In-Context Example Ordering Guided by Label Distributions

Zhichao Xu[1,2] **Daniel Cohen**[3] **Bei Wang**[1,2] **Vivek Srikumar**[1]

[1]Kahlert School of Computing, University of Utah
[2]Scientific Computing and Imaging Institute, University of Utah
[3]Dataminr, Inc.
zhichao.xu@utah.edu

Abstract

By allowing models to predict without task-specific training, in-context learning (ICL) with pretrained LLMs has enormous potential in NLP. However, a number of problems persist in ICL. In particular, its performance is sensitive to the choice and order of in-context examples. Given the same set of in-context examples with different orderings, model performance may vary from near random to near state-of-the-art. In this work, we formulate in-context example ordering as an optimization problem. We examine three problem settings that differ in the assumptions they make about what is known about the task. Inspired by the idea of learning from label proportions, we propose two principles for in-context example ordering guided by model's probability predictions. We apply our proposed principles to thirteen text classification datasets and nine different autoregressive LLMs with 700M to 13B parameters. We demonstrate that our approach outperforms the baselines by improving the classification accuracy, reducing model miscalibration, and also by selecting better in-context examples.

1 Introduction

An intriguing property of large language models like the GPT (Brown et al., 2020; OpenAI, 2023) and PaLM families of models (Chowdhery et al., 2022; Anil et al., 2023) is their ability to "learn in context". That is, the model can achieve competitive predictive performance with only a task description and a few training examples with no parameter updates (Brown et al., 2020; Min et al., 2022b; Xie et al., 2021). Model predictions can sometimes even match full fine-tuning performance (Lu et al., 2022).

In-context learning (ICL)—the idea of prompting LLMs with only a few examples, also known as few-shot prompting—has shown promise across NLP. Yet, many problems persist with this paradigm. Prior work has shown that ICL is sensitive to different natural language instructions and different orderings of in-context examples (Sorensen et al., 2022; Lu et al., 2022). Merely changing the ordering of a fixed set of examples can change the predictive performance from that of nearly fully-tuned models to random guessing. Lu et al. (2022) studied in-context ordering and proposed heuristics to select the performant orderings. However, prior work on example ordering assumes (to different degrees) that an additional dataset is available to help reorder the in-context examples.

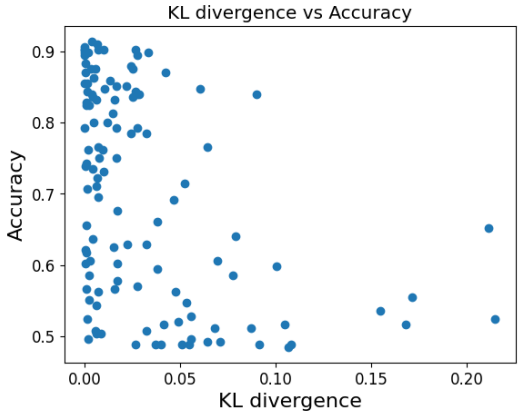
We ask: *Can we select the best in-context example orderings* ***with no labeled*** *data beyond the in-context ones?* We draw inspiration from the idea of learning from label distributions (Yu et al., 2014; Dulac-Arnold et al., 2019), which shows that the prior probability distributions of labels can weakly supervise label predictors. We build upon this insight to improve the quality of in-context predictions, and in particular, to select performant in-context example orderings.

We consider two cases: (a) when we only have in-context examples (FewShot), and (b) when we also have unlabeled examples (FewShotU) and additionally know the prior label distributions (FewShotUP). In all cases, we only use the model's output probability distributions over candidate outputs. These distributions serve as a direct indicator of the model's confidence as well as the bias carried from pretraining and in-context examples.

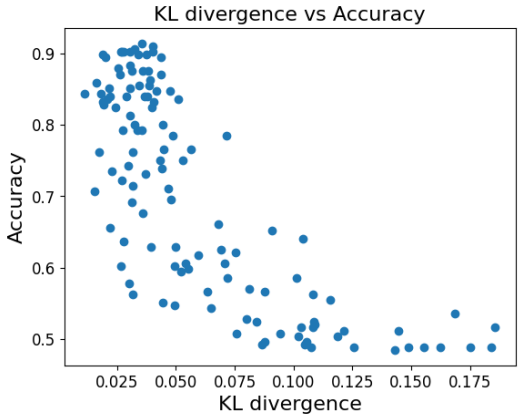
Given a set of in-context examples, we propose to select the best ordering that, on the corpus level, has a probability distribution over candidate labels, such that it is (a) less biased towards certain labels, or, (b) close to a prior label distribution, if known.

Fig. 1 illustrates the two criteria using OPT-1.3B as a backbone language model. Each point corresponds to a certain ordering of a fixed set of in-context examples. Fig. 1a corresponds to case (a). Its x-axis is the KL-divergence between the uniform distribution and the model's probability

Findings of the Association for Computational Linguistics: NAACL 2024, pages 2623–2640



(a) FewShot



(b) FewShotUP

Figure 1: KL-divergence vs accuracy for FewShot and FewShotUP on SST-2 dataset, with a backbone language model OPT-1.3B.

for the null input given in-context examples. This KL-divergence captures the model's bias towards certain labels; smaller values indicate less bias. Fig. 1b corresponds to case (b). Its x-axis is KL-divergence between model's average probability distribution over unlabeled samples and the prior label distribution. In Fig. 1a, accuracy is weakly inversely correlated with KL-divergence, indicating performant orderings tend to be less biased towards certain labels. In Fig. 1b, the negative correlation is stronger, indicating the marginal label probabilities of performant orderings tends to be close to informative priors.

Our approach, *Probability Distribution Ordering* (**PDO**), effectively improves in-context predictions on 13 text classification datasets and 9 language models with 700M–13B parameters. It not only improves the classification accuracy and reduces variance across all datasets and models, but also improves models' confidence calibration, making them more suitable for real-world deployment.

Finally, in analysis experiments, we study how well **PDO** can select in-context examples for a task. Prior work on task-level in-context example selection requires *labeled* development data (Chang and Jia, 2023; Nguyen and Wong, 2023). We show that **PDO** improves task-level example selection, matching CondAcc (Chang and Jia, 2023) without the need for a labeled development set.

2 Background & Notation

We seek to order in-context examples to improve both predictive accuracy and model calibration. This section reviews in-context learning (ICL) and model calibration, and introduce relevant notation. Through the paper, we use the word ordering and permutation interchangeably.

2.1 In-Context Learning

Consider the task of predicting a label $y \in \mathcal{Y}$ for an input $x \in \mathcal{X}$, where $\mathcal{X}$ and $\mathcal{Y}$ denote the textual input space and the label space, respectively. The label y can be *verbalized* into a natural language token. For example, for a sentiment classification task, the input space may be product reviews, and the label space $\mathcal{Y} = \{+, -\}$ may be verbalized to the words `positive` and `negative`.

In-context learning naturally applies to the few-shot setting. We have a small set of k training examples x_i paired with corresponding labels y_i, denoted by $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_k, y_k)\}$. To predict the label for a new example x, we construct an input for a language model by concatenating a certain ordering $\pi(\mathcal{D})$ of these k examples with x. Using this input $(\pi(\mathcal{D}), x)$, the model generates a probability distribution $P(y \mid \pi(\mathcal{D}), x)$ over the label set $\mathcal{Y}$ via the verbalized variants of each label $y \in \mathcal{Y}$. For a classification task, the predicted label for x is therefore $\arg\max_{y \in \mathcal{Y}} P(y \mid \pi(\mathcal{D}), x)$. Since the label for x is predicted using a probability distribution directly obtained from the pretrained language model without further processing, we follow prior works (e.g., Min et al., 2022a) and refer to this approach as the **Direct** method.

Holtzman et al. (2021) proposed an alternative scoring function to **Direct**, where they scored each label $y \in \mathcal{Y}$ for the unseen example x as:

$$\text{PMI}(x, y) = \log \frac{P(y \mid \pi(\mathcal{D}), x)}{P(y \mid \pi(\mathcal{D}), \text{null}).} \quad (1)$$

Here, the score $\text{PMI}(x, y)$ denotes the pointwise mutual information between a label y and input x. In practice, $P(y \mid \pi(\mathcal{D}), \text{null})$ requires simply

setting input x to an empty string. The predicted label is now $\arg\max_{y \in Y} \text{PMI}(x, y)$. The intuition is that a higher PMI value indicates a stronger association between the input x and a candidate label y. We refer to this approach as the **PMI** method.

The above scoring functions are agnostic to in-context example order. However, recent works (e.g., Lu et al., 2022; Wu et al., 2023) have shown that ICL performance is sensitive to the orderings of in-context examples. To address this issue, Lu et al. (2022) assumed a development set and presented a heuristic for ordering the prompts. Their heuristic (and also the approaches we present) can be used with both the **Direct** and **PMI** methods.

2.2 Confidence Calibration

Previous work on ICL have mainly evaluated their results with performance metrics such as accuracy for classification tasks. However, models can grow over-confident about their predictions, which is problematic for deployment. Prior works in calibrating neural networks (e.g. Guo et al., 2017) have argued that a network should provide a calibrated *confidence measure* with its prediction. Specifically, the mean probability of a correct prediction for x should be equal to its average accuracy, e.g., all predictions at 70% confidence level should have an average accuracy of 70%. A model's confidence calibration can be measured by the expected difference between its confidence and accuracy,

$$\mathop{\mathbb{E}}_{\hat{p}}\big[|P(\hat{y} = y|\hat{p} = p) - p|\big]. \tag{2}$$

Here, $\hat{y}$ denotes the predicted label, $\hat{p}$ denotes the associated confidence. In practice, this is often measured by Expected Calibration Error (ECE, Naeini et al., 2015). ECE approximates Eqn. (2) by partitioning predictions into a number of equally-spaced bins and taking a weighted average of the bins' accuracy-confidence difference.

Despite being effective in terms of performance, **PMI** skews the output probability distribution, leading to miscalibrated model outputs. In Eqn. (1), if the denominator $P(y_i \mid \pi(\mathcal{D}), \text{null})$ is already skewed by context $\pi(\mathcal{D})$, it can magnify the skewness of output probability distribution.

Consider the following example: for a binary sentiment classification task and a new input example x, the probability distribution over {`positive`, `negative`} is $(0.7, 0.3)$ for input $(\pi(\mathcal{D}), x)$, and $(0.3, 0.7)$ for input $(\pi(\mathcal{D}), \text{null})$, respectively. Taking softmax over the PMI-adjusted scores according to Eqn. (1) yields a final probability distribution of $(0.92, 0.08)$; whereas the model still predicts x as `positive`, it may have grown over-confident. In Sec. 4, we show empirically that **PMI** leads to higher miscalibration than **Direct**.

