL.

In-context Learning. Suppose a supervised classification dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, where x_i is an input, and $y_i \in \mathbb{U}$ is the corresponding label in a label space $\mathbb{U}$. For each query x_q to be predicted, we sample a demo sequence $G = \{(x_{a_j}, y_{a_j})\}_{j=1}^k$ from $\mathcal{D}$, where k is the number of demos, and construct a prompt input in natural language form $s = f(G, x_q)$ with a pattern f. Then, we input s into the LM $P_\theta(\cdot)$ with parameters θ and get an output token distribution $P_\theta(\cdot|s)$. We choose the label token l with the maximum probability **among the label space** as the prediction $\hat{y}_q$, that is:

$$\hat{y}_q = \underset{l \in \mathbb{U}}{\operatorname{argmax}}\, P_\theta(l|s). \quad (1)$$

Notice that we only construct prompts to drive the model to predict answers generatively, without any parameter updates.

NOISYICL. For each parameter matrix θ_i in the LM $P_\theta(\cdot)$ used for ICL, we simply do an interpolation between θ_i and a noise matrix sampled from an isotropic normal distribution $\mathcal{N}(0, \sigma^2)$ with intensity λ, that is:

$$\theta_i' = (1 - \lambda)\theta_i + \lambda\mathcal{N}(0, \sigma^2), \quad (2)$$

where λ and σ are model or task-specific hyperparameters. Then we perform the aforementioned ICL with the interpolated LM $P_{\theta'}(\cdot)$.

3 Experiments and Results

We investigate the effectiveness and calibration abilities of NOISYICL. We find that NOISYICL can improve ICL performance by an average of 10% (§3.2) and effectively calibrate the model's prediction bias and unfaithful confidence (§3.3).

3.1 Settings

Data. We use 12 commonly used classification datasets in our experiments. Since some of the datasets do not provide valid splitting, we randomly split these datasets into validation sets and test sets (see Appendix A for further details).

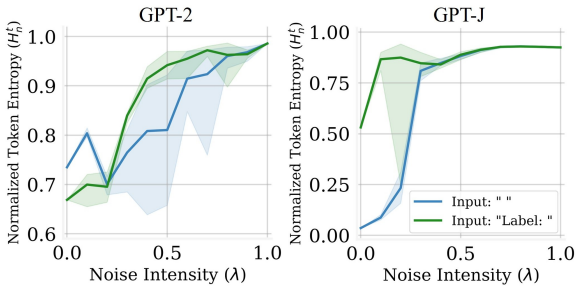

Figure 2: The correlation between the normalized token entropy H_n^t and the noise intensity λ with empty inputs. When the noise gets stronger, the H_n^t becomes higher, which indicates a fairer output.

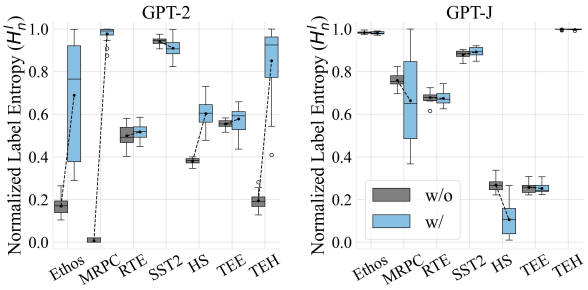

Figure 3: The normalized label entropy H_n^l on both models and 7 datasets with and without appropriate-noised NoisyICL. In most cases, the H_n^l with NoisyICL (**w/**) is greater than without NoisyICL (**w/o**).

Models. We use GPT-2 (parameters: 137M) (Radford et al., 2019) and GPT-J (parameters: 6053M) (Wang and Komatsuzaki, 2021). The model checkpoints are loaded from huggingface[2].

Hyperparameters. We fix σ, the standard deviation of the noise distribution, to 0.02, which is the same as the initialization distribution of both models. We believe the noise intensity λ should vary w.r.t. datasets and models. Therefore, we determine the most suitable λ by a simple search method for each dataset and model on the validation set (details in Appendix D). The selected intensities are shown in Table 1, which are concentrated in $(0, 0.3]$. Moreover, we confirm that the λ keeps relatively stable w.r.t. k given a dataset and model (Appendix D).

Other details. We default to use four demos and a simple prompt template as shown in Appendix B. Every labeled data is treated as the test query once, and for each query, we do 2 tries. Each experiment is repeated 10 times with different noise matrices.

3.2 NoisyICL Improves ICL Performance

We show accuracy and macro-F1 on the 12 downstream datasets with and without NoisyICL on the optimal λ in Table 1, and averaged results in Fig. 1 (see Appendix C for results of other numbers of demos (k)).

The results show that NoisyICL produces an average performance improvement of around 10%. This phenomenon preliminarily confirms our hypothesis: NoisyICL fits the pre-trained LMs towards ICL.

However, such gains vary depending on the dataset and model. In some combinations of datasets and models, significant performance improvement cannot be observed. We speculate the reason is that NoisyICL does not provide new knowledge from new training examples, and these datasets are too hard for the model intrinsically no matter whether NoisyICL is used.

[2] huggingface.co/gpt2, and huggingface.co/EleutherAI/gpt-j-6b

3.3 NoisyICL Is A Calibration

This section finds that NoisyICL conducts the following aspects of calibrations:

- **Lower prediction bias.** When no valid query is given, the predicted labels should have balanced frequencies. However, predictions conducted by under-calibrated LMs usually have significant bias, which is harmful to ICL (Fei et al., 2023; Zhao et al., 2021; Wei et al., 2023; Lu et al., 2022). Eliminating these biases can be seen as an aspect of calibration.

- **More faithful confidence.** In classification tasks, the outputted probability of the predicted label is called confidence. Suitable confidence should faithfully reflect the accuracy of outputs, that is, a prediction with greater accuracy should be assigned with greater confidence (Corbière et al., 2019). It has been proven that faithful confidence improves the stability of the model (Guo et al., 2017; Grabinski et al., 2022). Making the confidence more faithful is also an aspect of calibration (Guo et al., 2017; Tian et al., 2023).

Specifically, we find that NoisyICL induces the model to predict labels with more fair and faithful confidence.

NoisyICL reduces prediction bias. We investigate the following two types of prediction bias. (See Appendix E for further calculation details.)

(1) Intrinsic token bias. LMs are pre-trained with natural corpus, where different tokens have various frequencies. LMs learn such frequencies and act as biases among different tokens in prediction, which are deemed harmful to ICL (Fei et al., 2023). We believe that such token bias can be reduced by NOISYICL.

To quantify this, we calculate normalized token entropy H_n^t: the entropy of the predicted probability distribution among the whole vocabulary given an empty input (we use " ", and "Label: "), w.r.t. various noise intensities.

The results are shown in Fig. 2. A clear positive correlation between the noise intensity and H_n^t can be observed, meaning the model gives a fairer output when more noise is given.

(2) Overall label bias. In classification tasks, when no query is given, equal probabilities should be assigned to labels fairly, which benefits ICL (Lu et al., 2022). We believe that NOISYICL promotes such fairness.

To quantify this, we use normalized label entropy H_n^l: the entropy of the predicted probability distribution among the label space given some demos and an empty query. Notice that it is natural for a model to predict a "neutral" label when no query is given, so label bias on a dataset with a "neutral" label cannot be fairly measured with H_n^l. Therefore, we ignore these datasets with "neutral" labels (details in Appendix A).

The results are shown in Fig. 3, indicating that in most cases, NOISYICL calibrates the overall bias, especially on GPT-2.

NOISYICL promotes faithful confidence. The **E**xpected **C**alibration **Err**or (ECE_p) (Naeini et al., 2015) is a widely-used indicator for the faithfulness of confidence in classification tasks:

$$ECE_p = \mathbb{E}(|\max(\hat{z}) - \mathbb{E}_{y=\underset{i}{\operatorname{argmax}}\,\hat{z}_i}(1)|^p)^{\frac{1}{p}}, \quad (3)$$

where $\hat{z}$ is the predicted probability vector on the label space by a classification model, the final prediction $(\operatorname{argmax}_i \hat{z}_i)$ can be obtained with a confidence $(\max \hat{z})$, and the ground-truth label is y.

Let $p = 1$, we use the ECE_1 to investigate the faithfulness of the ICL output. The details of the calculation are shown in Appendix F. A lower ECE_1 means more faithful confidence, that is, the confidence becomes a more accurate prediction of accuracy (Corbière et al., 2019). Fig. 4 shows a case study of GPT-2 on the hate_speech18 dataset, indicating that the prediction is more faithful with NOISYICL (full visualized data are in Appendix G). We show the results with and without appropriate-noised NOISYICL for ECE_1 on the 12 datasets in Table 1.

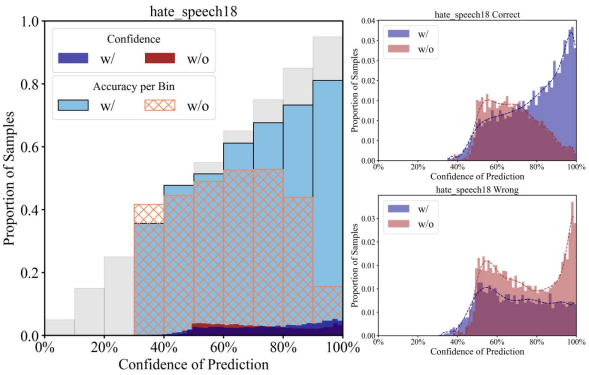

Figure 4: **Left**: Reliability diagrams (sparse bars) and global confidence distribution (dense bars) of GPT-2 on hate_speech18 with (**w/**, $ECE_1 = 9.01\%$) and without (**w/o**, $ECE_1 = 28.14\%$) NOISYICL. The predictions are divided into bins according to confidence, and we visualize the accuracy of each bin as a histogram. The grey bars are ideal, which is closer to the one with NOISYICL. **Right upper**: Confidence distribution on correct predictions. Relatively right-shifted with NOISYICL. **Right lower**: Confidence distribution on wrong predictions. Relatively left-shifted with NOISYICL. ($k = 4$)

In most cases, the ECE_1 is reduced by around 25% by NOISYICL, which means the confidence is more faithful with NOISYICL. This suggests that NOISYICL can drive the model to produce more faithful confidence, that is, less over-confidence in wrong predictions, and less under-confidence in correct predictions, which suggests that NOISYICL solves the confidence calibration.

Based on the above two investigations, we believe that NOISYICL is an effective calibration for the ICL scenario. This can be a reasonable explanation for the observed performance improvement in §3.2.

4 Conclusion

In this paper, we propose NOISYICL, which simply adds random noise to the parameters of LMs to fit them from pre-trained knowledge to ICL. We show that NOISYICL can improve the ICL performance and also calibrate the model for fairer outputs and more faithful confidence.

5 Limitations

As mentioned before, unlike the fine-tuning on additional ICL-style datasets (Min et al., 2022; Zhao et al., 2021; Wei et al., 2021), NOISYICL does not provide new knowledge for the model, so the calibrated model can not discover tasks that are not potentially included in the pre-training data (Gu et al., 2023). Moreover, a naive search for the best noise intensity is not efficient and satisfactory. An effective detection of λ should be developed. Currently, we can only confirm that for one dataset and model, the λ keeps relatively stable, especially when a large k is given (Appendix. D).

Future Works. Besides fixing the limits, future works can focus on where and how the noise should be introduced. In Transformer-based models, different layers have different abilities (Wang et al., 2023; Jawahar et al., 2019; Kobayashi et al., 2023). So, treating these layers differently may be an effective improvement of NOISYICL. An isotropic normal distribution is too simple, noise sampling methods should be discussed.

Moreover, adding noise to model parameters can be an erasing of pre-training (Ilharco et al., 2022), so, the search for λ is the search for the best checkpoint of pre-training. With these checkpoints, we can determine (Han et al., 2023; Han and Tsvetkov, 2022) which data is disadvantageous to ICL, and what knowledge is essential to ICL, to better reveal the essence of ICL.

References

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2009. The fifth pascal recognizing textual entailment challenge. *TAC*, 7:8.

Charles Corbière, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. 2019. Addressing failure prediction by learning model confidence. *Advances in Neural Information Processing Systems*, 32.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine learning challenges workshop*, pages 177–190. Springer.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate Speech Dataset from a White Supremacy Forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium. Association for Computational Linguistics.

Bill Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Third International Workshop on Paraphrasing (IWP2005)*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

Yu Fei, Yifan Hou, Zeming Chen, and Antoine Bosselut. 2023. Mitigating label biases for in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14014–14031, Toronto, Canada. Association for Computational Linguistics.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and William B Dolan. 2007. The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing*, pages 1–9.

Julia Grabinski, Paul Gavrikov, Janis Keuper, and Margret Keuper. 2022. Robust models are less overconfident. *Advances in Neural Information Processing Systems*, 35:39059–39075.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2023. Pre-training to learn in context. *arXiv preprint arXiv:2305.09137*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

R Bar Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The second pascal recognising textual entailment challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, volume 7, pages 785–794.

Xiaochuang Han, Daniel Simig, Todor Mihaylov, Yulia Tsvetkov, Asli Celikyilmaz, and Tianlu Wang. 2023. Understanding in-context learning via supportive pretraining data. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12660–12673.

Xiaochuang Han and Yulia Tsvetkov. 2022. Orca: Interpreting prompted language models via locating supporting data evidence in the ocean of pretraining data. *arXiv preprint arXiv:2205.12600*.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2022. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2023. Transformer language models handle word frequency in prediction head. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4523–4535, Toronto, Canada. Association for Computational Linguistics.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098.

P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. Metaicl: Learning to learn in context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2791–2809.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2020. Ethos: an online hate speech detection dataset.

Marius Mosbach, Tiago Pimentel, Shauli Ravfogel, Dietrich Klakow, and Yanai Elazar. 2023. Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation. *arXiv preprint arXiv:2305.16938*.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Jane Pan, Tianyu Gao, Howard Chen, and Danqi Chen. 2023. What in-context learning" learns" in-context: Disentangling task recognition and task learning. *arXiv preprint arXiv:2305.09731*.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pages 502–518.

