

In-context Example Ordering Guided by Label Distributions

Zhichao Xu, Daniel Cohen, Bei Wang and Vivek Srikumar



Abstract

In-context learning (ICL) allows the model to make predictions without task-specific training. Prior works show that in-context learning 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 between near random to near state-of-the-art.

In this work, we formulate in-context example ordering as an optimization problem, and propose two principles for in-context learning example ordering, guided by the label distributions. Experiments on 13 text classification datasets with 9 different autoregressive large language models (LLMs) show that our method can improve classification accuracy, reducing model miscalibration and can help selecting better in-context examples.

The Curious Case of In-context Learning

The three settings we explore for in-context learning

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

1 Translate English to French: ← task description
2 cheese => ← prompt

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

1 Translate English to French: ← task description
2 sea otter => loutre de mer ← example
3 cheese => ← prompt

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

1 Translate English to French: ← task description
2 sea otter => loutre de mer ← examples
3 peppermint => menthe poivrée
4 plush giraffe => girafe peluche
5 cheese => ← prompt

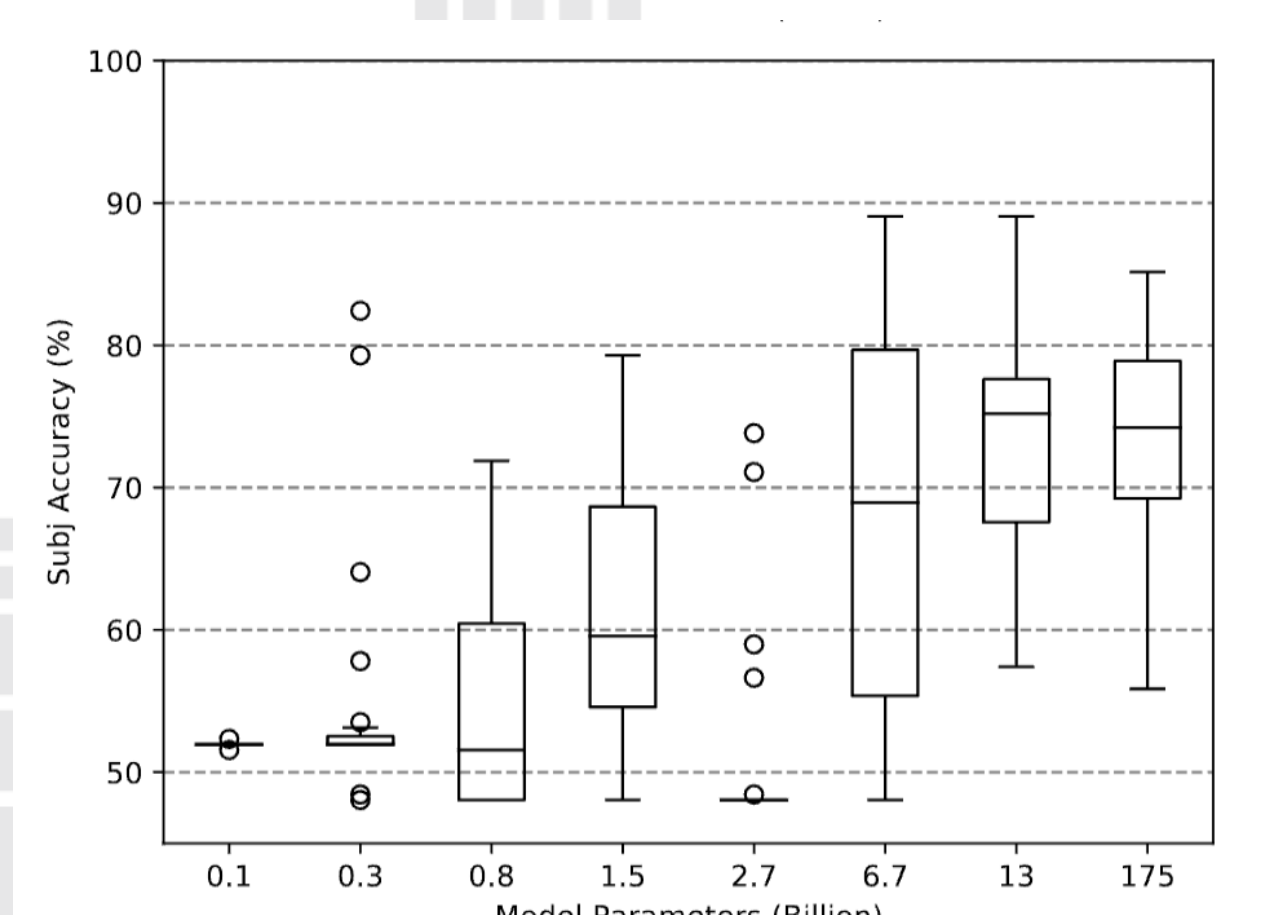
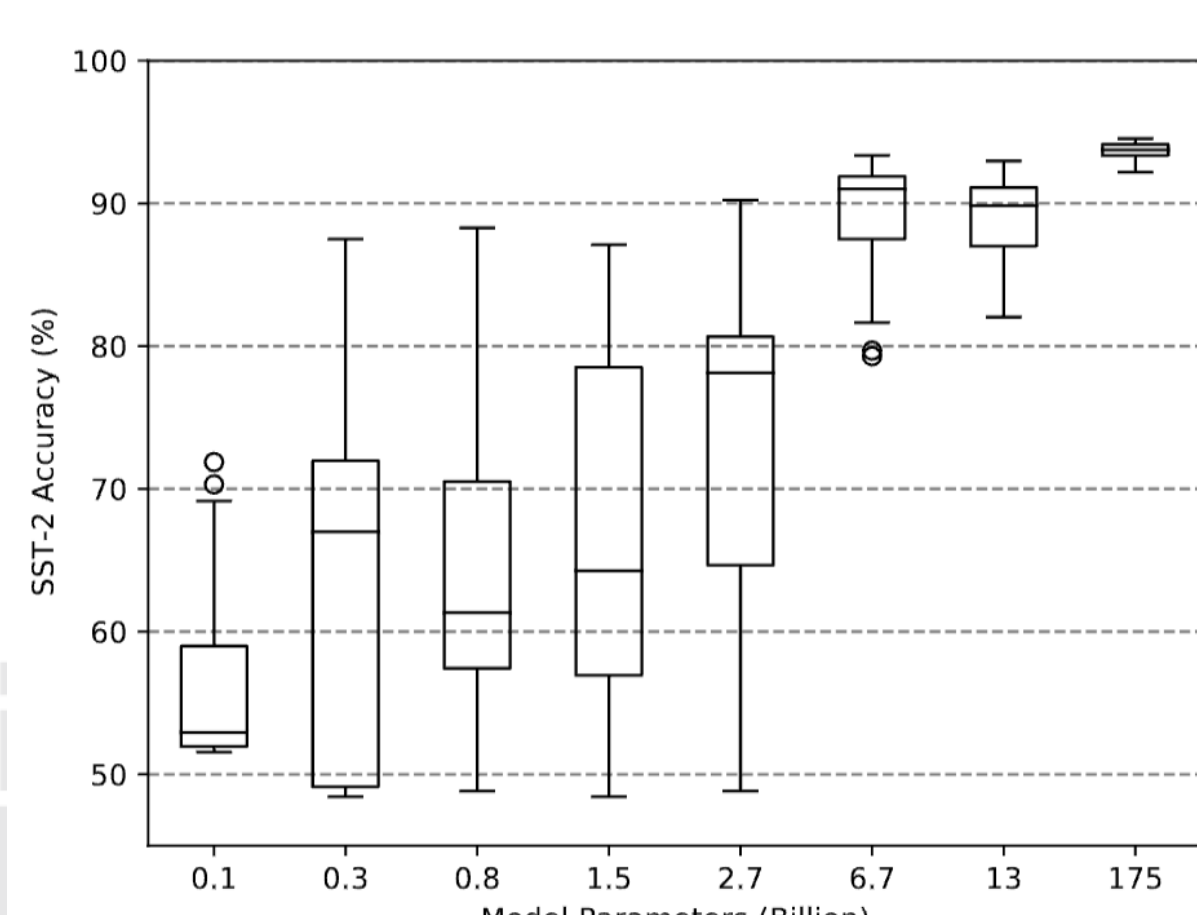
Traditional fine-tuning (not used for GPT-3)