3 A Proposal for Ordering Selection

We seek to find an ordering of the k in-context examples $\mathcal{D}$ that has the best predictive performance and leads to calibrated probabilities. To do so, we need to rank the $k!$ permutations and select the highest performing one. However, since we are operating in the few-shot setting, it is important to specify the task information we are allowed to use.

Prior efforts (Lu et al., 2022; Wu et al., 2023; Sorensen et al., 2022) on in-context example ordering operate under different resource settings, making it difficult to perform comprehensive comparisons. We study three settings, in the order of increased information:

1. FewShot: Only a few (typically 8 to 32) labeled in-context examples $\mathcal{D}$ are available.
2. FewShotU: In addition to a few labeled in-context examples $\mathcal{D}$, an *unlabeled* development set X is available.
3. FewShotUP: In addition to in-context examples $\mathcal{D}$ and an unlabeled development set X, we know the prior *probability distribution* $Q(\mathcal{Y})$ over the label space $\mathcal{Y}$.

In the FewShotUP setting, the prior over labels may be determined from prior information (e.g., previous experiments) or a subjective expert assessment (e.g., the probability of a certain disease in the population assessed by a clinician).

The FewShotU setting—where we do not know the prior label distribution—can be seen as a special case of FewShotUP that uses an *uninformative (or flat) prior*, i.e., a uniform prior distribution of $Q(\mathcal{Y}) = \text{Unif}(\mathcal{Y})$ over the development set. We therefore consider two cases separately: (1) when we only have the in-context examples (FewShot); and (2) when we also have an unlabeled set of examples (FewShotU) and additionally know the prior label distribution (FewShotUP).

3.1 FewShot with Only In-Context Examples

In this setting, we have no information about the task beyond the in-context examples. Thus, any label predictor should be maximally uncertain when

presented with no inputs. That is, a good in-context example ordering should lead to the model being unbiased towards certain labels with a null input (e.g., an empty string). We state our first principle:

> PRINCIPLE I: When unlabeled examples are not available, well-ordered in-context examples should lead to the probability distribution of a null input having the minimum KL divergence to a uniform distribution.

PRINCIPLE I can be instantiated as a function that scores an ordering $\pi := \pi(\mathcal{D})$ as follows:

$$\mathcal{L}(\pi) = D_{\text{KL}}\left(P(\mathcal{Y} \mid \pi, \text{null}) \,||\, \text{Unif}(\mathcal{Y})\right). \quad (3)$$

3.2 FewShotU and FewShotUP

We have unlabeled examples in FewShotU, and also the prior label distribution in FewShotUP.

Consider the prior distribution Q over the label space $\mathcal{Y}$. The distribution Q can be obtained by marginalizing out the input space $\mathcal{X}$ as:

$$Q(y) = \sum_{x \in \mathcal{X}} P_{\mathcal{Y}|\mathcal{X}}(y \mid x) P_{\mathcal{X}}(x) \quad (4)$$

$$= E_{\mathcal{X}}\left[P_{\mathcal{Y}|\mathcal{X}}(y \mid x)\right]. \quad (5)$$

The probability $P_{\mathcal{X}}$ in Eqn. (4) denotes the *unknown* distribution over the input space $\mathcal{X}$. The probability $P_{\mathcal{Y}|\mathcal{X}}$ in Eqn. (4) is the label distribution conditioned on the input $\mathcal{X}$. It is the object of study and is provided to us by the language model.

Therefore, if we have access to the unlabeled set X sampled i.i.d. from the natural data distribution, we can approximate the expectation in Eqn. (5) as an empirical mean $\hat{Q}(y) \approx Q(y)$:

$$\hat{Q}(y) = \frac{1}{|X|} \sum_{x \in X} P(y \mid x). \quad (6)$$

Now, suppose we know the prior distribution Q and have access to the *unlabeled* set X. Using X and any ordering $\pi := \pi(\mathcal{D})$ of the in-context examples, we can compute $\hat{Q}$ of Eqn. (6) and measure its difference from Q. Concretely, we define the *observed label distribution* $\hat{P}$ in terms of the model-induced label distributions:

$$\hat{P}(y \mid \pi) = \frac{1}{|X|} \sum_{x \in X} P(y \mid \pi, x). \quad (7)$$

Now, we can state our second principle:

> PRINCIPLE II: Given an unlabeled set of examples and the prior label distribution, well-ordered in-context examples should produce an observed label distribution that matches the prior label distribution.

PRINCIPLE II gives us a function that scores a permutation π as follows:

$$\mathcal{L}(\pi) = D_{\text{KL}}\left(\hat{P}\left(\mathcal{Y}|\pi\right) \,||\, Q(\mathcal{Y})\right) \quad (8)$$

The intuitive interpretation is that we expect the observed label probability $\hat{P}$ on set X to match the prior probability Q. Consequently, we should select an ordering that assigns probabilities labels that are similar to the prior.

As mentioned in Sec. 3.1, if we do not have access to a prior label distribution, we need to assume a uninformative prior and simply set $P(y) = 1/|\mathcal{Y}|$, i.e., uniform distribution $\text{Unif}(\mathcal{Y})$ over the label space $\mathcal{Y}$.

3.3 Selecting a Performant Ordering

The set $\mathcal{D}$ of k in-context examples lead to $k!$ possible orderings. Even for small values of k, we can end up with a prohibitively large number of orderings to score and rank, e.g., with 8 examples, we have to consider $8! = 40,320$ permutations. We propose a simple sample-then-select solution similar to Lu et al. (2022).[1] We first randomly sample K permutations from all possible $k!$ permutations, then rank them as in Eqn. (3) and Eqn. (8):

$$\pi^* = \arg\min_{\pi} \mathcal{L}(\pi) \quad (9)$$

We call our method *Probability Distribution Ordering* (**PDO**). This choice is independent of the **Direct** and **PMI** approaches (which use a given ordering). As a result, we can combine **Direct** and **PMI** approaches with **PDO**.

4 Experiments

Our experiments evaluate the effectiveness of the proposed principles and answer the following research questions:

1. Does **PDO** improve in-context learning accuracy and reduce variance?
2. While vanilla confidence calibration methods (such as temperature scaling) require a labeled development set, can **PDO** better calibrate a model without a labeled development set?

[1] Alternatively, we could seek to parameterize an ordering π; we leave this extension for future work.

4.1 Experimental Setup

We conduct experiments on 13 text classification datasets including binary and multi-label classifications, as well as balanced and imbalanced datasets. The details of these datasets and the prompt templates are in Appx. A. We also use 9 autoregressive language models of varying sizes to demonstrate the robustness of our proposed approach. Our approach falls into the category of *corpus-level* ICL (Wu et al., 2023) (or *task-level* ICL) where we select the best-performing template with or without a validation set and then equally apply this template to all test examples during in-context learning.

As scoring functions and ordering selection are two orthogonal procedures, we experiment with two different scoring approaches—**Direct** and **PMI**—to demonstrate the effectiveness of **PDO**.

4.2 Baselines

We compare against a number of baseline configurations detailed below.

Random. For a given set of in-context examples $\mathcal{D}$, we sample a set of orderings $\{\pi_1(\mathcal{D}), \pi_2(\mathcal{D}), \cdots, \pi_n(\mathcal{D})\}$, perform in-context learning with **Direct** or **PMI**, and average the performance metrics across the set of orderings. We do not perform any ordering selection. This configuration is as an important baseline for FewShot where we do not have an unlabeled set X.

GlobalE and LocalE. Following Lu et al. (2022), **GlobalE** gathers the predicted labels of all examples in unlabeled development set X, and selects the ordering with the minimum KL-divergence between uniform distribution and predicted label distribution. Enforcing a uniform distribution of predicted labels potentially degrades performance on imbalanced datasets. **LocalE** is similar to Eqn. (8), but instead computes KL-divergence between uniform distribution and the probability distribution of each sample in unlabeled development set X:

$$\mathcal{L}(\pi) = \sum_{x \in X} D_{\text{KL}}\big(P(\mathcal{Y}|\pi, x) \,||\, \text{Unif.}(\mathcal{Y})\big). \quad (10)$$

This criterion implicitly encourages the language model to predict a uniform distribution over individual samples, whereas Eqn. (8) minimizes the divergence globally. **LocalE** and **GlobalE** serve as important baselines for FewShotU and FewShotUP, where we assume unlabeled development set X.

Oracle. We select the orderings that lead to the best performance on X, assuming access to an oracle that provides ground truth labels. This configuration serves as an upper bound for all performant ordering selection approaches.

Combining PDO with Direct and PMI. Finally, for each model-dataset pair, as described in Sec. 3, we combine **PDO** with the **Direct** and **PMI** approaches. These configurations are referred to as **PDO-Direct** and **PDO-PMI**, respectively. Table 1, for example, shows the performance of OPT-13B and LLaMA-13B using both **Direct** and **PMI** combined with the three settings of **PDO**.

4.3 Evaluation

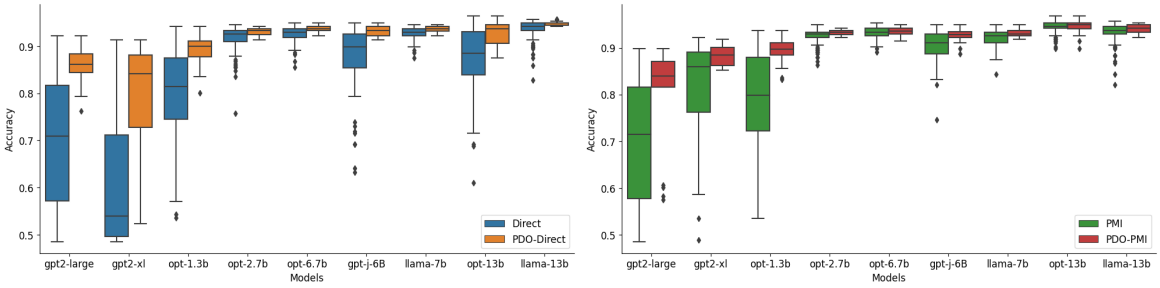
For each dataset, we use 8 in-context examples (shots) and 5 different random seeds (by default), i.e., 5 different sets of uniformly sampled in-context examples. For each set of in-context examples, we randomly sample 24 orderings and report the average accuracy and Expected Calibration Error (ECE) (Naeini et al., 2015) computed with a fixed number of 100 bins. For **GlobalE**, **LocalE**, **PDO** and **Oracle**, we select top-4 orderings out of 24 sampled orderings. The results of **Random** are averaged over $24 \times 5 = 120$ runs while the results of **GlobalE**, **LocalE**, **PDO** and **Oracle** are averaged over $4 \times 5 = 20$ runs. We randomly sample 256 instances from the training set (not overlapping with the 8 in-context examples) to form an unlabeled set X, and use the label distribution as the informative prior probability distribution.