Emily Sheng and David Uthus. 2020. Investigating societal biases in a poetry composition system. *arXiv preprint arXiv:2011.02686*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. `https://github.com/kingoflolz/mesh-transformer-jax`.

Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023. Label words are anchors: An information flow perspective for understanding in-context learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9840–9855, Singapore. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jerry Wei, Le Hou, Andrew Lampinen, Xiangning Chen, Da Huang, Yi Tay, Xinyun Chen, Yifeng Lu, Denny

Zhou, Tengyu Ma, et al. 2023. Symbol tuning improves in-context learning in language models. *arXiv preprint arXiv:2305.08298*.

Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2022. Noisytune: A little noise can help you finetune pretrained language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 680–685.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR.

A Datasets

The datasets used in this paper are shown in Table 2.

*To construct inputs of appropriate length, we remove data with a length exceeding 500 from the Ethos, and the number of the remaining data is 980.

Dataset Splitting. For each dataset, we first shuffle it with random seed 42. Then, we choose the 512 data at the tail as the testing data, and the 512 data at the head (if the total number of data is less than 1024, we choose the rest of the data.) as the validation data.

B Prompt Patterns

In this paper, we use a minimum prompt template. For each task category, we design a template as shown below.

For single-sentence classification datasets (x, y), we use:

```
Input: <x>, Label: <y> \n
...
Input: <x>, Label:
```

For aspect-based sentiment classification datasets $((x, a), y)$, we use:

```
Input: <x>, Aspect: <a>, Label: <y> \n
...
Input: <x>, Aspect: <a>, Label:
```

For double-sentence classification datasets $((x_1, x_2), y)$. we use:

```
Input: <x1>, Text 2: <x2>, Label: <y> \n
...
Input: <x1>, Text 2: <x2>, Label:
```

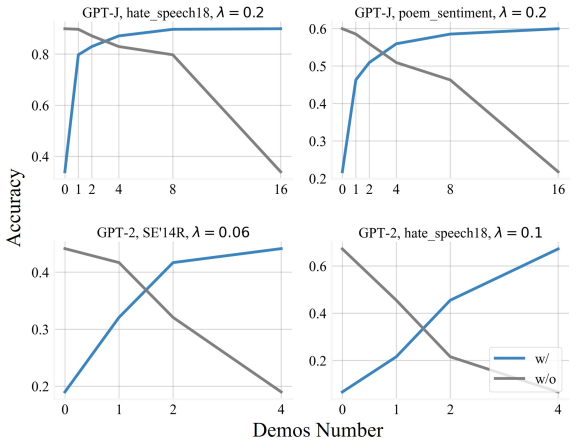

Figure 5: The relationship between demos quantity and accuracy in some cases. NOISYICL can make the model learn from the demos correctly.

Input: This is delicious!	Label: Positive
Input: I don't like it.	Label: Negative
Input: Boring.	Label: Negative
Input: It is fantastic!	Label: Positive
Input:	Label:

Figure 6: An example of inputs in the normalized label entropy calculation.

C Complete Results: Accuracy, Macro-F1, and ECE_1 With Various Demo Numbers (k)

We use various numbers of demos in the experiments shown in the §3.2.

The results of zero-shot ($k = 0$) are shown in Table 3.

The results of 1-shot ($k = 1$) are shown in Table 4.

The results of 2-shot ($k = 2$) are shown in Table 5.

The results of 8-shot ($k = 8$) are shown in Table 6.

The results of 16-shot ($k = 16$) are shown in Table 7.

Notice that in Table 6 and Table 7, some experiments are not feasible due to the length of sequences.

In these results, NOISYICL still outperforms the baseline, which proves that NOISYICL is generally applicable with various k.

D Searching and Stability of λ

We select a set of candidates of λ as {0, 0.002, 0.004, 0.006, 0.008, 0.01, 0.012, 0.014, 0.016, 0.018, 0.02, 0.04, 0.06, 0.08, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0}, test the performance of

Table 2: Datasets and Abbreviations used in this paper. **Abbr.**: Abbreviation, **Data#**: The total number of data, **Label#**: The number of classes, **Neutral?**: Does the dataset have a neutral label?, **Major.** $\sim$ **Minor.(%)**: The proportion of majority labels $\sim$ the proportion of minority labels.

Dataset	Abbr.	Data#	Label#	Neutral?	Major. $\sim$ Minor.(%)
single-sentence classification:					
poem_sentiment (Sheng and Uthus, 2020)	PS	1101	4	✓	62.2 $\sim$ 5.5
hate_speech18 (de Gibert et al., 2018)	HS	10944	4	×	86.9 $\sim$ 1.5
Ethos* (binary) (Mollas et al., 2020)	—	980	2	×	56.6 $\sim$ 43.4
financial_phrasebank (all agree) (Malo et al., 2014)	FP	2264	3	✓	61.4 $\sim$ 13.4
GLUE-SST2 (Wang et al., 2018; Socher et al., 2013)	SST2	68221	2	×	55.8 $\sim$ 44.2
tweet_eval_emotion (Mohammad et al., 2018)	TEE	5052	4	×	43 $\sim$ 9
tweet_eval_sentiment (Rosenthal et al., 2017)	TES	59899	3	✓	45.3 $\sim$ 15.5
tweet_eval_hate (Basile et al., 2019)	TEH	12970	2	×	58 $\sim$ 42
aspect-based sentiment classification:					
SemEval 2014-Task 4 Restaurants (Pontiki et al., 2014)	SE'14R	4722	3	✓	61.2 $\sim$ 17.6
SemEval 2014-Task 4 Laptops (Pontiki et al., 2014)	SE'14L	2951	3	✓	45.3 $\sim$ 15.5
double-sentence classification:					
GLUE-RTE (Wang et al., 2018; Dagan et al., 2005) (Haim et al., 2006; Giampiccolo et al., 2007; Bentivogli et al., 2009)	RTE	2767	2	×	50.2 $\sim$ 49.8
GLUE-MRPC (Wang et al., 2018; Dolan and Brockett, 2005)	MRPC	4076	2	×	67.4 $\sim$ 32.6

NOISYICL with this set of λ, and select the one with the best accuracy as the optimal λ.

Moreover, we find that given a dataset and model, λ remains relatively stable w.r.t the demo number k, especially when k is large. As shown in Fig. 7 and Fig. 8, we calculate $|\lambda_i - \lambda_j|$, the distance between the optimal λ of different demo numbers as the heatmap, and the normalized remaining range $R(\lambda_{i:})/\max(\lambda)$, where the $R(\lambda_{i:})$ is the range of all the optimal λ_c where the demo number $c \geq i$, and the $\max(\lambda)$ is the maximum of all the optimal λ on the dataset. It characterizes the dispersion of λ when the demo number is greater than the given one i.

The results show that while the k is increasing, the distance and the normalized remaining range quickly zero out, which supports our hypothesis: NOISYICL is a bridge from the pre-training to ICL. For that during this process, the demo number should have a minimal impact.

E Calculation of H_n^t and H_n^l

Normalized Token Entropy. We calculate the normalized token entropy H_n^t of the predicted probability distribution of tokens with an empty input as:

$$H_n^t = -\frac{\sum_{l\in\mathbb{V}} P_\theta(l|x_\varnothing)\ln P_\theta(l|x_\varnothing)}{\ln|\mathbb{V}|}, \quad (4)$$

where the $P_\theta(\cdot)$ is an LM with a vocabulary space $\mathbb{V}$ and size $|\mathbb{V}|$, the $x_\varnothing$ is an empty input, such as " ", or "Label: ". The model outputs a global probability distribution of tokens when the $x_\varnothing$ is given, and we calculate the normalized entropy on the distribution.

Normalized Label Entropy. We calculate the normalized label entropy H_n^l of the model prediction probability distribution on the label space with 4 demos and an empty queue from a dataset (Fig. 6 is an example) as:

$$H_n^l = -\frac{\sum_{l\in\mathbb{U}} r_{P,\theta}(l|x_*)\ln r_{P,\theta}(l|x_*)}{\ln|\mathbb{U}|}, \quad (5)$$

where the $r_{P,\theta}(l|x_*)$ is the predicted frequency of label l given the aforementioned query-less only-demos input x_*. We use 512 tries to estimate this frequency. The $\mathbb{U}$ and $|\mathbb{U}|$ are the label space and label space size, respectively.

F Calculation of ECE_1

Given a prediction set $\mathcal{Y} = \{(\hat{y}_i, z_i, y_i)\}_{i=1}^n$ of predictions $\hat{y}_i$, predicted confidence $z_i \in [0,1]$, and supervised label y_i, in the calculation of ECE_1, we first divide the confidence space $(0,1)$ into $m = 10$ equal bins $B_j = [0.1(j-1), 0.1j]_{j=1}^m$ according to the z_i as shown in Fig. 4. The amount of prediction in each bin is denoted as $|B_j|$. For each bin, we calculate the accuracy $\alpha_j = \frac{1}{|B_j|}\sum_{\hat{y}_i\in B_j, \hat{y}_i=y_1} 1$ of predictions in the jth bin, and assign a standard accuracy based on the average of confidence in the bin, that is, $\alpha_j^s = \frac{1}{|B_j|}\sum_{z_i\in B_j} z_i$. The ECE_1 can be described as:

$$ECE_1 = \sum_{j=1}^m \frac{|B_j|}{n}|\alpha_j - \alpha_j^s|, \quad (6)$$

Table 3: Accuracy and Macro-F1 results (%, $mean_{std}$, $k = 0$). A better result is in green. Notation is the same with the Tabel 1.