Fine-tuning

The model is trained via repeated gradient updates using a large corpus of example tasks.

1 sea otter => loutre de mer ← example #1
↓
gradient update
↓
1 peppermint => menthe poivrée ← example #2
↓
gradient update
↓
...
↓
1 plush giraffe => girafe peluche ← example #N
↓
gradient update
↓
1 cheese => ← prompt

Brown et al. [1] first demonstrate that LLMs can perform in-context learning with relatively good accuracy. Lu et al. [2] show that ICL performance of smaller LMs are sensitive to example orderings.



Probability Distribution Ordering (PDO)

We consider two problem settings—FewShot with only in-context examples, and FewShot with unlabeled examples. Denote input $x \in \mathcal{X}$, label $y \in \mathcal{Y}$, a small set of k training examples $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \dots, (x_k, y_k)\}$, and the ordering function $\pi := \pi(\mathcal{D})$.

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.

$$\mathcal{L}(\pi) = D_{KL}(P(\mathcal{Y} | \pi, null) || Unif.(\mathcal{Y})) \quad (1)$$

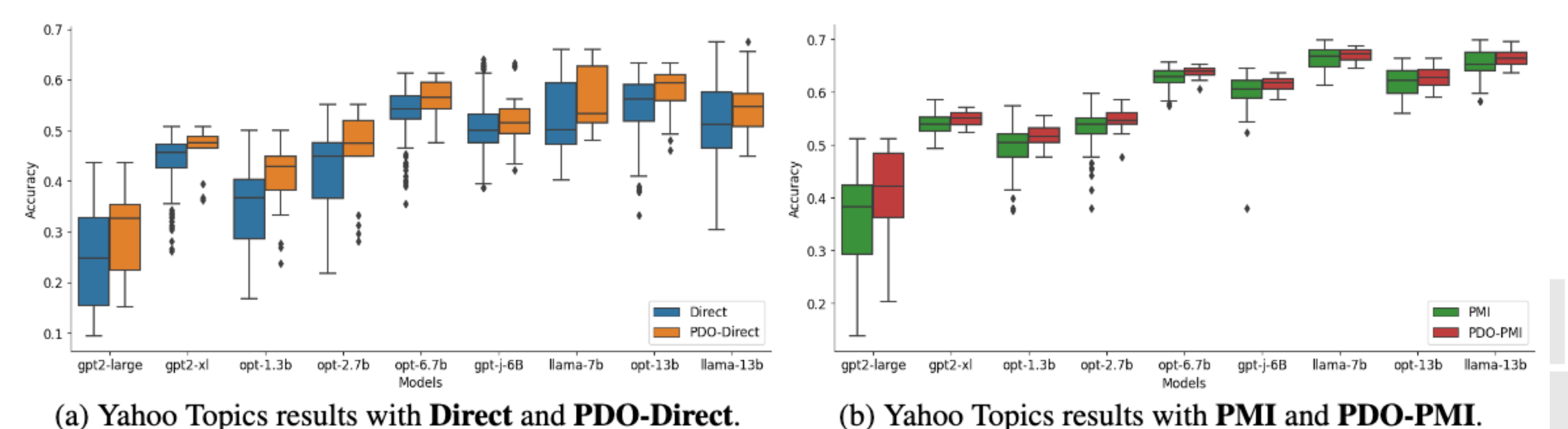
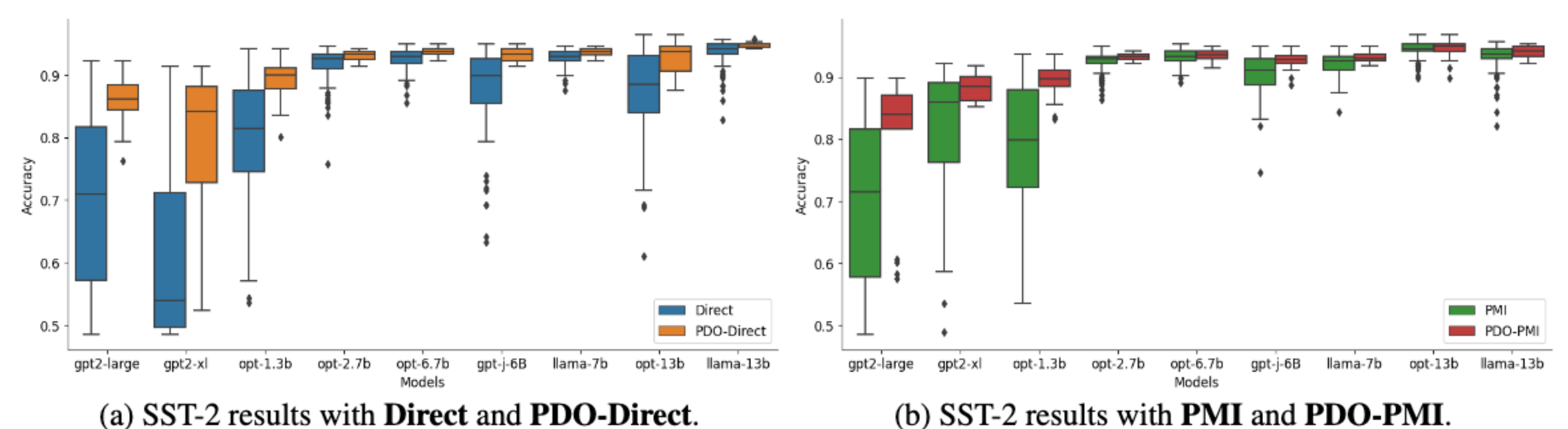
Consider we have unlabeled examples X , and the prior probability distribution Q over label space \mathcal{Y} . We can define the observed label distribution \hat{P} as:

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

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 probability distribution

$$\mathcal{L}(\pi) = D_{KL}(\hat{P}(\mathcal{Y} | \pi) | Q(\mathcal{Y})) \quad (2)$$

$$\pi^* = \arg \min \mathcal{L}(\pi) \quad (3)$$



References

- [1] Brown, Tom, et al. "Language models are few-shot learners." Advances in neural information processing systems 33 (2020): 1877-1901.
- [2] Lu, Yao, et al. "Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity." The 60th Annual Meeting of the Association for Computational Linguistics (2022)