We benchmark the performance of various approaches with 9 autoregressive LLMs with 770M to 13B parameters: GPT2-large (770M), GPT2-xl (1.5B) (Radford et al., 2019), OPT-1.3B, OPT-2.7B, OPT-6.7B, OPT-13B (Zhang et al., 2022b), GPT-J-6B (Wang and Komatsuzaki, 2021), and the more recent LLaMA-7B and LLaMA-13B (Touvron et al., 2023). For the rest of this section, we mainly discuss results on OPT-13B and LLaMA-13B; the results on smaller models show similar trends and we show complete results in Appx. D.[2]

4.4 Results and Analysis

Increasing model size improves ICL classification performance. Figures 2 and 3 show the effect of different model choices on predictive accuracy. As the model size increases, the classification performance also increases whereas the variance de-

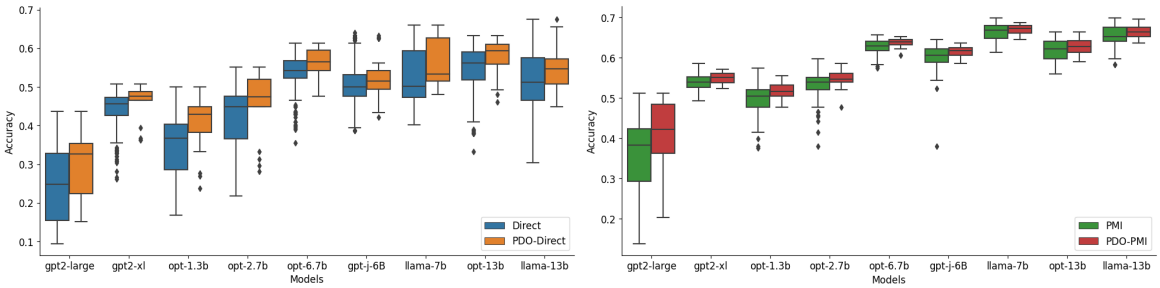
[2]Our code is available at Link to Github.



(a) SST-2 results with **Direct** and **PDO-Direct**.

(b) SST-2 results with **PMI** and **PDO-PMI**.

Figure 2: SST-2 results with different language models.



(a) Yahoo Topics results with **Direct** and **PDO-Direct**.

(b) Yahoo Topics results with **PMI** and **PDO-PMI**.

Figure 3: Yahoo topic results with different language models.

creases. This observation is consistent with prior works (Min et al., 2022b; Lu et al., 2022).

In FewShot where no unlabeled set is available, **PDO** is competitive (Table 1). In three out of four sections in Table 1, **PDO** outperforms the non-selective baselines, and is only slightly worse than **Random** with LLaMA-13B **PMI** (72.0% compared to 72.2%).

When an unlabeled set is available, but the label prior is unknown (FewShotU), **PDO** slightly outperforms **GlobalE** and **LocalE**. For example, for OPT-13B **Direct**, **PDO** achieves on average 66.8% compared to **LocalE**'s 65.2% and **GlobalE**'s 66.1%, whereas for OPT-13B **PMI**, the numbers are 69.2% compared to 67.8% and 67.2%.

Table 1 shows that **PDO** performance consistently improves with more information. For example, the average performance on 13 datasets is 71.2%, 73.1% and 74.0% in the FewShot, FewShotU, and FewShotUP settings respectively. With prior probability distribution known, **PDO** outperforms all baselines, and **PDO-PMI** further improves the classification performance

PDO's performance improvement is consistent across different numbers of in-context examples. Fig. 4 shows an ablation study with varying numbers of in-context examples. We report mean accuracy on 5 topic classification datasets with LLaMA-7B. We observe that as the number of samples increases, the mean accuracy also improves.

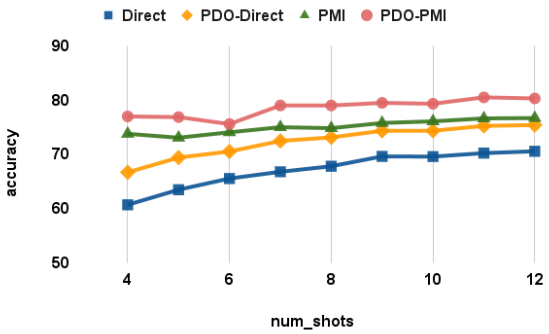


Figure 4: We show the mean accuracy over 5 topic classification datasets across different numbers of in-context training examples (from 4 to 12) under FewShotUP. The backbone LLM is LLaMA-7B. **PDO**'s improvement is consistent with different numbers of samples.

Further, using **PDO** consistently improves performance over the non-selective baselines.

PDO reduces in-context learning miscalibration. From Table 2 we notice that under three different settings, **PDO** can all reduce model miscalibration compared to the baselines. For example, for OPT-13B **Direct** FewShot, **PDO** reports an average ECE result of 17.9% compared to **Random**'s 20.4%; for FewShotU, **PDO** reports an average of 17.1% compared to **LocalE**'s 19.9% and **GlobalE**'s 17.3%. Notably, **Oracle** reports 17.4%, meaning that predictive accuracy and model calibration are not always positively correlated, whereas **PDO** can effectively prevent the model's prediction from being too confident.

Table 1: Accuracy and standard deviation measured in % on OPT-13B and LLaMA-13B respectively, e.g., $65.8_{7.2}$ means a 65.8% accuracy with a 7.2% variance.

Methods	Avg.	Avg. Balanced	Avg. Imbalanced
OPT-13B Direct			
Random	$65.8_{7.2}$	$80.9_{5.0}$	$52.7_{9.0}$
LocalE	$65.2_{6.6}$	$\mathbf{81.4}_{4.5}$	$\mathbf{54.2}_{8.4}$
GlobalE	$\mathbf{66.1}_{5.5}$	$80.4_{2.9}$	$53.8_{7.9}$
PDO (FewShot)	$66.8_{6.5}$	$82.7_{3.6}$	$53.2_{9.0}$
PDO (FewShotU)	$66.8_{6.2}$	$81.5_{4.1}$	$54.2_{8.0}$
PDO (FewShotUP)	$\mathbf{69.6}_{5.6}$	$\mathbf{84.2}_{2.9}$	$\mathbf{57.2}_{7.8}$
Oracle	$70.8_{4.9}$	$84.6_{2.1}$	$59.0_{7.3}$
OPT-13B PMI			
Random	$67.0_{8.1}$	$85.4_{6.9}$	$51.3_{9.1}$
LocalE	$\mathbf{67.8}_{4.5}$	$\mathbf{85.4}_{2.1}$	$\mathbf{52.6}_{6.7}$
GlobalE	$67.2_{3.4}$	$83.4_{1.4}$	$53.3_{5.0}$
PDO (FewShot)	$67.8_{7.6}$	$85.5_{6.9}$	$52.7_{8.1}$
PDO (FewShotU)	$69.2_{4.2}$	$85.5_{2.0}$	$55.3_{6.1}$
PDO (FewShotUP)	$\mathbf{71.1}_{3.5}$	$\mathbf{86.1}_{1.5}$	$\mathbf{58.3}_{5.3}$
Oracle	$73.1_{2.8}$	$87.4_{1.2}$	$60.9_{4.1}$
LLaMA-13B Direct			
Random	$70.8_{6.1}$	$83.4_{3.8}$	$59.8_{8.0}$
LocalE	$70.6_{5.8}$	$83.2_{4.4}$	$59.8_{7.0}$
GlobalE	$\mathbf{72.9}_{5.0}$	$\mathbf{85.1}_{2.8}$	$\mathbf{62.4}_{6.9}$
PDO (FewShot)	$71.2_{6.0}$	$83.9_{3.4}$	$60.3_{8.1}$
PDO (FewShotU)	$73.1_{4.8}$	$\mathbf{85.5}_{2.7}$	$62.5_{5.5}$
PDO (FewShotUP)	$\mathbf{74.0}_{3.3}$	$85.4_{2.7}$	$\mathbf{64.2}_{3.9}$
Oracle	$75.5_{3.2}$	$85.9_{2.5}$	$66.7_{3.7}$
LLaMA-13B PMI			
Random	$72.2_{6.1}$	$87.3_{3.8}$	$59.3_{8.0}$
LocalE	$71.4_{5.0}$	$86.5_{1.9}$	$58.5_{7.6}$
GlobalE	$\mathbf{73.6}_{3.4}$	$\mathbf{88.0}_{1.1}$	$\mathbf{61.3}_{5.4}$
PDO (FewShot)	$72.0_{6.0}$	$87.2_{3.4}$	$59.1_{8.1}$
PDO (FewShotU)	$73.9_{3.6}$	$\mathbf{87.9}_{1.4}$	$61.9_{5.5}$
PDO (FewShotUP)	$\mathbf{75.5}_{2.7}$	$87.8_{1.4}$	$\mathbf{65.0}_{3.9}$
Oracle	$77.4_{2.1}$	$89.0_{0.9}$	$67.4_{3.1}$

Table 2: Expected Calibration Error results (measured in %) on OPT-13B and LLaMA-13B, respectively.

Methods	Avg.	Avg. Balanced	Avg. Imbalanced
OPT-13B Direct			
Random	20.4	13.3	26.4
LocalE	19.9	13.4	25.5
GlobalE	**17.3**	**12.6**	**21.2**
PDO (FewShot)	17.9	13.0	22.2
PDO (FewShotU)	17.1	13.6	20.2
PDO (FewShotUP)	**16.6**	**12.5**	**20.1**
Oracle	17.4	12.7	21.5
OPT-13B PMI			
Random	20.2	18.2	21.8
LocalE	18.0	18.2	17.9
GlobalE	**17.7**	**17.9**	**17.7**
PDO (FewShot)	17.9	**17.4**	20.7
PDO (FewShotU)	**17.2**	18.1	**16.4**
PDO (FewShotUP)	17.6	**17.4**	17.7
Oracle	18.6	17.8	19.4
LLaMA-13B Direct			
Random	16.1	12.6	19.1
LocalE	15.4	13.2	17.4
GlobalE	**14.4**	**12.2**	**16.4**
PDO (FewShot)	14.8	12.7	16.5
PDO (FewShotU)	14.2	12.1	16.0
PDO (FewShotUP)	**13.9**	**12.0**	**15.5**
Oracle	15.4	12.0	18.3
LLaMA-13B PMI			
Random	17.8	16.9	18.7
LocalE	17.4	**16.5**	18.0
GlobalE	**16.9**	16.8	**17.0**
PDO (FewShot)	17.4	16.0	18.5
PDO (FewShotU)	16.8	**15.9**	17.5
PDO (FewShotUP)	**16.5**	**15.9**	**17.2**
Oracle	17.6	16.6	18.5

4.5 In-context example selection

So far, we have evaluated **PDO** for selecting performant orderings. Can the principles that drive it also be used to select in-context examples at the task-level? We follow a similar setup as in Sec. 4.3, i.e. we randomly sample 120 sets of in-context examples, each set consisting of 8 examples, and we sample one permutation from each set. Then we utilize **PDO** to rank them and select the top 20 sets of examples. We compare to CondAcc (Chang and Jia, 2023) for selecting in-context examples at the task level. CondAcc computes each in-context example's influence on *labeled development set* X^*. After determining the top-k influential examples, we place them in the increasing order of influence (Liu et al., 2022).