	Dataset			PS	HS	SE'14R	SE'14L	RTE	MRPC	Ethos	FP	SST2	TEE	TES	TEH	Mean
GPT-J	λ			0.1	0.1	0.1	0.1	0.06	0.02	0.004	0.02	0.1	0.2	0.008	0.3	—
	Acc. (%)	Val.	w/o	$16.60_{0.00}$	$37.30_{0.00}$	$33.98_{0.00}$	$38.87_{0.00}$	$52.54_{0.00}$	$65.82_{0.00}$	$51.07_{0.00}$	$61.13_{0.00}$	$57.42_{0.00}$	$39.84_{0.00}$	$49.22_{0.00}$	$43.36_{0.00}$	45.60
			w/	$21.88_{6.75}$	$50.70_{4.33}$	$40.74_{10.27}$	$45.94_{2.72}$	$53.40_{0.55}$	$67.77_{0.55}$	$51.07_{0.19}$	$61.37_{0.60}$	$60.98_{3.87}$	$39.57_{3.50}$	$49.53_{0.20}$	$54.14_{5.91}$	49.76
		Test	w/o	$17.38_{0.00}$	$29.88_{0.00}$	$28.91_{0.00}$	$39.45_{0.00}$	$49.80_{0.00}$	$59.18_{0.00}$	$54.10_{0.00}$	$61.91_{0.00}$	$66.41_{0.00}$	$41.60_{0.00}$	$49.22_{0.00}$	$46.88_{0.00}$	45.39
			w/	$22.97_{7.52}$	$38.30_{10.89}$	$31.25_{5.79}$	$40.84_{3.85}$	$49.77_{1.46}$	$59.38_{0.79}$	$54.36_{0.15}$	$61.37_{0.65}$	$65.96_{2.27}$	$41.15_{3.79}$	$49.14_{0.26}$	$49.86_{8.37}$	47.03
	MF1 (%)	Val.	w/o	$11.32_{0.00}$	$20.81_{0.00}$	$35.19_{0.00}$	$36.56_{0.00}$	$46.42_{0.00}$	$48.13_{0.00}$	$48.58_{0.00}$	$36.13_{0.00}$	$56.98_{0.00}$	$15.99_{0.00}$	$29.81_{0.00}$	$33.40_{0.00}$	34.94
			w/	$14.87_{2.10}$	$25.82_{2.65}$	$39.90_{8.62}$	$41.36_{3.98}$	$50.10_{1.31}$	$49.08_{0.54}$	$48.58_{0.25}$	$37.35_{0.74}$	$59.80_{4.50}$	$25.30_{2.60}$	$30.36_{0.60}$	$39.66_{6.23}$	38.62
		Test	w/o	$12.70_{0.00}$	$16.02_{0.00}$	$27.01_{0.00}$	$39.38_{0.00}$	$45.56_{0.00}$	$40.48_{0.00}$	$50.99_{0.00}$	$34.91_{0.00}$	$64.45_{0.00}$	$16.26_{0.00}$	$29.81_{0.00}$	$42.58_{0.00}$	35.85
			w/	$16.93_{4.12}$	$17.92_{3.40}$	$30.20_{7.55}$	$40.50_{4.43}$	$49.06_{2.03}$	$39.96_{0.59}$	$51.20_{0.12}$	$35.56_{1.38}$	$61.74_{5.39}$	$22.54_{5.12}$	$29.69_{0.59}$	$39.70_{8.87}$	36.25
	ECE_1 (%, ↓)	Val.	w/o	$36.56_{0.00}$	$22.03_{0.00}$	$28.11_{0.00}$	$17.80_{0.00}$	$29.83_{0.00}$	$7.24_{0.00}$	$28.08_{0.00}$	$13.45_{0.00}$	$5.45_{0.00}$	$48.09_{0.00}$	$22.10_{0.00}$	$46.21_{0.00}$	25.41
			w/	$22.52_{6.40}$	$16.20_{3.75}$	$16.77_{9.48}$	$8.39_{3.82}$	$23.90_{1.51}$	$9.20_{0.54}$	$28.12_{0.24}$	$12.75_{0.37}$	$4.22_{2.43}$	$18.66_{2.88}$	$22.53_{0.19}$	$21.99_{12.15}$	17.10
		Test	w/o	$35.33_{0.00}$	$26.92_{0.00}$	$33.41_{0.00}$	$16.92_{0.00}$	$31.46_{0.00}$	$13.80_{0.00}$	$24.56_{0.00}$	$15.09_{0.00}$	$3.49_{0.00}$	$46.49_{0.00}$	$22.10_{0.00}$	$36.32_{0.00}$	25.49
			w/	$24.86_{6.34}$	$23.67_{6.86}$	$26.19_{10.00}$	$12.16_{5.22}$	$24.43_{2.49}$	$13.65_{0.60}$	$24.49_{0.13}$	$13.34_{0.90}$	$5.91_{2.43}$	$25.21_{14.36}$	$22.37_{0.36}$	$24.95_{16.80}$	20.10
GPT-2	λ			0.04	0.002	0.1	0.1	0.012	0.06	0.1	0.06	0.02	0.02	0.002	0.1	—
	Acc. (%)	Val.	w/o	$4.30_{0.00}$	$15.43_{0.00}$	$23.44_{0.00}$	$39.84_{0.00}$	$49.41_{0.00}$	$69.53_{0.00}$	$44.23_{0.00}$	$43.55_{0.00}$	$60.55_{0.00}$	$32.03_{0.00}$	$46.68_{0.00}$	$41.99_{0.00}$	39.25
			w/	$5.00_{0.50}$	$15.23_{0.00}$	$40.90_{15.99}$	$40.35_{5.80}$	$50.82_{0.42}$	$63.52_{2.09}$	$55.98_{1.93}$	$58.13_{2.45}$	$63.87_{0.96}$	$38.71_{0.57}$	$46.41_{0.16}$	$49.38_{3.75}$	44.02
		Test	w/o	$5.86_{0.00}$	$10.74_{0.00}$	$22.27_{0.00}$	$35.35_{0.00}$	$48.63_{0.00}$	$63.48_{0.00}$	$46.88_{0.00}$	$45.51_{0.00}$	$64.84_{0.00}$	$31.25_{0.00}$	$37.50_{0.00}$	$40.23_{0.00}$	37.71
			w/	$6.52_{0.61}$	$9.39_{0.20}$	$41.89_{11.21}$	$37.19_{1.48}$	$47.95_{0.57}$	$58.81_{1.26}$	$53.89_{1.52}$	$46.45_{0.29}$	$59.39_{3.25}$	$41.48_{0.61}$	$37.21_{0.22}$	$61.04_{1.07}$	41.77
	MF1 (%)	Val.	w/o	$4.70_{0.00}$	$9.58_{0.00}$	$24.11_{0.00}$	$36.57_{0.00}$	$47.91_{0.00}$	$41.63_{0.00}$	$34.16_{0.00}$	$26.03_{0.00}$	$54.25_{0.00}$	$22.91_{0.00}$	$46.68_{0.00}$	$29.85_{0.00}$	31.53
			w	$5.25_{0.57}$	$9.46_{0.01}$	$27.57_{5.24}$	$31.40_{3.63}$	$50.31_{0.61}$	$50.06_{1.78}$	$51.44_{2.19}$	$33.29_{0.25}$	$61.13_{2.01}$	$27.20_{0.36}$	$38.58_{0.10}$	$41.68_{2.24}$	35.61
		Test	w/o	$6.08_{0.00}$	$6.50_{0.00}$	$20.45_{0.00}$	$34.84_{0.00}$	$46.00_{0.00}$	$38.83_{0.00}$	$35.37_{0.00}$	$28.04_{0.00}$	$64.67_{0.00}$	$22.73_{0.00}$	$32.56_{0.00}$	$29.22_{0.00}$	30.44
			w/	$5.68_{1.01}$	$5.78_{0.11}$	$27.89_{5.96}$	$27.69_{3.78}$	$46.66_{0.47}$	$49.68_{0.71}$	$48.89_{3.86}$	$29.66_{0.32}$	$58.07_{4.31}$	$29.69_{0.45}$	$32.17_{0.22}$	$54.16_{2.63}$	34.67
	ECE_1 (%, ↓)	Val.	w/o	$53.30_{0.00}$	$46.81_{0.00}$	$33.37_{0.00}$	$13.58_{0.00}$	$18.57_{0.00}$	$22.00_{0.00}$	$42.42_{0.00}$	$13.50_{0.00}$	$6.37_{0.00}$	$20.26_{0.00}$	$5.92_{0.00}$	$49.98_{0.00}$	27.17
			w/	$58.20_{2.84}$	$48.35_{0.11}$	$17.65_{7.90}$	$14.21_{4.88}$	$17.31_{0.55}$	$22.03_{1.67}$	$17.86_{2.16}$	$17.87_{2.35}$	$5.43_{1.65}$	$14.39_{0.67}$	$6.90_{0.26}$	$28.80_{1.95}$	22.42
		Test	w/o	$52.61_{0.00}$	$51.74_{0.00}$	$32.92_{0.00}$	$14.46_{0.00}$	$19.68_{0.00}$	$27.22_{0.00}$	$39.90_{0.00}$	$10.44_{0.00}$	$8.01_{0.00}$	$22.48_{0.00}$	$12.02_{0.00}$	$49.70_{0.00}$	28.43
			w/	$57.49_{2.32}$	$54.24_{0.30}$	$17.41_{10.77}$	$20.08_{5.50}$	$20.17_{0.46}$	$25.34_{0.77}$	$21.52_{3.83}$	$8.81_{0.48}$	$3.94_{2.29}$	$11.62_{0.67}$	$12.48_{0.25}$	$14.84_{2.55}$	22.33

In terms of implementation, we use the `MulticlassCalibrationError` module given by the TorchMetrics [3] with `n_bins=10`, `norm='l1'`.

G Complete Results: Reliability Diagrams and Confident Distribution

Reliability diagrams: as shown in Fig. 9 - 14 for GPT-J, and Fig. 21 - 24 for GPT-2.
Confidence distributions: as shown in Fig. 15 - 20 for GPT-J, and Fig. 25 - 28 for GPT-2.

H Case Analysis: NoisyICL Furtherance Correct ICL

Moreover, we find that in some cases, unperturbed ICL can't benefit correctly from scaling the number of demos, while, NoisyICL can help the model correct this issue, as shown in Fig. 5. These unperturbed models exhibit an overfitting-like phenomenon and also low accuracies, while NoisyICL can relieve it.

We speculate the reason is the mismatch between the pre-trained knowledge and ICL inputs. This leads to a decrease in the model's in-context task learning (Pan et al., 2023) ability, while NoisyICL reduces such a gap between pre-trained data and ICL style data, which makes models extract information from ICL inputs better.

[3] lightning.ai/docs/torchmetrics/stable/classification/calibration_error.html#multiclasscalibrationerror

I License for Artifacts

Here we discuss the license of the artifacts used in this paper.

Models. GPT-2 is under the MIT license, and GPT-J is under the apache-2.0 license.

Datasets. We list the open-source license for the datasets used in this paper as follows:

- cc-by-4.0: PS, SE'14R, SE'14L, TEE, TES, TEH
- cc-by-sa-3.0: HS, SST2, RTE, MRPC, FP
- agpl-3.0: Ethos

Consistency of Usage. Models are used with their original usage. Due to the different data splitting of these datasets, to ensure the consistency of the experiment methods, we use a re-splitting method as described in Appendix A. However, the overall usage is consistent with their intended use.

Table 4: Accuracy and Macro-F1 results (%, $mean_{std}$, $k = 1$). A better result is in green. Notation is the same with the Tabel 1.

	Dataset			PS	HS	SE'14R	SE'14L	RTE	MRPC	Ethos	FP	SST2	TEE	TES	TEH	Mean
		λ		0.2	0.012	0.1	0.02	0.2	0.016	0.016	0.014	0.018	0.014	0.002	0.08	—
GPT-J	Acc. (%)	Val.	w/o	$39.43_{0.99}$	$76.02_{0.86}$	$57.85_{1.45}$	$53.77_{1.82}$	$50.41_{1.15}$	$58.48_{1.48}$	$50.32_{1.81}$	$53.93_{0.82}$	$68.20_{0.52}$	$43.48_{0.80}$	$43.83_{2.45}$	$50.90_{1.02}$	52.22
			w/	$53.24_{4.76}$	$76.86_{0.54}$	$64.28_{2.15}$	$55.57_{1.05}$	$51.76_{1.55}$	$58.28_{1.35}$	$51.00_{1.77}$	$53.96_{1.27}$	$71.56_{1.07}$	$44.32_{0.50}$	$46.37_{1.09}$	$52.13_{0.96}$	56.61
		Test	w/o	$35.08_{1.74}$	$84.92_{0.94}$	$53.66_{1.12}$	$51.60_{1.29}$	$50.52_{1.31}$	$53.50_{2.26}$	$50.40_{1.37}$	$52.29_{1.48}$	$73.89_{0.84}$	$41.94_{0.92}$	$44.80_{1.26}$	$50.71_{1.46}$	53.61
			w/	$46.31_{3.18}$	$85.64_{1.06}$	$58.26_{1.49}$	$53.71_{1.08}$	$51.84_{0.51}$	$54.28_{1.10}$	$50.88_{2.16}$	$51.96_{1.10}$	$73.85_{1.67}$	$42.06_{0.93}$	$45.23_{1.75}$	$52.11_{1.38}$	55.51
	MF1 (%)	Val.	w/o	$26.01_{0.70}$	$24.73_{0.66}$	$50.78_{1.61}$	$52.26_{1.85}$	$50.36_{1.18}$	$51.11_{1.43}$	$49.04_{1.83}$	$38.25_{1.01}$	$67.39_{0.58}$	$26.71_{1.08}$	$38.75_{2.41}$	$49.42_{0.85}$	43.73
			w/	$23.18_{1.34}$	$24.49_{0.92}$	$52.49_{2.89}$	$53.61_{1.35}$	$48.28_{4.78}$	$50.61_{1.72}$	$49.88_{1.81}$	$38.94_{1.99}$	$71.02_{1.16}$	$27.65_{1.31}$	$41.43_{1.61}$	$50.75_{0.97}$	44.36
		Test	w/o	$25.30_{1.20}$	$24.73_{0.71}$	$47.91_{1.24}$	$50.62_{1.25}$	$50.17_{1.19}$	$49.97_{2.33}$	$49.81_{1.33}$	$37.90_{1.98}$	$72.25_{1.04}$	$24.67_{1.43}$	$40.10_{1.28}$	$48.45_{1.61}$	43.49
			w/	$24.76_{1.72}$	$25.13_{1.02}$	$47.53_{2.14}$	$52.18_{0.89}$	$47.10_{4.38}$	$50.65_{1.07}$	$50.36_{2.14}$	$38.93_{1.27}$	$72.22_{1.96}$	$24.97_{0.92}$	$40.45_{1.73}$	$49.94_{1.40}$	52.42
	ECE_1 (%, ↓)	Val.	w/o	$23.09_{1.25}$	$22.67_{0.82}$	$11.25_{1.73}$	$17.01_{1.66}$	$49.33_{1.14}$	$35.31_{1.39}$	$47.42_{1.82}$	$27.54_{0.97}$	$5.33_{1.23}$	$44.33_{1.12}$	$28.71_{2.16}$	$45.64_{1.19}$	29.80
			w/	$16.64_{4.79}$	$21.81_{0.59}$	$13.33_{2.62}$	$15.57_{0.64}$	$26.82_{5.71}$	$35.56_{1.89}$	$46.63_{1.82}$	$27.39_{1.11}$	$3.72_{1.19}$	$42.88_{0.69}$	$26.36_{1.40}$	$43.95_{1.10}$	26.72
		Test	w/o	$28.18_{1.95}$	$13.97_{0.95}$	$12.05_{1.52}$	$15.95_{1.05}$	$49.20_{1.34}$	$39.94_{2.33}$	$47.35_{1.40}$	$28.24_{1.42}$	$4.60_{0.95}$	$45.34_{0.90}$	$27.63_{0.85}$	$46.98_{1.37}$	29.95
			w/	$20.85_{5.28}$	$13.21_{1.07}$	$15.45_{1.84}$	$14.28_{0.93}$	$25.64_{3.69}$	$35.56_{1.89}$	$46.76_{2.10}$	$28.13_{0.97}$	$5.53_{1.44}$	$44.71_{0.93}$	$27.08_{1.93}$	$45.28_{1.33}$	26.87
		λ		0.02	0.002	0.016	0.004	0.002	0.002	0.2	0.06	0.02	0.02	0.06	0.1	—
GPT-2	Acc. (%)	Val.	w/o	$42.42_{0.97}$	$41.37_{1.11}$	$45.68_{1.09}$	$39.69_{1.69}$	$50.62_{1.35}$	$68.71_{0.20}$	$47.26_{0.72}$	$45.45_{1.81}$	$57.79_{1.23}$	$28.11_{0.82}$	$41.45_{1.36}$	$42.70_{0.32}$	45.94
			w/	$45.55_{3.18}$	$37.34_{1.33}$	$48.75_{1.80}$	$41.66_{1.04}$	$50.18_{2.11}$	$68.81_{0.18}$	$53.35_{2.99}$	$55.33_{2.72}$	$59.26_{1.38}$	$32.73_{0.53}$	$45.25_{1.34}$	$48.41_{2.15}$	48.88
		Test	w/o	$35.07_{1.44}$	$39.64_{0.61}$	$43.24_{1.43}$	$38.47_{0.97}$	$50.22_{1.27}$	$62.94_{0.42}$	$48.65_{0.59}$	$45.09_{1.51}$	$59.76_{1.67}$	$27.61_{1.20}$	$37.98_{1.24}$	$44.37_{0.81}$	44.42
			w/	$40.08_{2.91}$	$34.92_{1.12}$	$44.36_{0.95}$	$39.41_{1.30}$	$50.58_{2.31}$	$62.47_{0.29}$	$51.69_{2.95}$	$52.12_{1.60}$	$60.64_{1.78}$	$30.55_{0.79}$	$43.88_{1.53}$	$55.35_{0.79}$	47.17
	MF1 (%)	Val.	w/o	$26.44_{0.73}$	$20.32_{1.44}$	$32.96_{0.78}$	$33.69_{1.92}$	$50.59_{1.35}$	$41.67_{0.39}$	$43.18_{0.82}$	$30.19_{1.51}$	$56.71_{1.25}$	$22.59_{0.81}$	$36.97_{1.57}$	$32.23_{0.53}$	35.63
			w	$26.37_{1.41}$	$18.95_{1.29}$	$35.97_{1.24}$	$35.61_{1.25}$	$50.13_{2.10}$	$42.28_{0.48}$	$47.26_{2.81}$	$34.47_{0.77}$	$59.22_{1.39}$	$22.87_{1.07}$	$30.49_{1.97}$	$51.68_{2.03}$	37.94
		Test	w/o	$23.97_{1.23}$	$17.45_{0.85}$	$33.27_{1.93}$	$36.32_{1.03}$	$49.85_{1.26}$	$39.41_{0.66}$	$43.80_{0.77}$	$30.92_{1.38}$	$59.63_{1.61}$	$22.32_{1.38}$	$36.21_{1.30}$	$38.79_{1.06}$	36.00
			w/	$26.31_{1.58}$	$15.87_{1.14}$	$33.78_{1.33}$	$37.11_{1.53}$	$50.18_{2.40}$	$39.10_{0.41}$	$44.93_{3.66}$	$34.58_{1.06}$	$59.52_{1.82}$	$21.69_{1.08}$	$33.40_{1.36}$	$50.57_{1.50}$	37.25
	ECE_1 (%, ↓)	Val.	w/o	$14.45_{1.49}$	$25.96_{1.14}$	$16.68_{1.33}$	$25.55_{1.99}$	$42.92_{1.41}$	$19.22_{0.49}$	$35.57_{1.01}$	$21.11_{1.76}$	$5.40_{1.16}$	$31.62_{1.17}$	$20.72_{1.70}$	$46.15_{0.38}$	25.44
			w/	$9.64_{1.86}$	$30.37_{1.36}$	$17.27_{2.01}$	$23.65_{0.78}$	$43.31_{2.03}$	$17.87_{0.49}$	$19.53_{3.89}$	$14.80_{0.62}$	$3.94_{0.41}$	$30.66_{0.71}$	$24.11_{2.54}$	$26.44_{2.47}$	21.80
		Test	w/o	$18.82_{1.59}$	$26.53_{0.96}$	$18.64_{1.41}$	$24.56_{0.96}$	$43.56_{1.28}$	$24.09_{0.58}$	$35.15_{0.76}$	$22.13_{1.42}$	$2.83_{1.04}$	$32.31_{1.34}$	$22.45_{1.38}$	$39.96_{0.77}$	25.92
			w/	$13.20_{2.03}$	$32.02_{1.47}$	$20.45_{1.07}$	$23.83_{1.22}$	$43.07_{2.30}$	$23.01_{0.87}$	$23.65_{6.52}$	$14.48_{1.15}$	$3.22_{1.50}$	$33.43_{1.60}$	$19.37_{1.64}$	$23.89_{1.50}$	22.80

Table 5: Accuracy and Macro-F1 results (%, $mean_{std}$, $k = 2$). A better result is in green. Notation is the same with the Tabel 1.

	Dataset			PS	HS	SE'14R	SE'14L	RTE	MRPC	Ethos	FP	SST2	TEE	TES	TEH	Mean
		λ		0.2	0.2	0.1	0.1	0.08	0.2	0.014	0.08	0.016	0.018	0.02	0.008	—
GPT-J	Acc. (%)	Val.	w/o	$42.54_{1.49}$	$72.03_{0.66}$	$41.50_{0.97}$	$40.88_{0.49}$	$51.31_{1.19}$	$55.74_{1.24}$	$49.81_{2.28}$	$61.84_{0.60}$	$72.05_{0.96}$	$44.47_{0.62}$	$48.85_{0.75}$	$49.08_{1.16}$	52.51
			w/	$57.11_{4.74}$	$77.34_{3.29}$	$58.57_{5.39}$	$51.25_{2.71}$	$50.76_{2.09}$	$58.98_{6.13}$	$51.37_{1.15}$	$60.47_{1.77}$	$73.38_{0.70}$	$44.67_{0.63}$	$49.14_{0.78}$	$50.62_{2.11}$	56.97
		Test	w/o	$39.23_{0.98}$	$82.42_{1.21}$	$36.69_{1.30}$	$39.26_{1.67}$	$48.13_{2.60}$	$50.97_{1.43}$	$51.41_{1.93}$	$60.77_{0.64}$	$72.60_{1.66}$	$44.28_{0.73}$	$48.79_{1.00}$	$51.68_{1.95}$	52.19
			w/	$50.95_{4.52}$	$82.99_{5.51}$	$53.91_{3.42}$	$49.30_{3.09}$	$49.45_{1.30}$	$53.43_{5.23}$	$51.46_{1.62}$	$59.51_{1.22}$	$72.83_{1.36}$	$44.21_{0.53}$	$48.77_{1.20}$	$51.02_{1.84}$	55.65
	MF1 (%)	Val.	w/o	$25.17_{1.10}$	$25.03_{0.33}$	$37.42_{1.13}$	$39.11_{0.83}$	$50.44_{1.23}$	$51.35_{1.45}$	$49.78_{2.28}$	$31.83_{1.31}$	$71.32_{1.07}$	$25.43_{1.03}$	$36.69_{1.16}$	$49.05_{1.17}$	41.05
			w/	$22.75_{1.84}$	$24.15_{0.56}$	$50.06_{5.11}$	$49.26_{2.50}$	$50.47_{2.22}$	$48.95_{0.99}$	$51.36_{1.15}$	$35.39_{3.22}$	$72.77_{0.78}$	$25.16_{0.84}$	$37.72_{1.56}$	$50.56_{2.12}$	43.22
		Test	w/o	$25.51_{1.25}$	$24.95_{0.85}$	$34.99_{1.38}$	$39.75_{1.63}$	$47.82_{2.64}$	$49.56_{1.48}$	$51.18_{1.96}$	$30.66_{0.98}$	$70.34_{1.87}$	$24.05_{0.97}$	$36.78_{1.28}$	$51.49_{1.96}$	40.59
			w/	$22.98_{1.69}$	$25.10_{0.79}$	$47.51_{2.96}$	$48.50_{3.12}$	$49.35_{1.25}$	$46.51_{1.56}$	$51.19_{1.62}$	$32.85_{2.38}$	$70.54_{1.76}$	$24.05_{0.67}$	$36.63_{0.53}$	$50.87_{1.80}$	42.17
	ECE_1 (%, ↓)	Val.	w/o	$15.11_{1.29}$	$21.52_{0.58}$	$23.45_{1.46}$	$27.07_{0.89}$	$39.26_{0.83}$	$23.82_{1.59}$	$38.51_{2.24}$	$22.54_{0.77}$	$3.56_{0.89}$	$44.33_{0.66}$	$19.22_{0.62}$	$40.33_{1.30}$	26.56
			w/	$13.69_{3.31}$	$11.25_{5.10}$	$8.03_{4.49}$	$15.54_{3.79}$	$39.53_{1.65}$	$12.12_{2.82}$	$37.12_{0.98}$	$21.10_{1.70}$	$4.80_{0.74}$	$43.87_{0.96}$	$18.64_{1.08}$	$38.66_{2.22}$	22.03
		Test	w/o	$18.59_{1.01}$	$13.15_{1.15}$	$27.87_{1.35}$	$26.71_{1.83}$	$41.77_{2.91}$	$28.29_{1.47}$	$36.62_{1.76}$	$22.97_{0.87}$	$4.77_{0.84}$	$44.53_{0.98}$	$19.92_{1.25}$	$36.63_{1.86}$	26.82
			w/	$13.49_{2.88}$	$6.52_{2.25}$	$11.44_{3.91}$	$15.80_{2.76}$	$41.33_{2.22}$	$20.24_{6.56}$	$36.61_{1.67}$	$21.88_{1.75}$	$5.32_{1.16}$	$44.40_{0.77}$	$19.98_{0.70}$	$37.48_{1.92}$	22.87
		λ		0.02	0.1	0.08	0.08	0.006	0.002	0.2	0.04	0.02	0.02	0.04	0.2	—
GPT-2	Acc. (%)	Val.	w/o	$41.86_{1.78}$	$48.24_{1.37}$	$46.99_{0.84}$	$42.42_{1.10}$	$50.88_{1.05}$	$68.55_{0.16}$	$45.85_{1.27}$	$49.10_{1.15}$	$56.56_{1.31}$	$29.06_{0.71}$	$39.39_{1.51}$	$42.81_{0.63}$	46.81
			w/	$47.23_{3.35}$	$48.73_{6.40}$	$52.85_{5.89}$	$46.62_{1.59}$	$51.80_{1.48}$	$68.89_{0.43}$	$52.84_{2.43}$	$61.31_{0.81}$	$59.45_{1.20}$	$30.29_{0.82}$	$45.70_{1.46}$	$51.09_{4.07}$	51.40
		Test	w/o	$38.26_{1.37}$	$52.04_{0.91}$	$42.52_{1.15}$	$39.71_{1.47}$	$49.67_{0.80}$	$62.77_{0.39}$	$48.26_{1.00}$	$47.42_{1.45}$	$58.96_{1.07}$	$28.47_{0.97}$	$32.11_{1.37}$	$44.74_{0.82}$	45.41
			w/	$40.95_{1.84}$	$45.52_{3.99}$	$50.12_{3.58}$	$40.31_{0.90}$	$50.64_{1.87}$	$57.34_{1.01}$	$51.76_{2.27}$	$59.41_{1.11}$	$61.33_{1.51}$	$28.49_{1.04}$	$40.29_{3.18}$	$53.01_{3.31}$	48.26
	MF1 (%)	Val.	w/o	$25.20_{1.70}$	$21.61_{1.20}$	$36.33_{1.10}$	$36.50_{0.51}$	$50.38_{1.03}$	$41.38_{0.39}$	$41.44_{1.75}$	$33.91_{1.20}$	$55.30_{1.30}$	$20.40_{1.31}$	$35.73_{1.63}$	$33.12_{1.17}$	35.94
			w	$25.40_{0.74}$	$21.77_{1.89}$	$32.89_{2.37}$	$28.58_{2.09}$	$51.21_{1.53}$	$42.70_{0.82}$	$48.94_{1.40}$	$31.66_{2.16}$	$59.17_{1.11}$	$19.12_{0.69}$	$30.66_{1.33}$	$46.38_{4.14}$	36.54
		Test	w/o	$24.91_{1.27}$	$20.54_{0.43}$	$35.57_{0.99}$	$38.02_{1.35}$	$48.42_{0.93}$	$39.80_{0.37}$	$42.83_{1.35}$	$33.29_{1.18}$	$58.84_{1.10}$	$20.96_{1.15}$	$31.99_{1.36}$	$40.79_{1.13}$	36.33
			w/	$25.63_{1.20}$	$18.73_{0.73}$	$31.81_{1.94}$	$28.97_{2.76}$	$49.44_{1.83}$	$44.06_{1.11}$	$45.99_{4.55}$	$33.03_{1.60}$	$59.72_{1.13}$	$18.40_{1.10}$	$33.02_{1.48}$	$50.92_{1.60}$	36.64
	ECE_1 (%, ↓)	Val.	w/o	$10.21_{1.71}$	$25.07_{1.21}$	$14.93_{1.00}$	$21.00_{1.10}$	$35.37_{1.19}$	$20.19_{0.40}$	$41.29_{1.34}$	$15.87_{1.07}$	$6.43_{1.22}$	$30.94_{0.72}$	$18.02_{1.74}$	$46.90_{0.60}$	23.85
			w/	$5.92_{1.26}$	$29.23_{4.69}$	$12.70_{1.65}$	$22.24_{1.86}$	$34.69_{1.23}$	$19.18_{0.61}$	$21.02_{2.77}$	$10.53_{2.79}$	$3.97_{1.86}$	$35.02_{0.56}$	$16.78_{1.71}$	$27.18_{4.94}$	19.87
		Test	w/o	$12.48_{1.44}$	$19.52_{0.68}$	$17.95_{1.41}$	$20.35_{1.46}$	$36.73_{0.99}$	$23.63_{0.44}$	$39.93_{1.27}$	$17.40_{1.25}$	$4.08_{1.05}$	$31.48_{1.10}$	$24.73_{1.44}$	$42.57_{1.24}$	24.24
			w/	$9.58_{1.56}$	$33.12_{4.08}$	$16.31_{1.69}$	$28.04_{3.60}$	$36.06_{1.95}$	$17.86_{0.86}$	$24.58_{6.04}$	$8.99_{2.05}$	$2.09_{0.84}$	$37.68_{1.94}$	$15.45_{1.96}$	$21.12_{2.08}$	20.91

Table 6: Accuracy and Macro-F1 results (%, $mean_{std}$, $k = 8$). A better result is in green. Notation is the same with the Tabel 1. Some experiments can't be conducted due to the length of the input sequence.