From Table 3, we see that in 3 out of 4 sections, FewShot improves performance compared to Random. FewShotU and FewShotUP consistently outperform Random in all 4 sections. FewShotUP also matches CondAcc's performance despite not using a labeled development set. These observations show that **PDO**'s can help select performant samples at the task level.

5 Related Work

Brown et al. (2020) first demonstrated that autoregressive LLMs are able to "learn in context". Various strategies have since been proposed to improve in-context learning performance. Wei et al. (2022); Kojima et al. (2022) showed that chain-of-thought prompting can improve LLM's performance on reasoning tasks. A different line of works (Su et al., 2022; Liu et al., 2022; Gao et al., 2020; Lyu et al., 2023) proposes to augment ICL performance via examples/evidence retrieval.

Existing works on prompt engineering for few-shot ICL can be broadly divided into the following

Table 3: Sample selection results (accuracy and standard deviation measured in %) on OPT-13B and LLaMA-13B respectively.

Methods	Avg.	Avg. Balanced	Avg. Imbalanced
OPT-13B Direct			
Random	$65.9_{7.4}$	$80.8_{5.7}$	$53.1_{8.8}$
PDO (FewShot)	$68.3_{5.3}$	$84.2_{2.3}$	$54.7_{7.9}$
PDO (FewShotU)	$72.3_{4.1}$	$86.5_{1.6}$	$60.1_{6.2}$
PDO (FewShotUP)	$73.0_{2.9}$	$86.5_{1.5}$	$61.5_{4.1}$
CondAcc	$72.7_{0.0}$	$86.7_{0.0}$	$61.1_{0.0}$
Oracle	$75.3_{1.5}$	$87.1_{1.0}$	$65.1_{1.9}$
OPT-13B PMI			
Random	$67.1_{5.6}$	$85.2_{2.4}$	$51.5_{8.3}$
PDO (FewShot)	$67.7_{5.2}$	$85.8_{1.9}$	$52.3_{8.1}$
PDO (FewShotU)	$71.6_{3.7}$	$86.8_{1.4}$	$58.6_{5.6}$
PDO (FewShotUP)	$73.1_{3.4}$	$86.7_{1.5}$	$61.4_{5.1}$
CondAcc	$72.9_{0.0}$	$86.5_{0.0}$	$61.1_{0.0}$
Oracle	$75.2_{1.7}$	$88.3_{0.5}$	$64.1_{2.7}$
LLaMA-13B Direct			
Random	$70.2_{6.7}$	$82.4_{4.5}$	$59.6_{8.5}$
PDO (FewShot)	$71.3_{5.6}$	$84.3_{2.8}$	$60.2_{8.1}$
PDO (FewShotU)	$73.6_{4.5}$	$87.4_{1.2}$	$61.8_{7.3}$
PDO (FewShotUP)	$75.4_{2.1}$	$87.2_{1.3}$	$65.4_{2.9}$
CondAcc	$75.0_{0.0}$	$86.5_{0.0}$	$65.5_{0.0}$
Oracle	$77.6_{1.4}$	$87.8_{0.9}$	$68.9_{1.8}$
LLaMA-13B PMI			
Random	$72.1_{5.4}$	$87.1_{2.1}$	$59.3_{8.1}$
PDO (FewShot)	$72.0_{5.5}$	$87.0_{2.2}$	$59.1_{8.2}$
PDO (FewShotU)	$74.1_{3.3}$	$88.3_{1.2}$	$61.8_{5.1}$
PDO (FewShotUP)	$76.1_{2.4}$	$88.1_{1.2}$	$65.7_{3.4}$
CondAcc	$75.8_{0.0}$	$87.5_{0.0}$	$65.6_{0.0}$
Oracle	$78.8_{1.2}$	$89.7_{0.5}$	$69.4_{1.8}$

directions: (1) sample selection (Su et al., 2022; Gao et al., 2020); (2) scoring functions, to propose alternative scoring functions to replace raw probability; examples include PMI (Holtzman et al., 2021), Noisy Channel Classification (Min et al., 2022a), Contextual Calibration (Zhao et al., 2021); (3) prompt instruction/order selection (Lu et al., 2022; Wu et al., 2023; Sorensen et al., 2022).

Prior works have noticed that ICL performance is sensitive to sample choices and ordering. Zhao et al. (2021) noticed that models can be biased (recency bias, majority bias) towards certain labels from in-context examples. Min et al. (2022b) conducted an empirical study to discuss factors important to ICL performance. Lu et al. (2022) pointed out ICL performance can vary from near full fine-tuning to random across different orderings of the same set of examples. Wu et al. (2023) combined retrieval augmentation and leveraged the backbone LLM's confidence to rank and select performant orderings. Our work closely follows Lu et al. (2022)'s approach and generalizes to different resource settings.

Per Wu et al. (2023), our method can be categorized to *corpus-level* approaches, i.e., selecting a universal in-context example ordering for all instances. There are in fact *instance-level* approaches, i.e., selecting performant orderings for each single test instance (Su et al., 2022; Liu et al., 2022). Unlike task-level selection/ordering, instance-level approaches require more computational effort because the selection/ordering needs to be performed for every instance.

Limited works have discussed the confidence calibration problem in ICL. As argued by prior works (Guo et al., 2017; Niculescu-Mizil and Caruana, 2005; Platt et al., 1999), reliable confidence measurement is critical in classification problems, especially high-risk decision scenarios. Existing confidence calibration methods including temperature scaling and Platt scaling mostly require labeled samples to tune hyperparameters while our method do not require labeled samples. Our work can serve as a motivation for future works to add confidence calibration as an evaluation metric for ICL.

Prior works (Yu et al., 2014; Dulac-Arnold et al., 2019) on *Learning from Label Proportions* (LLP) focus on the settings where an instance-level labeling is unavailable. Considering we have N bags, each consisting of n_i examples, we do not have access to each corresponding label, but instead, we have access to the label proportions of each bag. In this case, we can still learn a classifier using the label proportions of all N bags as weak supervision signals (see theoretical proof in Yu et al. (2014); Zhang et al. (2022a)). Dulac-Arnold et al. (2019) discussed choices of empirical loss functions for image classification task and found a classifier could be learned by minimizing KL divergence between predicted label proportions and true label proportions.

6 Conclusions

In this paper, we aim to optimize in-context example ordering to improve ICL performance. We rigorously examine three problem settings based on the availability of labeled examples and propose two principles in selecting performant orderings. Our approach, referred to as the Probability Distribution Ordering (PDO), leverages the model's output probability distributions. Via extensive experiments, we demonstrate that our approach requires a trivial amount of extra computation and outperforms the baselines by improving classification accuracy and reducing model miscalibration.

7 Limitations

Due to limited bandwidth and budget, we only experiment with autoregressive LLMs no larger than 13B. The effectiveness with encoder-decoder models such as those from T5 family (Raffel et al., 2020) are not studied. Examining our findings on commercial language models such as GPT-4 requires further experiments.

The proposed principles require computing the output probability distributions, thus they are not trivial to extend to generation tasks such as Open QA and summarization. We believe a potential future direction is to generalize the proposed approach to natural language generation tasks.

Acknowledgements

We would thank members of UtahNLP and anonymous reviewers for their constructive feedback. This material is based upon work supported by NSF under grants #2134223, #2205418, #2007398, #2217154, #1822877 and #1801446. The support and resources from the Center for High Performance Computing at the University of Utah are gratefully acknowledged.

References

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Ting-Yun Chang and Robin Jia. 2023. Data curation alone can stabilize in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8123–8144, Toronto, Canada. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Gabriel Dulac-Arnold, Neil Zeghidour, Marco Cuturi, Lucas Beyer, and Jean-Philippe Vert. 2019. Deep multi-class learning from label proportions. *arXiv preprint arXiv:1905.12909*.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*.

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.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7038–7051, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, 6(2):167–195.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. 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, Dublin, Ireland. Association for Computational Linguistics.

Xinxi Lyu, Sewon Min, Iz Beltagy, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Z-ICL: Zero-shot in-context learning with pseudo-demonstrations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2304–2317, Toronto, Canada. Association for Computational Linguistics.

Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796.

Sewon Min, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022a. Noisy channel language model prompting for few-shot text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5316–5330, Dublin, Ireland. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022b. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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, pages 2901–2907.

Tai Nguyen and Eric Wong. 2023. In-context example selection with influences. *arXiv preprint arXiv:2302.11042*.

Alexandru Niculescu-Mizil and Rich Caruana. 2005. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632.

R OpenAI. 2023. Gpt-4 technical report. *arXiv*, pages 2303–08774.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 115–124, Ann Arbor, Michigan. Association for Computational Linguistics.

John Platt et al. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew 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, Seattle, Washington, USA. Association for Computational Linguistics.

Taylor Sorensen, Joshua Robinson, Christopher Rytting, Alexander Shaw, Kyle Rogers, Alexia Delorey, Mahmoud Khalil, Nancy Fulda, and David Wingate. 2022. An information-theoretic approach to prompt engineering without ground truth labels. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 819–862, Dublin, Ireland. Association for Computational Linguistics.

Hongjin Su, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, et al. 2022. Selective annotation makes language models better few-shot learners. In *The Eleventh International Conference on Learning Representations*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ellen M Voorhees and Dawn M Tice. 2000. Building a question answering test collection. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 200–207.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Zhiyong Wu, Yaoxiang Wang, Jiacheng Ye, and Lingpeng Kong. 2023. Self-adaptive in-context learning: An information compression perspective for in-context example selection and ordering. In *Proceedings of the 61st Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), pages 1423–1436, Toronto, Canada. Association for Computational Linguistics.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2021. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*.

Felix X Yu, Krzysztof Choromanski, Sanjiv Kumar, Tony Jebara, and Shih-Fu Chang. 2014. On learning from label proportions. *arXiv preprint arXiv:1402.5902*.

Jianxin Zhang, Yutong Wang, and Clay Scott. 2022a. Learning from label proportions by learning with label noise. *Advances in Neural Information Processing Systems*, 35:26933–26942.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022b. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

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 Details on Datasets and Templates

We provide details on the 13 datasets and templates in Table 4 and Table 5, respectively. All datasets licenses are available for public use.

Table 4: Details on datasets.

Dataset	# Classes	Balanced
Sentiment Classification		
SST2 (Socher et al., 2013)	2	✓
SST5 (Socher et al., 2013)	5	✗
CR (Hu and Liu, 2004)	2	✓
MR (Pang and Lee, 2005)	2	✓
financial_phrasebank (Malo et al., 2014)	3	✗
Topic Classification		
AG News (Zhang et al., 2015)	4	✓
TREC (Voorhees and Tice, 2000)	6	✓
Yahoo Topics (Zhang et al., 2015)	10	✗
Dbpedia (Lehmann et al., 2015)	14	✗
Subj (Pang and Lee, 2005)	2	✗
Toxicity Detection		
Tweet Offensive (Barbieri et al., 2020)	2	✗
Tweet Irony (Barbieri et al., 2020)	2	✗
Tweet Hate (Barbieri et al., 2020)	2	✗

B Computational Complexity of PDO

For a fixed set of K in-context examples, we sample up to k permutations from $K!$ possible permutations. The random baseline does not incur any additional computation cost.

In FewShot (no unlabeled development set available), we perform in total k forward passes to select the permutations that are less biased towards certain labels.

For FewShotU and FewShotUP, for each permutation, we perform forward passes on all instances in the development set X, therefore requiring in total $k \times |X|$ forward passes, where $|X|$ denotes the size of the unlabeled development set. The computational cost for FewShotU and FewShotUP is the same as GlobalE and LocalE baselines (Lu et al., 2022).

C Extended Related Work

Existing works on selecting in-context examples can be categorized into two classes: (i) instance-level example selection, i.e. to select a set of in-context examples (and its ordering) for each instance in the test set, and (ii) corpus-level/task-level example selection, i.e. to select a set of high quality in-context examples and apply them equality to all test instances. Wu et al. (2023) argue that instance-level example selection can achieve high performance. On the other hand, instance-level example selection incurs additional computational costs at the inference time, and its performance suffers from potential degradation when the size of high-quality annotated examples is small. Only a few prior works focus on task-level example selection. For example, Chang and Jia (2023) propose to utilize influence functions (Koh and Liang, 2017) to calculate the influence score for each individual instance in the training set, and select the most influential ones as in-context examples. Their experiment results show that by carefully selecting in-context examples with high influence scores, the performance of in-context learning can be improved while the variance can be reduced. A concurrent work (Nguyen and Wong, 2023) show that increasing the number of influential examples can further improve performance. VoteK (Su et al., 2022) can be seen as a special combination of instance-level and task-level ICL, where first in the task-level, a relatively large set of unlabeled instances is selected to be annotated, then at the inference time, for each test instance, a specific

Table 5: Templates and label tokens. We use minimum templates and single token labels similar to (Lu et al., 2022; Wu et al., 2023).

Dataset	Template	Label Tokens
SST2, CR, MR	Review: [INPUT]\nSentiment: [LABEL]	positive, negative
SST5	Review: [INPUT]\nSentiment: [LABEL]	terrible, bad, okay, good, great
financial_phrasebank	News: [INPUT]\nSentiment: [LABEL]	positive, negative
Subj	Input: [INPUT]\nType: [LABEL]	subjective, objective
AG News	Input: [INPUT]\nType: [LABEL]	sports, business, world, technology
TREC	Question: [INPUT]\nType: [LABEL]	description, entity, expression location, number, human
Dbpedia	Input: [INPUT]\nType: [LABEL]	company, school, artist, athlete politics, transportation, building nature, village, animal, plant album, film, book
Yahoo Topics	Question: [INPUT]\nTopic: [LABEL]	culture, science, health, politics education, electronics, entertainment business, sports, relationship
Tweet Irony	Tweet: [INPUT]\nLabel: [LABEL]	ironic, neutral
Tweet Hate	Tweet: [INPUT]\nLabel: [LABEL]	hate, neutral
Tweet Offensive	Tweet: [INPUT]\nLabel: [LABEL]	offensive, neutral

small set of annotated examples are selected as in-context examples (instance level). In Sec. 4.5 we show that PDO can achieve comparable performance to a task-level example selection method CondAcc (Chang and Jia, 2023), while requiring no labeled development set X^* to compute the influence scores.

D Complete Results

Table 6: Complete classification results (measured by accuracy), Part 1.

	SST2	CR	MR	SUBJ	SST5	AGNews	TREC	Yahoo Topics	Dbpedia	FPB	Tweet Offensive	Tweet Irony	Tweet Hate
GPT2-large Direct													
Random	0.698	0.606	0.690	0.541	0.332	0.491	0.386	0.247	0.444	0.461	0.503	0.481	0.461
LocalE	0.798	0.785	0.837	0.656	0.369	0.550	0.402	0.260	0.421	0.595	0.524	0.501	0.478
GlobalE	0.812	0.784	0.821	0.632	0.376	0.558	0.423	0.275	0.369	0.544	0.475	0.492	0.484
PDO (FewShot)	0.677	0.754	0.721	0.555	0.364	0.525	0.394	0.275	0.475	0.562	0.458	0.497	0.479
PDO (FewShotU)	0.797	0.784	0.837	0.655	0.369	0.550	0.439	0.263	0.421	0.594	0.522	0.500	0.477
PDO (FewShotUP)	0.859	0.727	0.841	0.666	0.381	0.609	0.439	0.291	0.565	0.618	0.624	0.508	0.497
Oracle	0.867	0.794	0.854	0.670	0.429	0.618	0.456	0.317	0.575	0.647	0.637	0.528	0.511
GPT2-large PMI													
Random	0.694	0.727	0.718	0.623	0.375	0.658	0.423	0.362	0.766	0.605	0.419	0.500	0.508
LocalE	0.768	0.733	0.789	0.735	0.425	0.659	0.464	0.364	0.709	0.637	0.493	0.495	0.509
GlobalE	0.755	0.712	0.799	0.735	0.411	0.652	0.492	0.370	0.685	0.612	0.501	0.500	0.475
PDO (FewShot)	0.658	0.796	0.723	0.599	0.386	0.627	0.430	0.372	0.763	0.635	0.408	0.504	0.516
PDO (FewShotU)	0.769	0.733	0.790	0.730	0.425	0.660	0.488	0.364	0.709	0.637	0.494	0.495	0.513
PDO (FewShotUP)	0.801	0.811	0.832	0.746	0.412	0.730	0.488	0.407	0.812	0.721	0.584	0.499	0.556
Oracle	0.811	0.831	0.851	0.766	0.457	0.752	0.517	0.436	0.829	0.775	0.589	0.540	0.581
GPT2-xl Direct													
Random	0.603	0.576	0.582	0.600	0.352	0.674	0.425	0.434	0.706	0.483	0.404	0.515	0.434
LocalE	0.727	0.661	0.671	0.711	0.371	0.660	0.415	0.430	0.673	0.555	0.529	0.504	0.466
GlobalE	0.719	0.671	0.705	0.689	0.362	0.656	0.422	0.423	0.671	0.533	0.489	0.495	0.463
PDO (FewShot)	0.752	0.658	0.647	0.653	0.361	0.719	0.441	0.447	0.715	0.468	0.429	0.491	0.431
PDO (FewShotU)	0.726	0.661	0.672	0.710	0.372	0.660	0.452	0.432	0.673	0.554	0.529	0.504	0.466
PDO (FewShotUP)	0.788	0.683	0.704	0.708	0.425	0.759	0.452	0.457	0.769	0.638	0.592	0.517	0.495
Oracle	0.790	0.700	0.714	0.733	0.464	0.768	0.489	0.464	0.778	0.650	0.595	0.545	0.515
GPT2-xl PMI													
Random	0.818	0.801	0.770	0.617	0.267	0.799	0.473	0.539	0.818	0.441	0.414	0.468	0.451
LocalE	0.828	0.813	0.784	0.718	0.318	0.764	0.480	0.542	0.804	0.526	0.483	0.469	0.461
GlobalE	0.811	0.825	0.792	0.733	0.285	0.766	0.505	0.530	0.801	0.519	0.490	0.441	0.460
PDO (FewShot)	0.798	0.780	0.736	0.654	0.282	0.805	0.500	0.539	0.833	0.479	0.429	0.474	0.429
PDO (FewShotU)	0.825	0.816	0.786	0.719	0.318	0.763	0.523	0.543	0.804	0.525	0.482	0.464	0.463
PDO (FewShotUP)	0.884	0.878	0.864	0.726	0.327	0.825	0.523	0.551	0.840	0.555	0.515	0.481	0.534
Oracle	0.896	0.887	0.867	0.733	0.345	0.839	0.551	0.562	0.847	0.561	0.533	0.505	0.541
OPT-1.3B Direct													
Random	0.800	0.879	0.794	0.600	0.404	0.725	0.384	0.348	0.809	0.686	0.504	0.471	0.535
LocalE	0.829	0.902	0.833	0.682	0.391	0.732	0.405	0.360	0.791	0.707	0.554	0.481	0.488
GlobalE	0.828	0.894	0.821	0.687	0.395	0.727	0.409	0.355	0.745	0.705	0.511	0.480	0.471
PDO (FewShot)	0.826	0.899	0.846	0.585	0.400	0.729	0.378	0.360	0.810	0.713	0.533	0.478	0.528
PDO (FewShotU)	0.830	0.902	0.835	0.682	0.391	0.731	0.419	0.361	0.791	0.706	0.553	0.484	0.491
PDO (FewShotUP)	0.891	0.913	0.879	0.699	0.427	0.785	0.419	0.407	0.830	0.714	0.667	0.487	0.566
Oracle	0.901	0.918	0.883	0.721	0.479	0.791	0.455	0.409	0.835	0.756	0.683	0.506	0.590
OPT-1.3B PMI													
Random	0.791	0.907	0.868	0.576	0.364	0.752	0.418	0.496	0.871	0.639	0.523	0.504	0.497
LocalE	0.825	0.897	0.846	0.699	0.385	0.776	0.419	0.478	0.872	0.701	0.520	0.492	0.481
GlobalE	0.805	0.899	0.852	0.695	0.390	0.766	0.453	0.470	0.868	0.701	0.505	0.497	0.481
PDO (FewShot)	0.825	0.904	0.855	0.588	0.330	0.764	0.425	0.482	0.873	0.699	0.498	0.501	0.496
PDO (FewShotU)	0.811	0.897	0.846	0.699	0.382	0.774	0.463	0.479	0.872	0.700	0.513	0.500	0.486
PDO (FewShotUP)	0.896	0.901	0.882	0.706	0.428	0.791	0.463	0.519	0.882	0.695	0.676	0.519	0.567
Oracle	0.901	0.927	0.897	0.723	0.442	0.817	0.513	0.535	0.890	0.745	0.682	0.547	0.592

Table 7: Complete classification results (measured by accuracy), Part 2.