	Dataset			PS	HS	SE'14R	SE'14L	RTE	MRPC	Ethos	FP	SST2	TEE	TES	TEH	Mean
	λ			0.2	0.2	0.1	0.1	0.1	0.2	0.004	0.012	0.002	0.006	0.08	0.002	—
GPT-J	Acc. (%)	Val.	w/o	$27.32_{0.84}$	$68.87_{1.30}$	$35.61_{0.57}$	$33.57_{0.74}$	$50.12_{0.29}$	$35.86_{0.55}$	$57.29_{1.69}$	$62.34_{0.05}$	$83.96_{0.86}$	$47.64_{0.52}$	$51.13_{0.66}$	$50.62_{1.62}$	50.37
			w/	$62.01_{4.95}$	$85.29_{0.95}$	$45.84_{11.13}$	$40.84_{6.41}$	$50.49_{0.89}$	$58.87_{7.93}$	$57.95_{1.35}$	$62.44_{0.08}$	$83.89_{0.79}$	$47.81_{0.90}$	$51.35_{1.67}$	$52.05_{0.75}$	58.24
		Test	w/o	$25.70_{0.84}$	$77.10_{1.75}$	$33.30_{0.63}$	$30.80_{0.71}$	$46.35_{1.39}$	$39.35_{0.84}$	$61.67_{1.38}$	$61.76_{0.06}$	$80.59_{1.02}$	$46.37_{0.50}$	$50.68_{0.35}$	$54.57_{1.41}$	50.69
			w/	$58.54_{3.27}$	$89.74_{3.97}$	$41.68_{6.66}$	$35.66_{4.65}$	$47.53_{1.26}$	$54.57_{8.48}$	$61.46_{0.96}$	$61.78_{0.08}$	$79.78_{0.69}$	$46.42_{0.63}$	$50.62_{0.41}$	$54.68_{0.92}$	56.87
	MF1 (%)	Val.	w/o	$16.07_{0.47}$	$23.66_{0.32}$	$30.15_{1.14}$	$24.48_{1.40}$	$45.66_{0.49}$	$33.37_{0.56}$	$56.68_{1.81}$	$25.91_{0.26}$	$83.80_{0.91}$	$27.20_{0.59}$	$33.72_{1.47}$	$48.52_{1.98}$	37.44
			w/	$20.91_{1.79}$	$27.76_{1.35}$	$41.70_{10.30}$	$35.28_{8.07}$	$45.89_{3.44}$	$46.69_{2.90}$	$57.45_{1.37}$	$26.24_{0.36}$	$83.72_{0.82}$	$27.39_{0.84}$	$36.84_{5.55}$	$49.94_{0.99}$	41.65
		Test	w/o	$16.01_{0.94}$	$26.26_{1.29}$	$28.72_{0.67}$	$27.21_{0.78}$	$42.58_{1.37}$	$33.58_{0.98}$	$60.88_{1.44}$	$25.60_{0.25}$	$78.98_{1.18}$	$24.50_{0.70}$	$33.19_{0.40}$	$54.36_{1.42}$	37.66
			w/	$20.81_{1.17}$	$24.29_{0.56}$	$39.15_{6.23}$	$34.28_{5.55}$	$42.47_{4.45}$	$42.45_{4.20}$	$60.52_{1.02}$	$25.67_{0.31}$	$78.06_{0.79}$	$24.80_{0.92}$	$33.85_{3.11}$	$54.45_{0.96}$	40.07
	ECE_1 (%, ↓)	Val.	w/o	$42.98_{0.85}$	$10.43_{1.20}$	$32.79_{0.53}$	$42.43_{0.99}$	$21.59_{0.34}$	$33.19_{0.72}$	$13.63_{1.44}$	$24.54_{0.13}$	$12.61_{0.98}$	$40.10_{0.65}$	$13.99_{0.17}$	$21.11_{1.51}$	25.78
			w/	$11.81_{4.56}$	$9.51_{2.14}$	$19.27_{11.45}$	$25.96_{10.94}$	$23.13_{2.74}$	$8.10_{2.84}$	$12.47_{1.21}$	$24.33_{0.69}$	$12.80_{0.75}$	$39.95_{1.00}$	$7.25_{4.79}$	$20.01_{0.66}$	17.88
		Test	w/o	$44.92_{0.80}$	$7.93_{1.41}$	$36.15_{0.98}$	$43.80_{1.01}$	$25.13_{1.62}$	$29.83_{0.74}$	$9.35_{1.45}$	$25.00_{0.23}$	$10.52_{0.94}$	$41.83_{0.54}$	$14.13_{0.43}$	$16.78_{1.45}$	25.45
			w/	$8.86_{3.57}$	$6.04_{1.94}$	$22.26_{9.50}$	$29.44_{8.54}$	$28.55_{6.11}$	$15.28_{7.58}$	$9.77_{0.86}$	$24.92_{0.52}$	$9.88_{0.66}$	$41.83_{0.80}$	$9.24_{2.47}$	$16.77_{0.95}$	18.57

Table 7: Accuracy and Macro-F1 results (%, $mean_{std}$, $k = 16$). A better result is in green. Notation is the same with the Tabel 1. Some experiments can't be conducted due to the length of the input sequence.

	Dataset			PS	HS	SE'14R	SE'14L	RTE	FP	SST2	TEE	TES	TEH	Mean
	λ			0.2	0.2	0.1	0.1	0.2	0.1	0.004	0.008	0.08	0.002	—
GPT-J	Acc. (%)	Val.	w/o	$18.52_{0.78}$	$70.02_{0.52}$	$40.86_{0.90}$	$34.59_{0.41}$	$48.91_{0.62}$	$62.42_{0.10}$	$88.24_{0.70}$	$49.22_{0.47}$	$51.84_{0.97}$	$54.59_{0.87}$	51.92
			w/	$61.78_{5.94}$	$85.98_{0.61}$	$46.84_{12.92}$	$43.01_{8.67}$	$50.76_{1.15}$	$62.66_{0.35}$	$88.83_{0.71}$	$49.30_{0.53}$	$53.11_{1.99}$	$55.20_{0.91}$	59.75
		Test	w/o	$18.75_{0.38}$	$74.97_{1.43}$	$34.98_{1.10}$	$30.00_{0.58}$	$45.95_{0.45}$	$61.82_{0.09}$	$81.88_{1.00}$	$47.11_{0.35}$	$51.70_{0.42}$	$55.90_{0.81}$	50.31
			w/	$59.97_{2.40}$	$89.96_{4.15}$	$42.56_{9.98}$	$38.55_{7.65}$	$52.96_{1.39}$	$61.92_{0.49}$	$82.20_{0.86}$	$47.06_{0.65}$	$52.13_{0.88}$	$56.29_{1.21}$	58.36
	MF1 (%)	Val.	w/o	$10.33_{0.46}$	$30.44_{0.84}$	$32.79_{0.78}$	$24.15_{0.66}$	$38.73_{0.71}$	$26.08_{0.46}$	$88.22_{0.70}$	$28.75_{0.40}$	$36.01_{1.76}$	$53.61_{0.88}$	36.91
			w/	$20.00_{0.71}$	$23.54_{0.57}$	$41.11_{11.38}$	$35.07_{10.38}$	$42.85_{7.50}$	$27.18_{1.53}$	$88.81_{0.72}$	$29.04_{0.55}$	$40.22_{6.60}$	$54.02_{1.02}$	40.18
		Test	w/o	$9.61_{0.26}$	$27.64_{1.69}$	$28.80_{0.97}$	$25.05_{0.94}$	$35.57_{0.69}$	$25.80_{0.31}$	$80.33_{1.12}$	$25.50_{0.54}$	$35.44_{0.90}$	$55.76_{0.81}$	34.95
			w/	$19.70_{1.28}$	$23.96_{0.08}$	$37.93_{9.63}$	$36.06_{8.97}$	$41.17_{5.54}$	$26.27_{1.96}$	$80.61_{1.12}$	$25.45_{0.95}$	$37.20_{3.58}$	$56.13_{1.25}$	38.45
	ECE_1 (%, ↓)	Val.	w/o	$54.29_{0.95}$	$4.01_{0.77}$	$27.53_{1.27}$	$40.90_{0.52}$	$26.13_{0.54}$	$26.34_{0.14}$	$14.43_{0.75}$	$38.97_{0.58}$	$11.46_{0.99}$	$13.66_{0.72}$	25.77
			w/	$11.75_{5.60}$	$9.70_{2.72}$	$18.56_{10.81}$	$23.33_{13.20}$	$21.01_{13.24}$	$18.07_{5.22}$	$15.49_{0.58}$	$38.61_{0.69}$	$6.12_{2.85}$	$13.25_{0.68}$	17.59
		Test	w/o	$53.72_{0.37}$	$4.51_{0.87}$	$34.09_{1.27}$	$43.06_{0.62}$	$29.93_{0.68}$	$26.90_{0.19}$	$9.96_{1.00}$	$40.84_{0.45}$	$11.24_{0.41}$	$13.45_{0.86}$	26.77
			w/	$8.33_{4.47}$	$5.63_{2.27}$	$20.29_{12.35}$	$24.94_{10.05}$	$17.53_{7.30}$	$20.44_{3.90}$	$9.99_{0.89}$	$41.00_{0.83}$	$5.19_{2.52}$	$13.03_{1.33}$	16.64

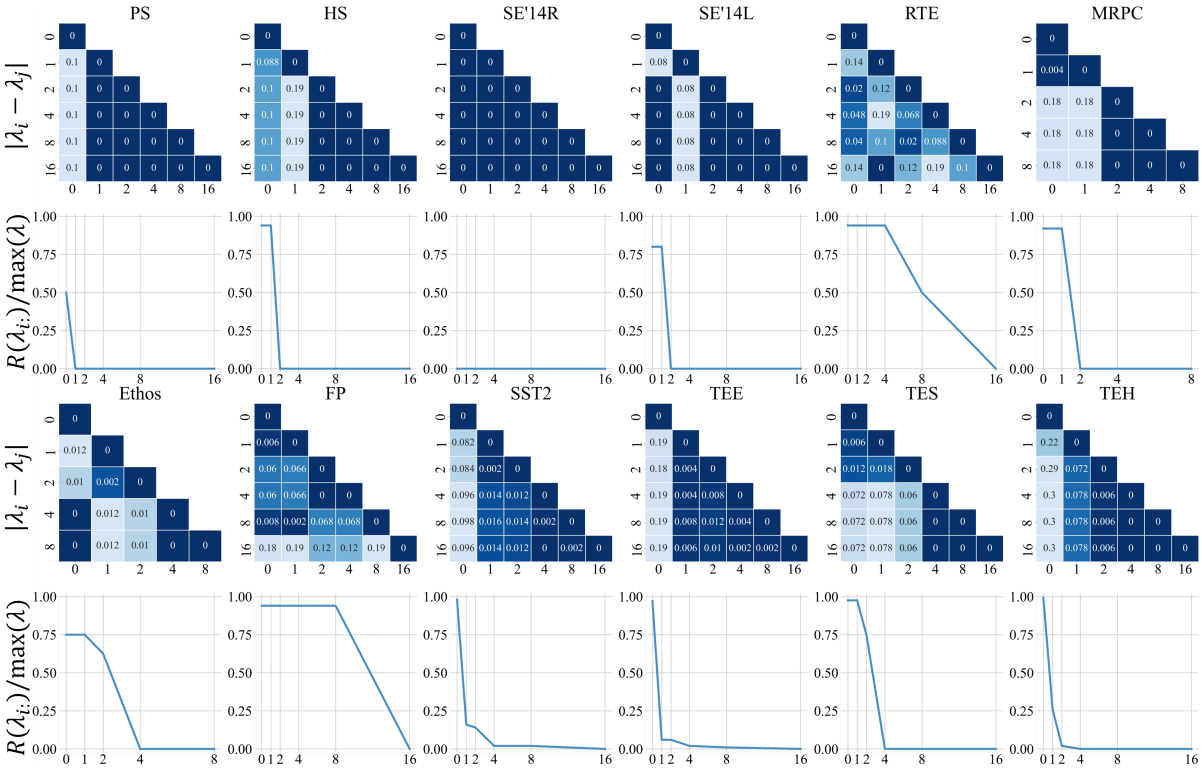

Figure 7: The stability of the optimal λ on various k for GPT-J. Upper figure: the distance between the optimal λ of different demo numbers; Lower figure: the normalized remaining range.

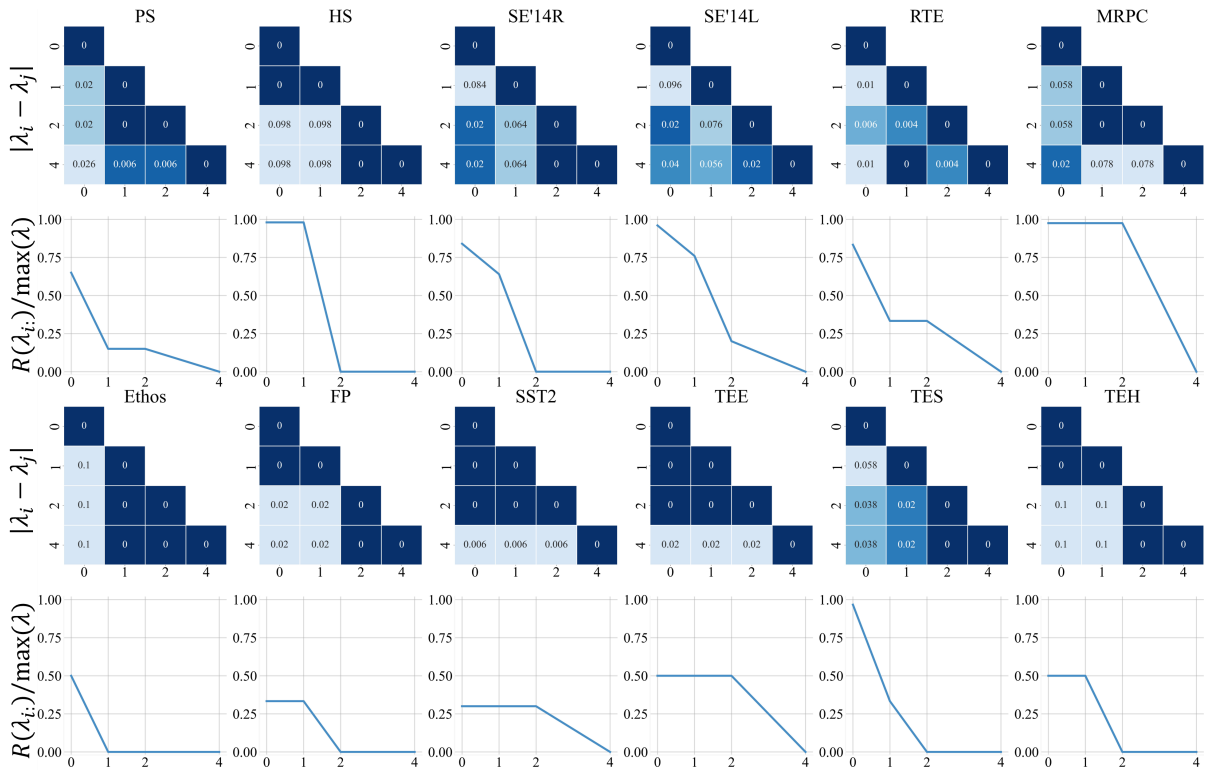

Figure 8: The stability of the optimal λ on various k for GPT-2. Upper figure: the distance between the optimal λ of different demo numbers; Lower figure: the normalized remaining range.

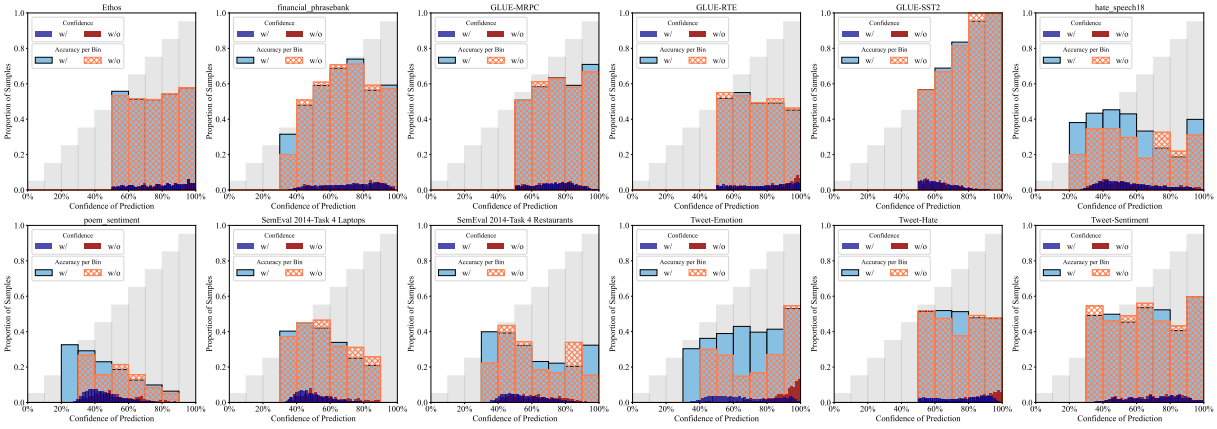

Figure 9: The reliability diagrams of GPT-J ($k = 0$).

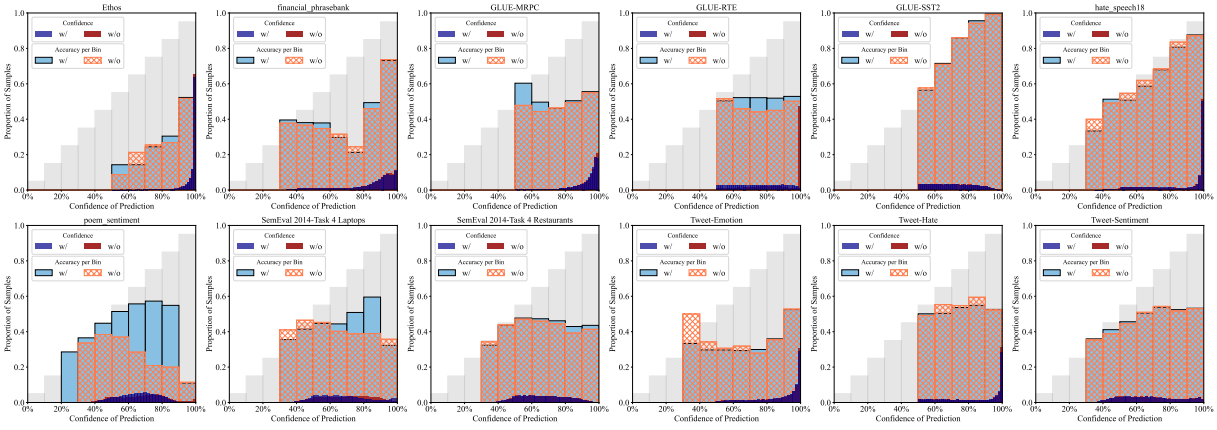

Figure 10: The reliability diagrams of GPT-J ($k = 1$).

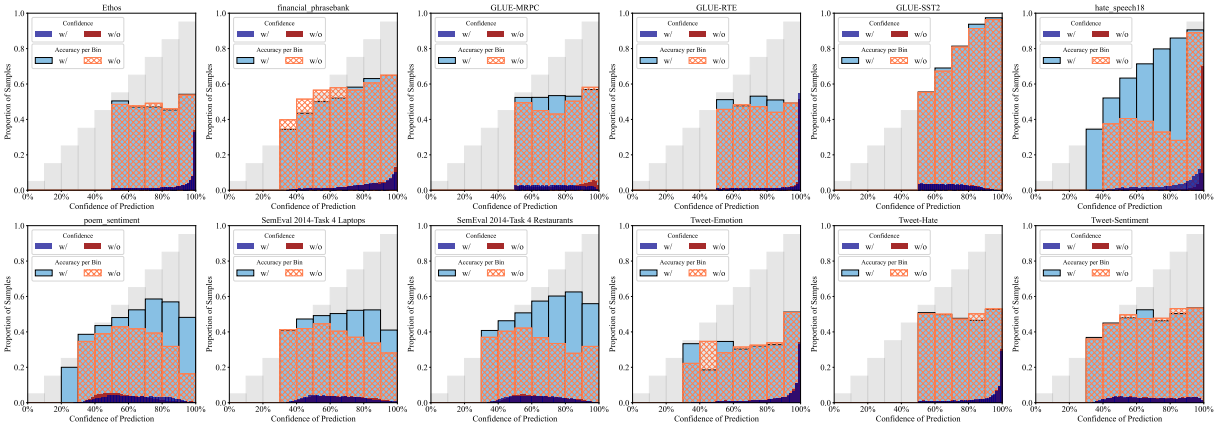

Figure 11: The reliability diagrams of GPT-J ($k = 2$).

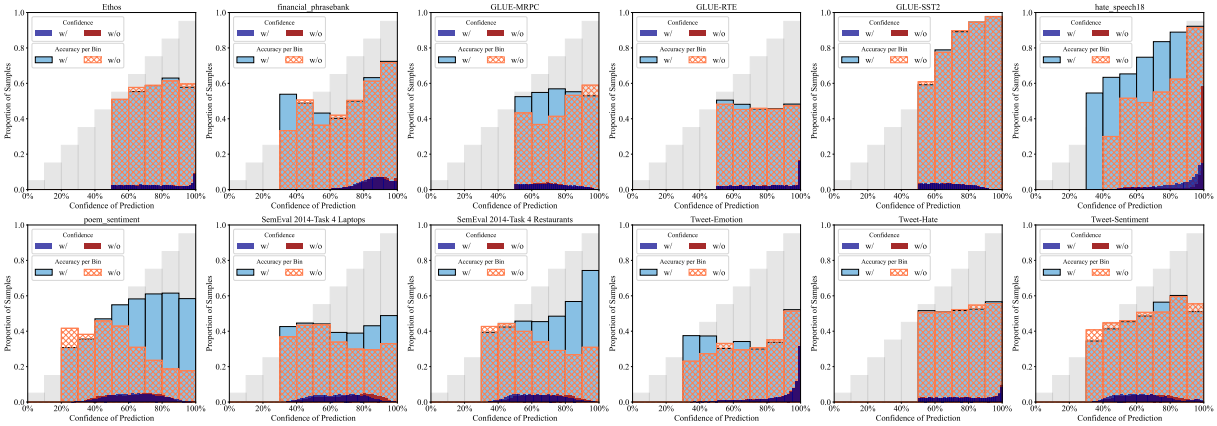

Figure 12: The reliability diagrams of GPT-J ($k = 4$).

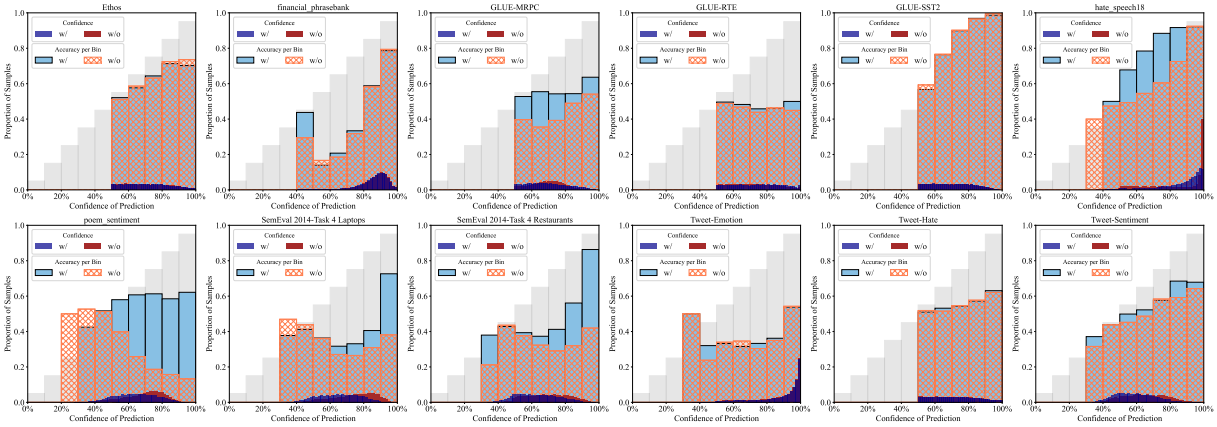

Figure 13: The reliability diagrams of GPT-J ($k = 8$).

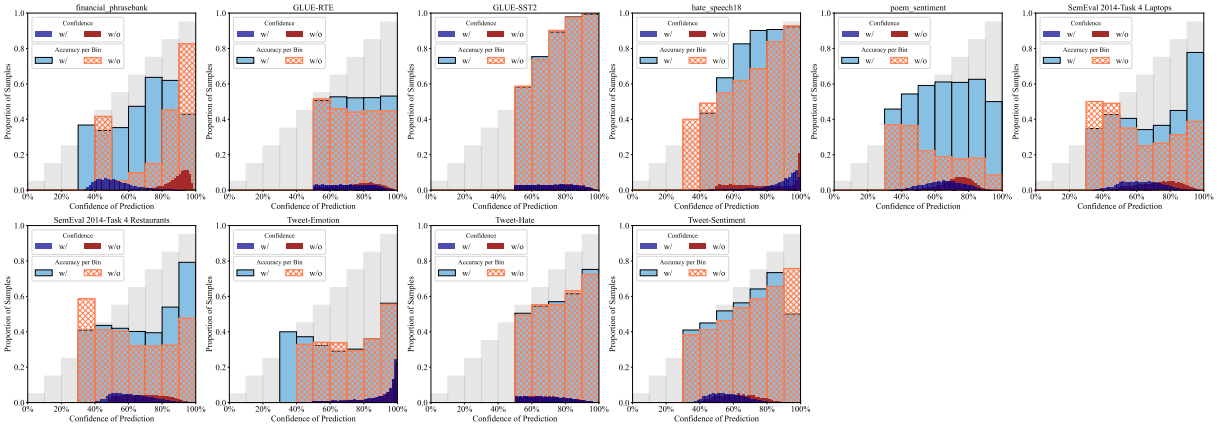

Figure 14: The reliability diagrams of GPT-J ($k = 16$).

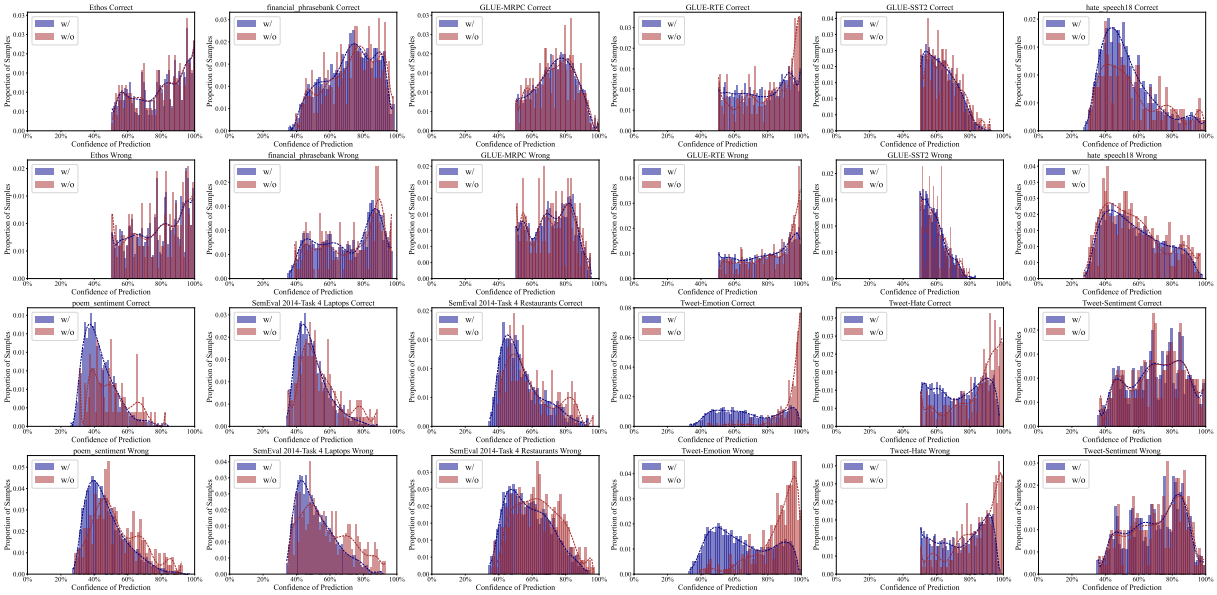

Figure 15: The confidence distribution of GPT-J ($k = 0$).