	SST2	CR	MR	SUBJ	SST5	AGNews	TREC	Yahoo Topics	Dbpedia	FPB	Tweet Offensive	Tweet Irony	Tweet Hate
OPT-2.7B Direct													
Random	0.917	0.895	0.891	0.653	0.455	0.747	0.418	0.420	0.844	0.626	0.563	0.537	0.516
LocalE	0.922	0.905	0.903	0.783	0.449	0.767	0.432	0.410	0.823	0.719	0.545	0.542	0.543
GlobalE	0.929	0.925	0.919	0.803	0.475	0.811	0.457	0.461	0.859	0.791	0.564	0.539	0.567
PDO (FewShot)	0.920	0.919	0.899	0.710	0.479	0.725	0.418	0.433	0.840	0.551	0.561	0.539	0.539
PDO (FewShotU)	0.922	0.905	0.903	0.782	0.449	0.768	0.467	0.407	0.823	0.719	0.546	0.539	0.544
PDO (FewShotUP)	0.929	0.925	0.917	0.800	0.493	0.816	0.467	0.458	0.856	0.786	0.673	0.547	0.588
Oracle	0.937	0.929	0.921	0.809	0.508	0.818	0.488	0.464	0.864	0.834	0.684	0.553	0.611
OPT-2.7B PMI													
Random	0.925	0.921	0.902	0.693	0.399	0.728	0.409	0.532	0.865	0.567	0.564	0.532	0.547
LocalE	0.927	0.918	0.906	0.760	0.448	0.778	0.431	0.523	0.855	0.669	0.585	0.550	0.547
GlobalE	0.912	0.911	0.922	0.751	0.433	0.761	0.442	0.52	0.851	0.667	0.571	0.551	0.540
PDO (FewShot)	0.917	0.922	0.900	0.715	0.416	0.776	0.394	0.543	0.877	0.548	0.575	0.539	0.555
PDO (FewShotU)	0.927	0.918	0.907	0.759	0.446	0.779	0.447	0.522	0.855	0.679	0.585	0.548	0.547
PDO (FewShotUP)	0.932	0.927	0.911	0.789	0.462	0.790	0.447	0.546	0.892	0.741	0.659	0.558	0.605
Oracle	0.939	0.937	0.922	0.815	0.475	0.803	0.521	0.563	0.898	0.762	0.671	0.573	0.636
OPT-6.7B Direct													
Random	0.924	0.871	0.911	0.690	0.450	0.703	0.448	0.532	0.868	0.742	0.620	0.525	0.501
LocalE	0.927	0.867	0.909	0.773	0.450	0.697	0.439	0.530	0.855	0.784	0.622	0.531	0.521
GlobalE	0.938	0.915	0.915	0.788	0.452	0.771	0.437	0.557	0.878	0.768	0.613	0.538	0.533
PDO (FewShot)	0.927	0.895	0.911	0.637	0.468	0.703	0.464	0.549	0.872	0.750	0.557	0.514	0.480
PDO (FewShotU)	0.928	0.867	0.909	0.773	0.450	0.697	0.467	0.532	0.855	0.784	0.623	0.530	0.520
PDO (FewShotUP)	0.938	0.911	0.921	0.786	0.482	0.769	0.467	0.561	0.882	0.780	0.682	0.538	0.568
Oracle	0.942	0.917	0.927	0.808	0.516	0.773	0.517	0.565	0.886	0.832	0.687	0.546	0.583
OPT-6.7B PMI													
Random	0.932	0.897	0.914	0.638	0.415	0.823	0.418	0.627	0.874	0.760	0.459	0.504	0.489
LocalE	0.927	0.897	0.913	0.749	0.433	0.830	0.455	0.627	0.870	0.807	0.602	0.540	0.526
GlobalE	0.932	0.875	0.901	0.744	0.432	0.811	0.471	0.620	0.861	0.801	0.591	0.525	0.512
PDO (FewShot)	0.929	0.901	0.910	0.644	0.419	0.804	0.451	0.630	0.880	0.795	0.434	0.510	0.450
PDO (FewShotU)	0.927	0.896	0.914	0.749	0.433	0.830	0.477	0.626	0.870	0.808	0.603	0.535	0.531
PDO (FewShotUP)	0.935	0.912	0.918	0.758	0.448	0.853	0.477	0.638	0.889	0.845	0.647	0.553	0.584
Oracle	0.944	0.919	0.929	0.768	0.465	0.870	0.513	0.646	0.896	0.869	0.651	0.565	0.597
GPT-J-6B Direct													
Random	0.879	0.831	0.876	0.728	0.447	0.795	0.498	0.509	0.851	0.535	0.563	0.486	0.469
LocalE	0.892	0.842	0.880	0.784	0.445	0.804	0.507	0.510	0.835	0.516	0.607	0.522	0.485
GlobalE	0.928	0.883	0.905	0.794	0.438	0.813	0.536	0.524	0.866	0.505	0.593	0.521	0.494
PDO (FewShot)	0.922	0.842	0.902	0.775	0.460	0.800	0.485	0.511	0.860	0.497	0.488	0.520	0.457
PDO (FewShotU)	0.892	0.842	0.880	0.785	0.445	0.805	0.534	0.513	0.836	0.516	0.606	0.523	0.484
PDO (FewShotUP)	0.931	0.890	0.903	0.799	0.472	0.816	0.534	0.525	0.863	0.577	0.684	0.522	0.501
Oracle	0.935	0.893	0.908	0.823	0.489	0.820	0.548	0.533	0.871	0.607	0.694	0.546	0.510
GPT-J-6B PMI													
Random	0.903	0.848	0.900	0.765	0.450	0.747	0.589	0.601	0.912	0.461	0.407	0.521	0.513
LocalE	0.903	0.848	0.898	0.752	0.445	0.788	0.602	0.608	0.908	0.468	0.581	0.535	0.580
GlobalE	0.895	0.831	0.879	0.773	0.436	0.799	0.629	0.590	0.891	0.459	0.581	0.532	0.551
PDO (FewShot)	0.919	0.848	0.900	0.780	0.450	0.773	0.574	0.606	0.908	0.471	0.416	0.538	0.507
PDO (FewShotU)	0.903	0.848	0.899	0.753	0.445	0.792	0.644	0.608	0.910	0.471	0.591	0.540	0.568
PDO (FewShotUP)	0.926	0.883	0.906	0.805	0.466	0.830	0.644	0.616	0.922	0.500	0.614	0.547	0.593
Oracle	0.957	0.883	0.912	0.826	0.488	0.834	0.655	0.625	0.931	0.522	0.620	0.561	0.604

Table 8: Complete classification results (measured by accuracy), Part 3.

	SST2	CR	MR	SUBJ	SST5	AGNews	TREC	Yahoo Topics	Dbpedia	FPB	Tweet Offensive	Tweet Irony	Tweet Hate
LLaMA-7B Direct													
Random	0.928	0.901	0.913	0.594	0.471	0.849	0.607	0.527	0.813	0.648	0.683	0.538	0.558
LocalE	0.927	0.901	0.911	0.627	0.467	0.853	0.589	0.530	0.809	0.656	0.684	0.540	0.560
GlobalE	0.932	0.917	0.927	0.732	0.467	0.869	0.639	0.559	0.849	0.689	0.679	0.536	0.572
PDO (FewShot)	0.931	0.911	0.910	0.689	0.455	0.846	0.587	0.552	0.815	0.678	0.679	0.538	0.548
PDO (FewShotU)	0.933	0.915	0.926	0.733	0.459	0.865	0.642	0.564	0.849	0.687	0.679	0.540	0.572
PDO (FewShotUP)	0.936	0.915	0.927	0.733	0.487	0.865	0.656	0.565	0.850	0.668	0.684	0.541	0.591
Oracle	0.940	0.922	0.930	0.735	0.507	0.874	0.668	0.569	0.854	0.690	0.684	0.542	0.598
LLaMA-7B PMI													
Random	0.920	0.927	0.889	0.728	0.426	0.855	0.591	0.664	0.902	0.779	0.487	0.488	0.498
LocalE	0.922	0.918	0.904	0.652	0.434	0.855	0.637	0.658	0.896	0.782	0.567	0.491	0.485
GlobalE	0.936	0.932	0.914	0.808	0.453	0.878	0.649	0.670	0.930	0.762	0.564	0.518	0.548
PDO (FewShot)	0.925	0.921	0.917	0.720	0.424	0.855	0.614	0.670	0.916	0.787	0.599	0.488	0.518
PDO (FewShotU)	0.931	0.917	0.912	0.800	0.454	0.870	0.659	0.666	0.925	0.795	0.561	0.516	0.552
PDO (FewShotUP)	0.931	0.917	0.912	0.803	0.447	0.866	0.681	0.670	0.929	0.787	0.652	0.518	0.579
Oracle	0.940	0.939	0.916	0.823	0.470	0.881	0.699	0.684	0.934	0.823	0.656	0.534	0.591
OPT-13B Direct													
Random	0.871	0.905	0.868	0.672	0.485	0.835	0.393	0.541	0.837	0.684	0.494	0.476	0.558
LocalE	0.884	0.907	0.861	0.628	0.481	0.857	0.398	0.540	0.838	0.680	0.454	0.466	0.484
GlobalE	0.893	0.891	0.847	0.695	0.457	0.840	0.392	0.544	0.814	0.674	0.527	0.498	0.524
PDO (FewShot)	0.925	0.917	0.879	0.694	0.472	0.836	0.400	0.560	0.842	0.698	0.487	0.480	0.548
PDO (FewShotU)	0.897	0.904	0.860	0.707	0.479	0.849	0.403	0.553	0.828	0.692	0.501	0.498	0.515
PDO (FewShotUP)	0.931	0.922	0.899	0.792	0.493	0.866	0.435	0.575	0.857	0.671	0.586	0.503	0.525
Oracle	0.935	0.925	0.901	0.813	0.512	0.872	0.457	0.580	0.861	0.715	0.590	0.506	0.534
OPT-13B PMI													
Random	0.944	0.916	0.918	0.647	0.430	0.835	0.399	0.618	0.893	0.579	0.534	0.488	0.498
LocalE	0.941	0.915	0.911	0.650	0.443	0.842	0.422	0.623	0.895	0.630	0.552	0.475	0.513
GlobalE	0.913	0.895	0.889	0.656	0.441	0.837	0.402	0.610	0.869	0.627	0.492	0.534	0.583
PDO (FewShot)	0.949	0.911	0.913	0.692	0.418	0.838	0.410	0.623	0.895	0.640	0.533	0.488	0.518
PDO (FewShotU)	0.942	0.916	0.911	0.688	0.459	0.843	0.432	0.623	0.895	0.649	0.578	0.505	0.558
PDO (FewShotUP)	0.944	0.922	0.911	0.747	0.462	0.857	0.448	0.628	0.902	0.656	0.652	0.536	0.582
Oracle	0.957	0.931	0.932	0.790	0.481	0.873	0.473	0.639	0.910	0.694	0.667	0.549	0.613
LLaMA-13B Direct													
Random	0.936	0.894	0.927	0.724	0.495	0.848	0.628	0.519	0.885	0.675	0.646	0.510	0.511
LocalE	0.926	0.893	0.917	0.729	0.482	0.853	0.624	0.510	0.892	0.660	0.661	0.511	0.521
GlobalE	0.943	0.920	0.943	0.811	0.490	0.864	0.670	0.551	0.887	0.735	0.562	0.539	0.560
PDO (FewShot)	0.936	0.894	0.930	0.751	0.492	0.849	0.608	0.526	0.898	0.703	0.624	0.525	0.521
PDO (FewShotU)	0.946	0.918	0.943	0.804	0.488	0.862	0.679	0.551	0.909	0.735	0.560	0.545	0.564
PDO (FewShotUP)	0.948	0.917	0.940	0.788	0.502	0.860	0.697	0.552	0.909	0.719	0.684	0.540	0.565
Oracle	0.951	0.921	0.945	0.844	0.512	0.867	0.706	0.559	0.911	0.768	0.685	0.555	0.594
LLaMA-13B PMI													
Random	0.932	0.928	0.928	0.767	0.426	0.851	0.634	0.655	0.942	0.686	0.629	0.511	0.501
LocalE	0.932	0.913	0.926	0.681	0.435	0.863	0.622	0.653	0.905	0.713	0.638	0.494	0.514
GlobalE	0.947	0.938	0.940	0.840	0.459	0.872	0.655	0.670	0.921	0.697	0.552	0.541	0.545
PDO (FewShot)	0.934	0.928	0.930	0.755	0.432	0.847	0.622	0.647	0.943	0.739	0.618	0.504	0.464
PDO (FewShotU)	0.943	0.921	0.933	0.825	0.449	0.868	0.650	0.660	0.946	0.740	0.569	0.542	0.558
PDO (FewShotUP)	0.942	0.921	0.934	0.823	0.460	0.861	0.701	0.665	0.946	0.766	0.681	0.546	0.574
Oracle	0.949	0.942	0.941	0.858	0.477	0.877	0.720	0.678	0.955	0.801	0.688	0.576	0.600

Table 9: Complete confidence calibration results (measured by ECE), Part 1.

	SST2	CR	MR	SUBJ	SST5	AGNews	TREC	Yahoo Topics	Dbpedia	FPB	Tweet Offensive	Tweet Irony	Tweet Hate
GPT2-large Direct													
Random	0.242	0.314	0.240	0.254	0.273	0.268	0.325	0.246	0.263	0.319	0.289	0.265	0.292
LocalE	0.250	0.221	0.232	0.193	0.209	0.223	0.272	0.208	0.225	0.257	0.164	0.129	0.175
GlobalE	0.278	0.224	0.235	0.197	0.180	0.216	0.27	0.216	0.19	0.256	0.179	0.136	0.204
PDO (FewShot)	0.260	0.226	0.208	0.264	0.224	0.237	0.301	0.215	0.213	0.242	0.241	0.163	0.185
PDO (FewShotU)	0.247	0.221	0.23	0.195	0.21	0.223	0.275	0.208	0.225	0.258	0.160	0.128	0.175
PDO (FewShotUP)	0.275	0.223	0.233	0.197	0.202	0.220	0.275	0.231	0.195	0.243	0.125	0.129	0.183
Oracle	0.028	0.127	0.023	0.107	0.054	0.140	0.097	0.085	0.166	0.157	0.08	0.027	0.082
GPT2-large PMI													
Random	0.243	0.224	0.221	0.205	0.150	0.225	0.340	0.262	0.222	0.238	0.245	0.141	0.132
LocalE	0.250	0.221	0.232	0.193	0.209	0.223	0.272	0.208	0.225	0.257	0.164	0.129	0.175
GlobalE	0.278	0.224	0.235	0.197	0.180	0.216	0.270	0.216	0.19	0.256	0.179	0.136	0.204
PDO (FewShot)	0.264	0.217	0.218	0.240	0.144	0.209	0.278	0.235	0.217	0.254	0.226	0.142	0.122
PDO (FewShotU)	0.247	0.221	0.230	0.195	0.210	0.223	0.275	0.208	0.225	0.258	0.160	0.128	0.175
PDO (FewShotUP)	0.261	0.228	0.232	0.197	0.157	0.221	0.233	0.21	0.221	0.262	0.136	0.108	0.109
Oracle	0.272	0.224	0.236	0.213	0.187	0.234	0.254	0.235	0.211	0.314	0.138	0.124	0.105
GPT2-xl Direct													
Random	0.277	0.289	0.264	0.192	0.192	0.179	0.27	0.230	0.168	0.258	0.348	0.199	0.265
LocalE	0.215	0.23	0.205	0.173	0.151	0.179	0.225	0.213	0.177	0.216	0.162	0.121	0.158
GlobalE	0.237	0.198	0.196	0.172	0.138	0.168	0.236	0.213	0.16	0.217	0.162	0.125	0.160
PDO (FewShot)	0.240	0.216	0.209	0.158	0.177	0.169	0.239	0.216	0.165	0.248	0.289	0.190	0.207
PDO (FewShotU)	0.213	0.233	0.205	0.170	0.148	0.179	0.242	0.213	0.177	0.217	0.163	0.122	0.157
PDO (FewShotUP)	0.239	0.212	0.200	0.165	0.157	0.164	0.242	0.212	0.159	0.239	0.165	0.121	0.151
Oracle	0.237	0.199	0.197	0.182	0.136	0.174	0.247	0.213	0.16	0.233	0.173	0.196	0.187
GPT2-xl PMI													
Random	0.248	0.218	0.214	0.191	0.228	0.184	0.304	0.345	0.197	0.263	0.297	0.234	0.172
LocalE	0.258	0.233	0.226	0.172	0.149	0.177	0.237	0.322	0.198	0.231	0.202	0.153	0.123
GlobalE	0.265	0.224	0.245	0.181	0.162	0.199	0.252	0.330	0.211	0.234	0.21	0.155	0.149
PDO (FewShot)	0.252	0.221	0.187	0.159	0.174	0.165	0.271	0.333	0.173	0.255	0.28	0.186	0.196
PDO (FewShotU)	0.255	0.234	0.224	0.174	0.147	0.177	0.246	0.322	0.198	0.233	0.204	0.159	0.125
PDO (FewShotUP)	0.284	0.239	0.241	0.176	0.153	0.173	0.246	0.325	0.166	0.236	0.204	0.153	0.115
Oracle	0.287	0.248	0.241	0.182	0.173	0.185	0.244	0.333	0.175	0.235	0.200	0.183	0.128
OPT-1.3B Direct													
Random	0.163	0.126	0.147	0.217	0.240	0.167	0.318	0.229	0.133	0.220	0.270	0.349	0.241
LocalE	0.165	0.156	0.159	0.189	0.281	0.173	0.256	0.194	0.139	0.214	0.157	0.181	0.165
GlobalE	0.177	0.155	0.148	0.194	0.272	0.187	0.271	0.204	0.148	0.226	0.155	0.192	0.181
PDO (FewShot)	0.157	0.135	0.152	0.209	0.213	0.171	0.319	0.216	0.130	0.210	0.193	0.225	0.169
PDO (FewShotU)	0.165	0.148	0.148	0.189	0.284	0.169	0.269	0.194	0.139	0.213	0.143	0.175	0.167
PDO (FewShotUP)	0.169	0.133	0.147	0.192	0.205	0.169	0.269	0.21	0.129	0.195	0.121	0.181	0.152
Oracle	0.169	0.132	0.139	0.200	0.212	0.162	0.270	0.205	0.127	0.227	0.157	0.270	0.244
OPT-1.3B PMI													
Random	0.180	0.171	0.173	0.230	0.284	0.196	0.460	0.294	0.143	0.206	0.192	0.168	0.167
LocalE	0.187	0.181	0.183	0.155	0.215	0.187	0.342	0.244	0.151	0.237	0.144	0.112	0.127
GlobalE	0.182	0.192	0.185	0.149	0.230	0.195	0.355	0.241	0.159	0.244	0.145	0.112	0.135
PDO (FewShot)	0.181	0.175	0.177	0.194	0.300	0.173	0.370	0.269	0.131	0.231	0.190	0.161	0.165
PDO (FewShotU)	0.179	0.185	0.187	0.152	0.218	0.189	0.345	0.244	0.151	0.238	0.140	0.103	0.127
PDO (FewShotUP)	0.195	0.171	0.180	0.154	0.225	0.180	0.345	0.265	0.127	0.204	0.126	0.112	0.128
Oracle	0.193	0.178	0.172	0.159	0.213	0.190	0.386	0.297	0.127	0.262	0.139	0.169	0.139

Table 10: Complete confidence calibration results (measured by ECE), Part 2.