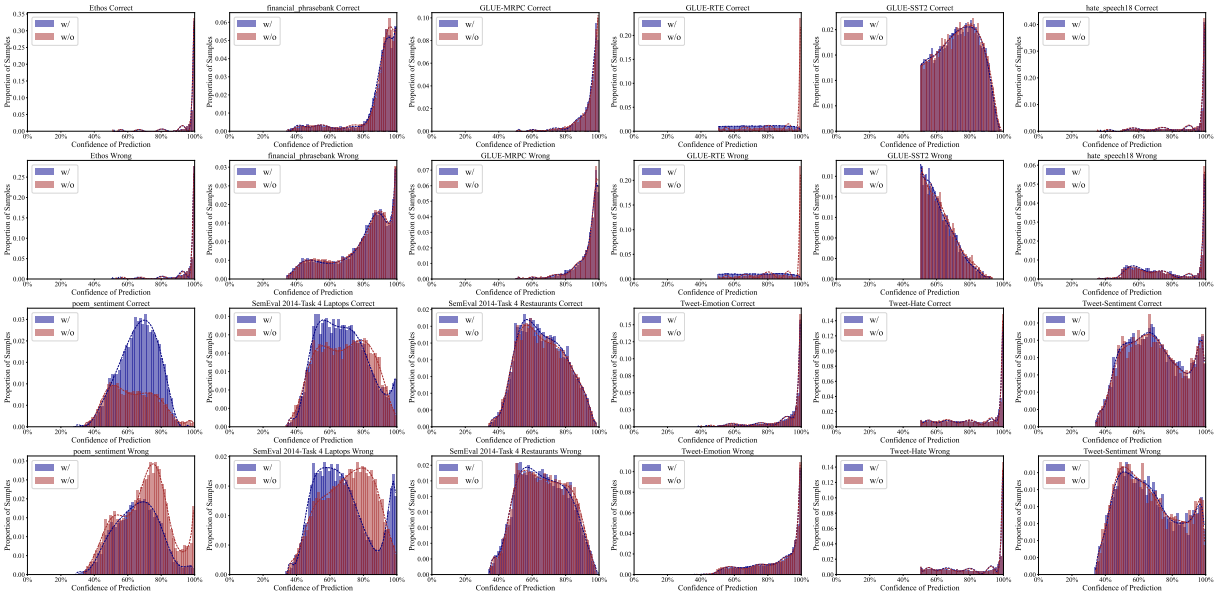

Figure 16: The confidence distribution of GPT-J ($k = 1$).

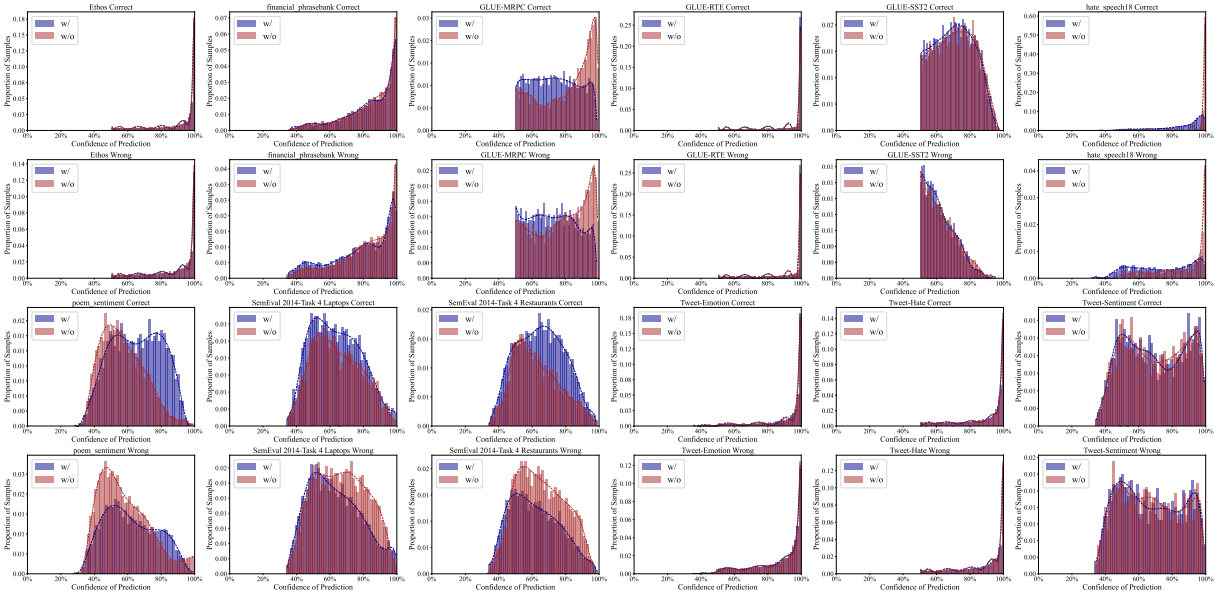

Figure 17: The confidence distribution of GPT-J ($k = 2$).

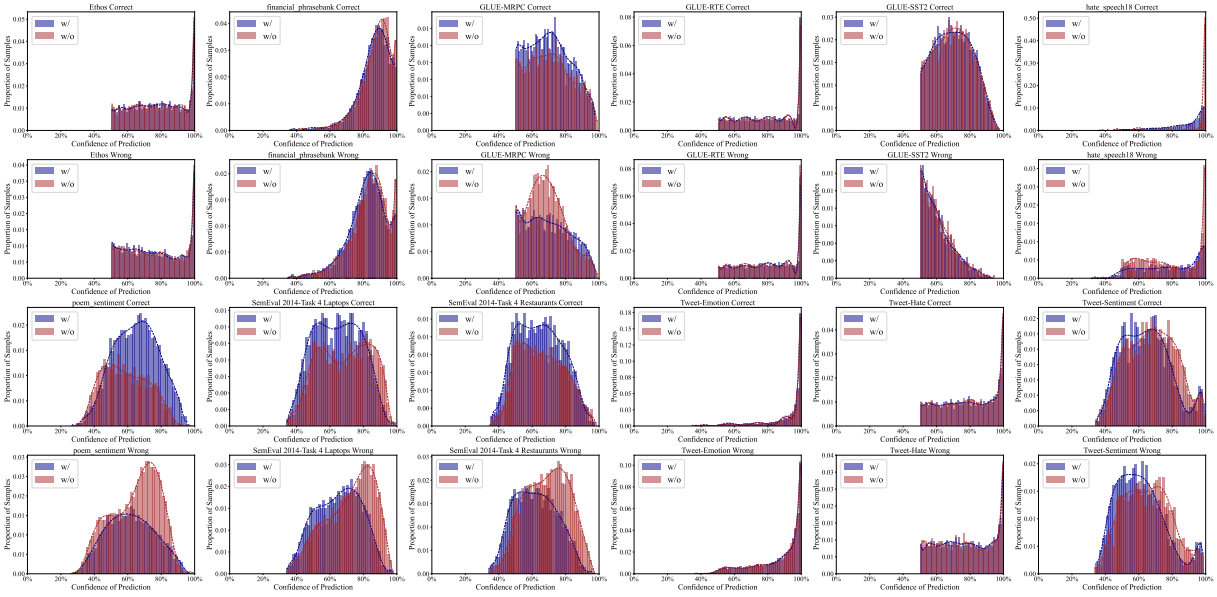

Figure 18: The confidence distribution of GPT-J ($k = 4$).

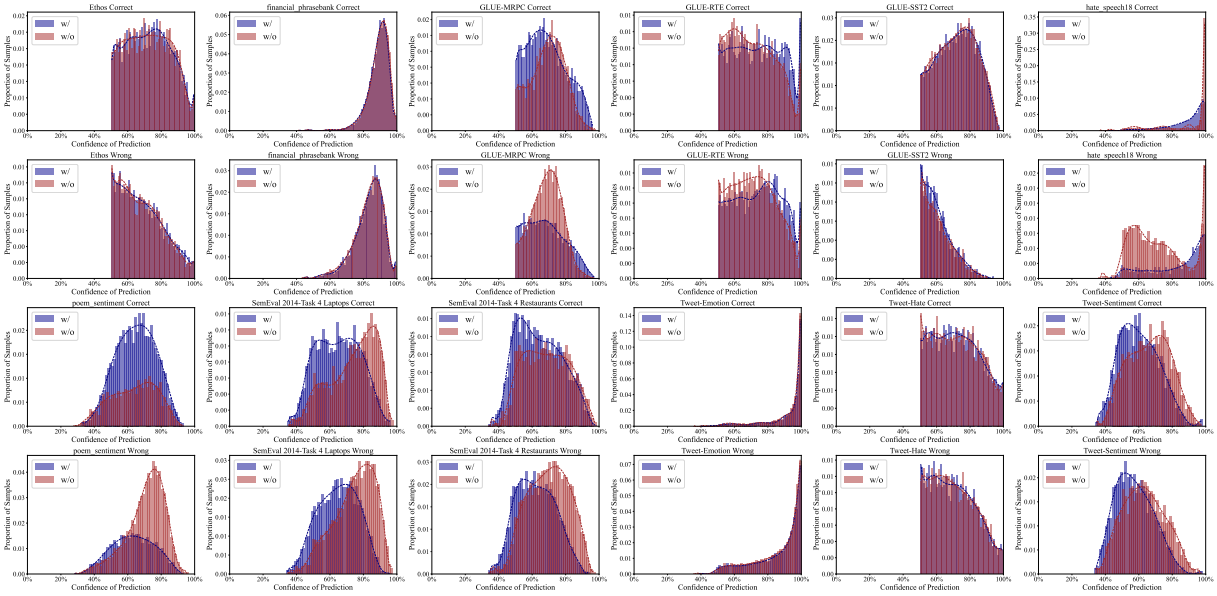

Figure 19: The confidence distribution of GPT-J ($k = 8$).

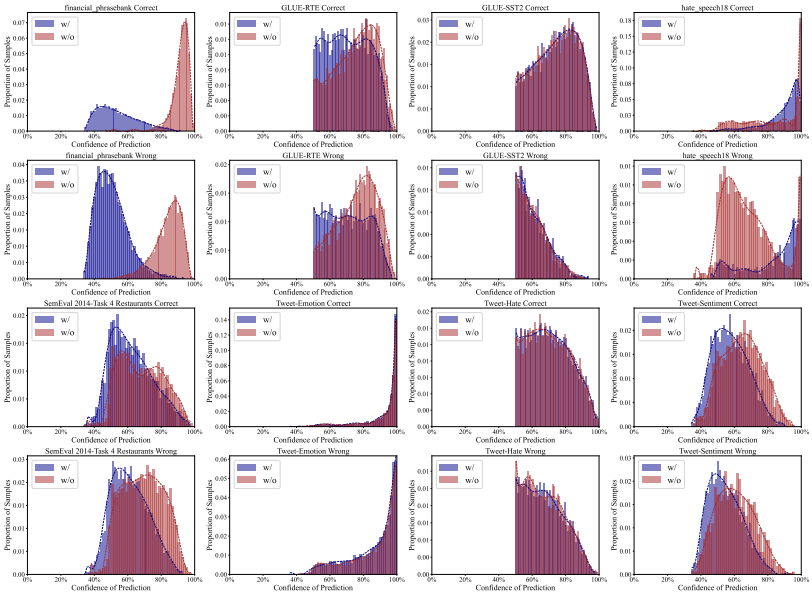

Figure 20: The confidence distribution of GPT-J ($k = 16$).

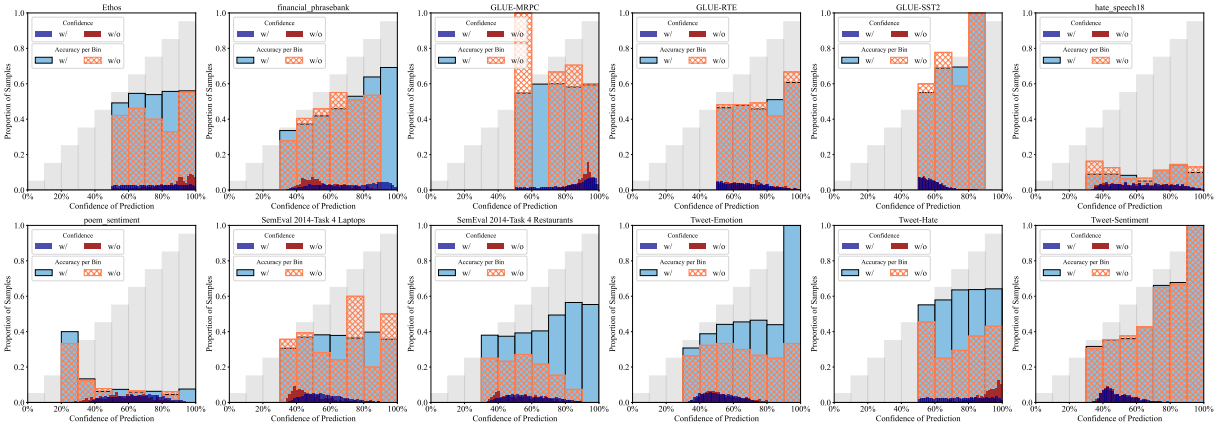

Figure 21: The reliability diagrams of GPT-2 ($k = 0$).

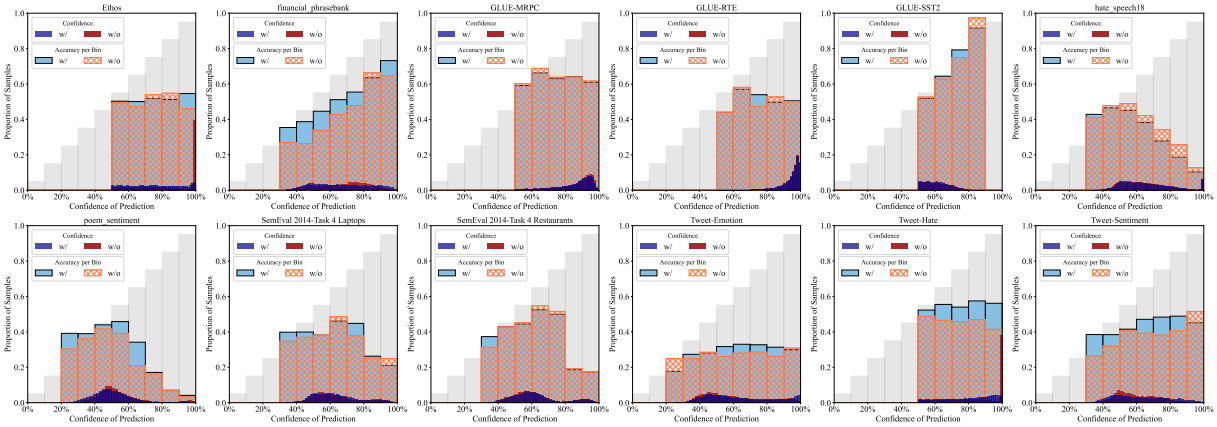

Figure 22: The reliability diagrams of GPT-2 ($k = 1$).