	SST2	CR	MR	SUBJ	SST5	AGNews	TREC	Yahoo Topics	Dbpedia	FPB	Tweet Offensive	Tweet Irony	Tweet Hate
OPT-2.7B Direct													
Random	0.204	0.131	0.153	0.226	0.218	0.177	0.276	0.230	0.113	0.269	0.260	0.313	0.251
LocalE	0.234	0.148	0.174	0.195	0.197	0.191	0.232	0.211	0.126	0.282	0.174	0.163	0.140
GlobalE	0.216	0.145	0.165	0.207	0.188	0.189	0.226	0.218	0.106	0.280	0.166	0.178	0.137
PDO (FewShot)	0.211	0.143	0.158	0.199	0.203	0.182	0.278	0.228	0.116	0.282	0.192	0.189	0.158
PDO (FewShotU)	0.235	0.148	0.173	0.199	0.196	0.193	0.237	0.211	0.126	0.282	0.179	0.160	0.142
PDO (FewShotUP)	0.223	0.145	0.165	0.206	0.199	0.184	0.237	0.224	0.105	0.260	0.143	0.163	0.141
Oracle	0.218	0.146	0.160	0.209	0.213	0.185	0.235	0.222	0.107	0.305	0.204	0.202	0.144
OPT-2.7B PMI													
Random	0.229	0.176	0.185	0.200	0.247	0.208	0.510	0.322	0.156	0.248	0.182	0.159	0.147
LocalE	0.250	0.190	0.201	0.200	0.184	0.187	0.403	0.281	0.177	0.250	0.137	0.106	0.112
GlobalE	0.255	0.191	0.199	0.205	0.189	0.201	0.405	0.302	0.195	0.255	0.145	0.131	0.135
PDO (FewShot)	0.227	0.184	0.187	0.198	0.204	0.177	0.485	0.299	0.127	0.267	0.159	0.143	0.144
PDO (FewShotU)	0.251	0.186	0.202	0.198	0.182	0.192	0.393	0.281	0.177	0.243	0.134	0.112	0.113
PDO (FewShotUP)	0.242	0.186	0.191	0.214	0.187	0.181	0.393	0.300	0.119	0.258	0.133	0.106	0.123
Oracle	0.234	0.181	0.194	0.217	0.218	0.184	0.425	0.310	0.125	0.274	0.142	0.117	0.120
OPT-6.7B Direct													
Random	0.142	0.120	0.111	0.199	0.253	0.174	0.269	0.201	0.103	0.237	0.195	0.207	0.238
LocalE	0.162	0.128	0.124	0.185	0.232	0.171	0.246	0.190	0.108	0.280	0.127	0.107	0.099
GlobalE	0.155	0.126	0.120	0.192	0.228	0.150	0.262	0.197	0.097	0.261	0.137	0.117	0.098
PDO (FewShot)	0.147	0.120	0.111	0.189	0.237	0.171	0.273	0.195	0.103	0.243	0.193	0.173	0.193
PDO (FewShotU)	0.161	0.129	0.124	0.188	0.233	0.170	0.250	0.190	0.108	0.280	0.128	0.103	0.095
PDO (FewShotUP)	0.144	0.169	0.112	0.192	0.228	0.150	0.250	0.197	0.095	0.217	0.131	0.107	0.117
Oracle	0.149	0.125	0.111	0.200	0.256	0.150	0.255	0.193	0.099	0.262	0.162	0.178	0.195
OPT-6.7B PMI													
Random	0.187	0.166	0.157	0.212	0.274	0.209	0.406	0.282	0.156	0.262	0.263	0.148	0.188
LocalE	0.194	0.171	0.166	0.189	0.220	0.190	0.299	0.262	0.172	0.299	0.123	0.093	0.107
GlobalE	0.205	0.177	0.192	0.195	0.205	0.188	0.313	0.271	0.175	0.305	0.134	0.101	0.111
PDO (FewShot)	0.186	0.168	0.157	0.185	0.243	0.204	0.321	0.278	0.130	0.289	0.232	0.143	0.220
PDO (FewShotU)	0.196	0.170	0.166	0.186	0.219	0.190	0.305	0.260	0.172	0.299	0.129	0.094	0.108
PDO (FewShotUP)	0.191	0.125	0.155	0.183	0.239	0.187	0.305	0.265	0.128	0.304	0.124	0.093	0.113
Oracle	0.190	0.167	0.160	0.186	0.237	0.190	0.313	0.273	0.132	0.282	0.126	0.094	0.125
GPT-J-6B Direct													
Random	0.221	0.154	0.144	0.183	0.246	0.142	0.230	0.212	0.109	0.177	0.204	0.252	0.291
LocalE	0.243	0.166	0.162	0.183	0.221	0.148	0.203	0.196	0.118	0.157	0.124	0.142	0.209
GlobalE	0.247	0.167	0.148	0.178	0.224	0.142	0.198	0.200	0.102	0.144	0.124	0.171	0.254
PDO (FewShot)	0.243	0.149	0.155	0.162	0.231	0.145	0.231	0.208	0.104	0.166	0.183	0.144	0.225
PDO (FewShotU)	0.242	0.166	0.160	0.184	0.221	0.148	0.205	0.192	0.117	0.154	0.120	0.145	0.202
PDO (FewShotUP)	0.252	0.199	0.153	0.180	0.225	0.136	0.205	0.203	0.101	0.175	0.127	0.142	0.212
Oracle	0.254	0.167	0.151	0.186	0.228	0.137	0.195	0.204	0.101	0.189	0.147	0.180	0.260
GPT-J-6B PMI													
Random	0.247	0.176	0.182	0.172	0.218	0.256	0.252	0.317	0.126	0.176	0.227	0.088	0.139
LocalE	0.265	0.189	0.193	0.179	0.202	0.218	0.206	0.290	0.140	0.154	0.101	0.066	0.097
GlobalE	0.275	0.199	0.187	0.175	0.213	0.234	0.201	0.292	0.145	0.149	0.095	0.075	0.083
PDO (FewShot)	0.255	0.177	0.182	0.177	0.202	0.229	0.229	0.298	0.115	0.179	0.213	0.085	0.126
PDO (FewShotU)	0.263	0.190	0.190	0.180	0.202	0.213	0.210	0.288	0.140	0.154	0.100	0.071	0.088
PDO (FewShotUP)	0.263	0.168	0.187	0.190	0.203	0.200	0.210	0.293	0.109	0.164	0.106	0.066	0.113
Oracle	0.261	0.196	0.186	0.195	0.214	0.200	0.214	0.303	0.114	0.175	0.107	0.084	0.114

Table 11: Complete confidence calibration results (measured by ECE), Part 3.

	SST2	CR	MR	SUBJ	SST5	AGNews	TREC	Yahoo Topics	Dbpedia	FPB	Tweet Offensive	Tweet Irony	Tweet Hate
LLaMA-7B Direct													
Random	0.122	0.094	0.091	0.269	0.263	0.122	0.205	0.222	0.140	0.224	0.130	0.217	0.175
LocalE	0.142	0.102	0.100	0.218	0.253	0.127	0.192	0.209	0.147	0.206	0.107	0.192	0.153
GlobalE	0.118	0.088	0.085	0.155	0.245	0.119	0.187	0.197	0.130	0.202	0.138	0.177	0.147
PDO (FewShot)	0.125	0.09	0.094	0.164	0.261	0.119	0.216	0.204	0.139	0.199	0.108	0.158	0.130
PDO (FewShotU)	0.118	0.09	0.087	0.157	0.249	0.118	0.185	0.199	0.128	0.197	0.109	0.141	0.117
PDO (FewShotUP)	0.121	0.09	0.088	0.157	0.239	0.117	0.186	0.203	0.127	0.189	0.099	0.140	0.126
Oracle	0.125	0.085	0.087	0.155	0.255	0.117	0.181	0.209	0.131	0.201	0.134	0.200	0.172
LLaMA-7B PMI													
Random	0.160	0.151	0.141	0.170	0.243	0.161	0.318	0.241	0.135	0.277	0.151	0.123	0.134
LocalE	0.171	0.152	0.145	0.190	0.209	0.161	0.229	0.234	0.160	0.292	0.107	0.096	0.111
GlobalE	0.165	0.149	0.133	0.170	0.245	0.170	0.250	0.236	0.105	0.265	0.082	0.084	0.084
PDO (FewShot)	0.164	0.149	0.137	0.157	0.228	0.142	0.267	0.239	0.198	0.276	0.105	0.115	0.121
PDO (FewShotU)	0.164	0.149	0.140	0.171	0.232	0.154	0.236	0.233	0.109	0.289	0.092	0.080	0.082
PDO (FewShotUP)	0.164	0.149	0.140	0.171	0.220	0.142	0.220	0.236	0.102	0.283	0.123	0.084	0.094
Oracle	0.167	0.151	0.138	0.173	0.223	0.167	0.224	0.244	0.105	0.307	0.130	0.087	0.098
OPT-13B Direct													
Random	0.132	0.100	0.110	0.196	0.248	0.130	0.240	0.211	0.115	0.224	0.282	0.379	0.280
LocalE	0.140	0.105	0.116	0.155	0.237	0.129	0.215	0.199	0.112	0.237	0.290	0.401	0.248
GlobalE	0.137	0.097	0.101	0.201	0.238	0.128	0.218	0.189	0.105	0.217	0.173	0.235	0.204
PDO (FewShot)	0.134	0.100	0.109	0.177	0.243	0.128	0.234	0.197	0.111	0.220	0.215	0.253	0.212
PDO (FewShotU)	0.141	0.106	0.116	0.159	0.236	0.131	0.207	0.198	0.121	0.217	0.191	0.219	0.184
PDO (FewShotUP)	0.135	0.094	0.104	0.198	0.227	0.127	0.214	0.187	0.104	0.200	0.170	0.219	0.180
Oracle	0.138	0.098	0.101	0.211	0.242	0.125	0.200	0.197	0.102	0.219	0.186	0.248	0.201
OPT-13B PMI													
Random	0.189	0.154	0.163	0.186	0.258	0.153	0.361	0.304	0.130	0.208	0.159	0.180	0.169
LocalE	0.202	0.156	0.166	0.133	0.223	0.145	0.254	0.281	0.140	0.225	0.133	0.149	0.134
GlobalE	0.185	0.155	0.163	0.190	0.235	0.157	0.276	0.290	0.121	0.201	0.114	0.114	0.106
PDO (FewShot)	0.190	0.152	0.160	0.179	0.249	0.141	0.299	0.288	0.111	0.216	0.156	0.180	0.167
PDO (FewShotU)	0.203	0.156	0.163	0.146	0.209	0.147	0.243	0.279	0.139	0.223	0.115	0.110	0.105
PDO (FewShotUP)	0.196	0.155	0.161	0.194	0.221	0.140	0.252	0.282	0.112	0.216	0.130	0.110	0.117
Oracle	0.192	0.154	0.163	0.206	0.24	0.156	0.295	0.282	0.118	0.229	0.134	0.117	0.138
LLaMA-13B Direct													
Random	0.101	0.115	0.093	0.205	0.251	0.121	0.184	0.213	0.112	0.220	0.137	0.153	0.186
LocalE	0.113	0.128	0.104	0.185	0.247	0.119	0.184	0.207	0.121	0.247	0.106	0.110	0.137
GlobalE	0.098	0.121	0.088	0.192	0.239	0.116	0.175	0.194	0.115	0.233	0.124	0.093	0.089
PDO (FewShot)	0.100	0.124	0.091	0.184	0.242	0.123	0.191	0.216	0.107	0.221	0.117	0.099	0.104
PDO (FewShotU)	0.101	0.121	0.087	0.191	0.235	0.115	0.171	0.198	0.101	0.241	0.110	0.081	0.093
PDO (FewShotUP)	0.093	0.122	0.092	0.181	0.236	0.116	0.176	0.197	0.101	0.213	0.095	0.085	0.096
Oracle	0.096	0.120	0.089	0.203	0.244	0.115	0.179	0.197	0.104	0.255	0.137	0.139	0.127
LLaMA-13B PMI													
Random	0.154	0.170	0.152	0.187	0.283	0.178	0.277	0.262	0.096	0.239	0.129	0.087	0.106
LocalE	0.161	0.178	0.161	0.197	0.235	0.156	0.256	0.238	0.098	0.253	0.137	0.075	0.106
GlobalE	0.158	0.175	0.156	0.204	0.258	0.174	0.239	0.242	0.103	0.230	0.105	0.069	0.082
PDO (FewShot)	0.152	0.172	0.151	0.191	0.264	0.151	0.261	0.257	0.079	0.257	0.131	0.081	0.109
PDO (FewShotU)	0.159	0.176	0.150	0.206	0.247	0.153	0.241	0.234	0.081	0.276	0.103	0.074	0.078
PDO (FewShotUP)	0.159	0.176	0.157	0.200	0.246	0.146	0.220	0.239	0.079	0.253	0.115	0.079	0.088
Oracle	0.156	0.168	0.154	0.211	0.256	0.176	0.223	0.250	0.091	0.292	0.144	0.074	0.098