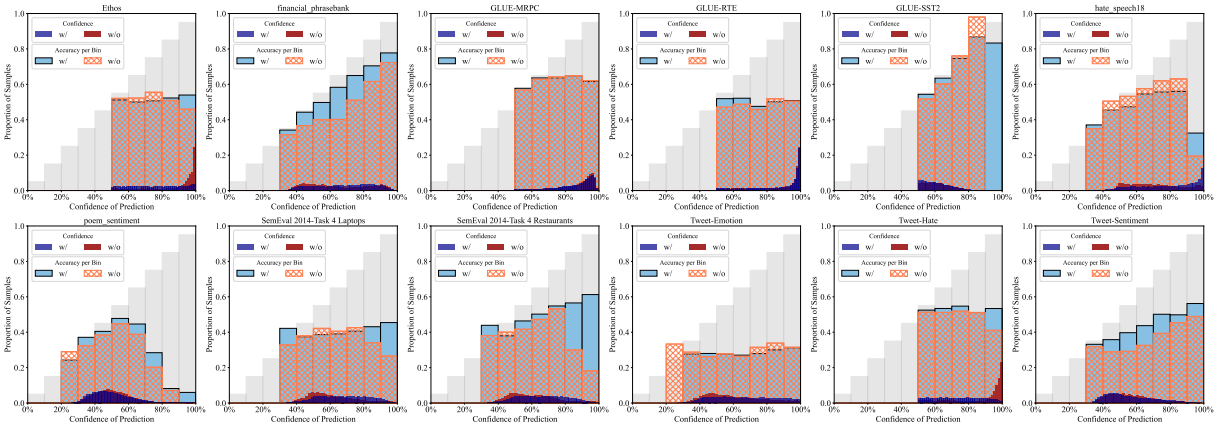

Figure 23: The reliability diagrams of GPT-2 ($k = 2$).

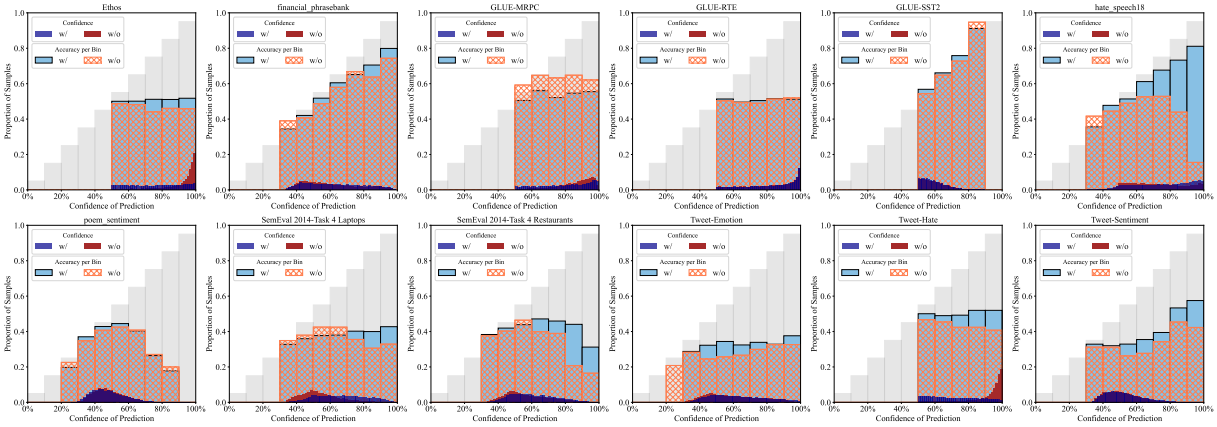

Figure 24: The reliability diagrams of GPT-2 ($k = 4$).

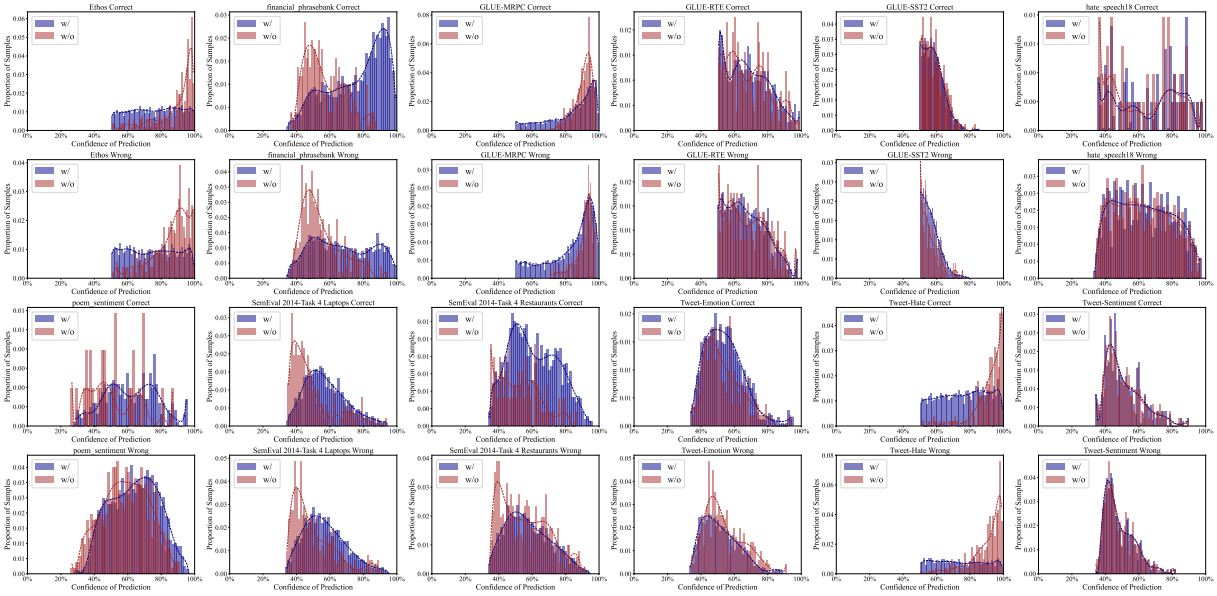

Figure 25: The confidence distribution of GPT-2 ($k = 0$).

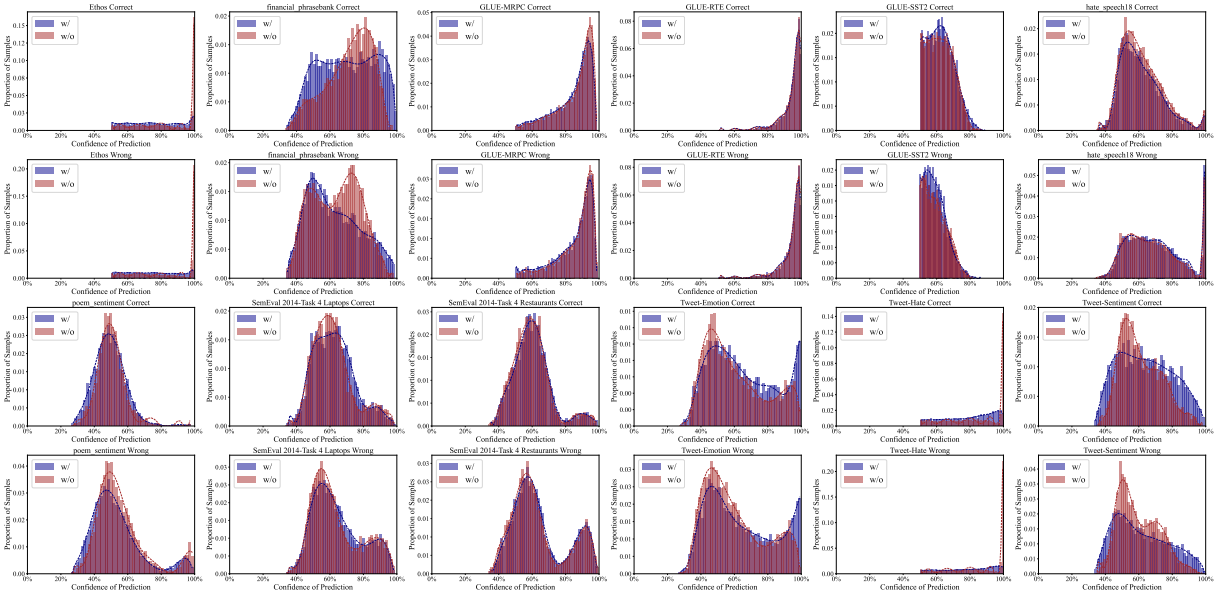

Figure 26: The confidence distribution of GPT-2 ($k = 1$).

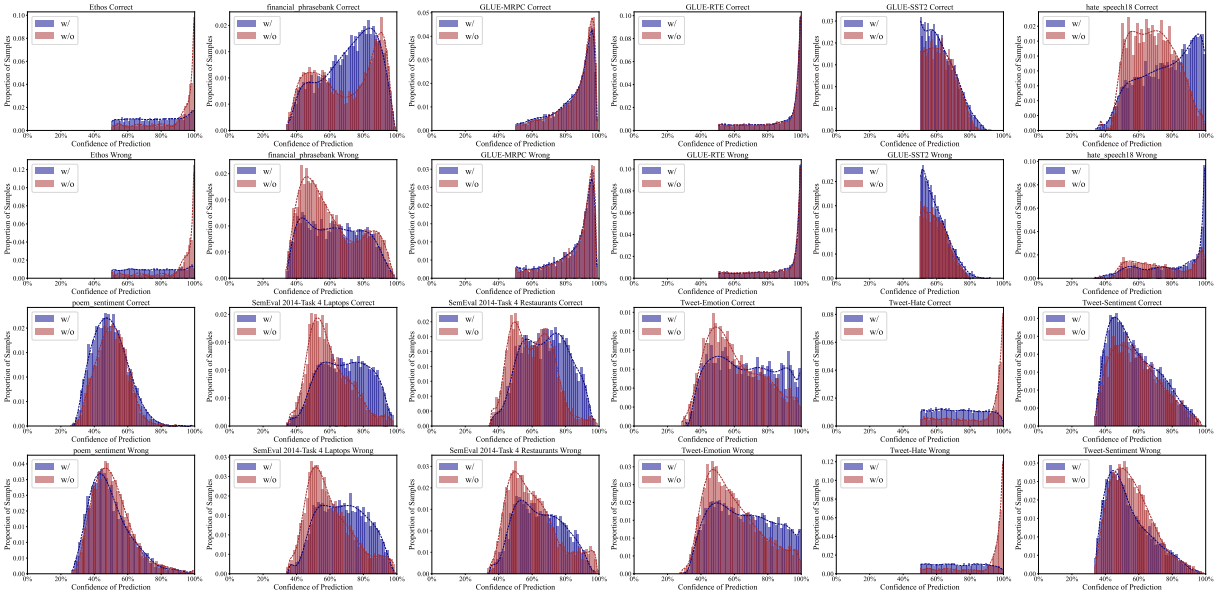

Figure 27: The confidence distribution of GPT-2 ($k = 2$).

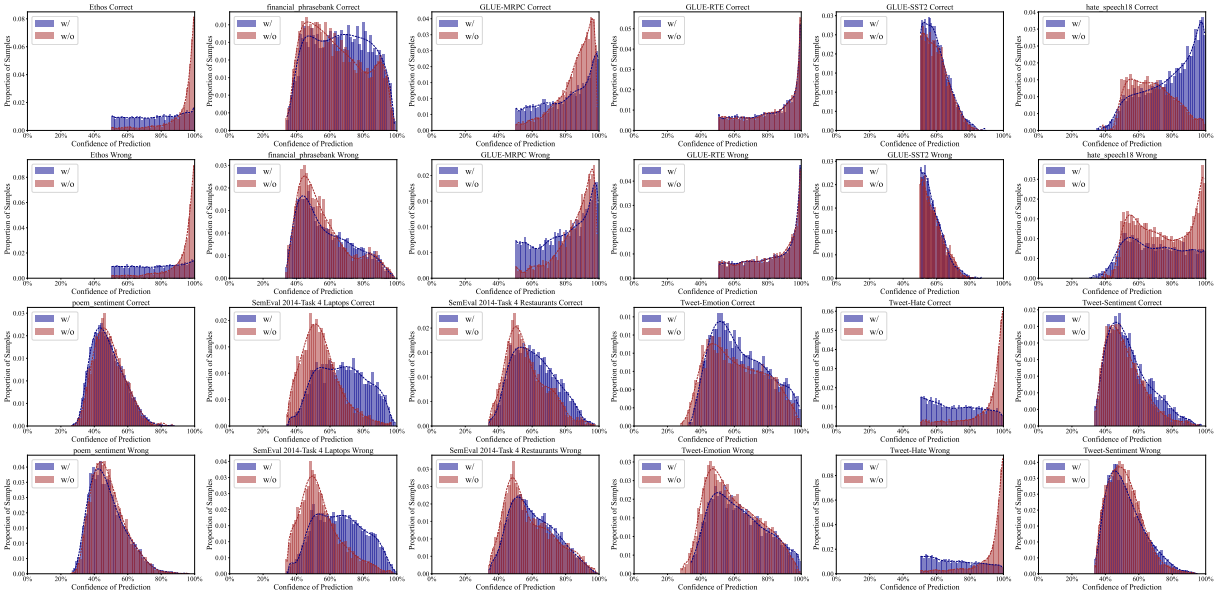

Figure 28: The confidence distribution of GPT-2 ($k = 4$).