In-context Example Ordering Guided by Label Distributions

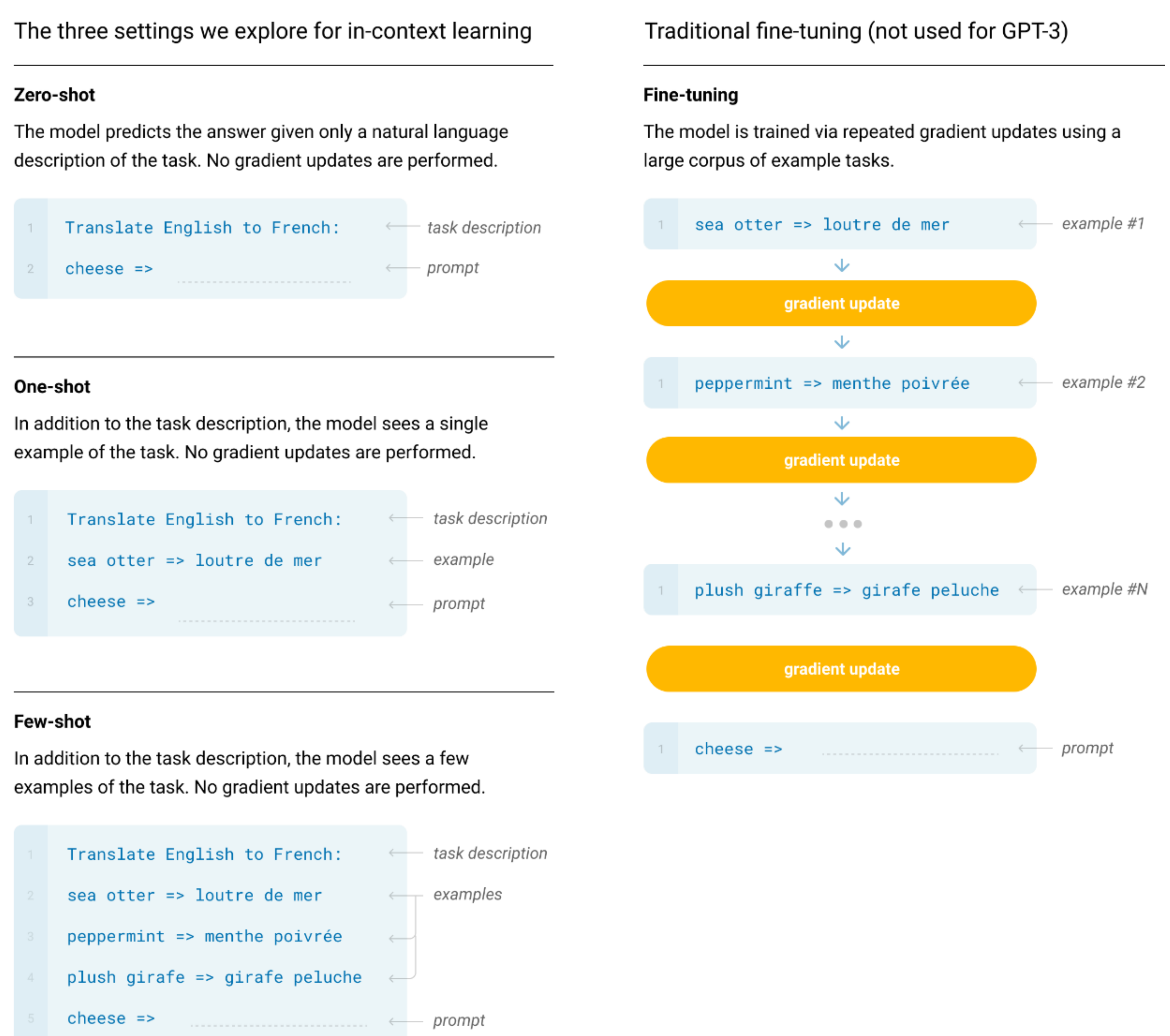
Zhichao Xu, Daniel Cohen, Bei Wang and Vivek Srikumar

Abstract

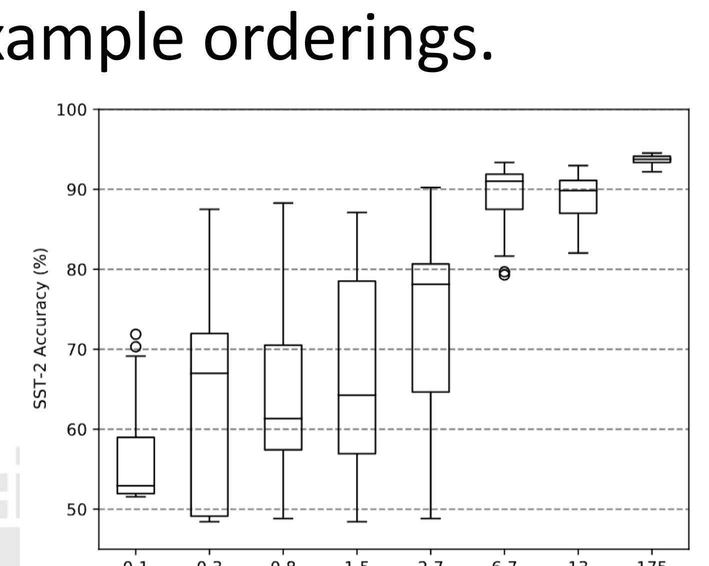
In-context learning (ICL) allows the model to make predictions without task-specific training. Prior works show that in-context learning performance is sensitive to the choice and order of in-context examples. Given the same set of in-context examples with different orderings, model performance may vary between near random to near state-of-the-art.

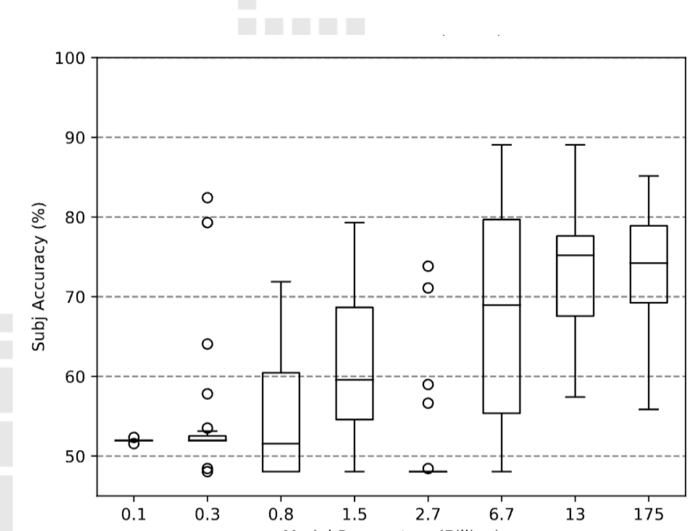
In this work, we formulate in-context example ordering as an optimization problem, and propose two principles for in-context learning example ordering, guided by the label distributions. Experiments on 13 text classification datasets with 9 different autoregressive large language models (LLMs) show that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

The Curious Case of In-context Learning



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.

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

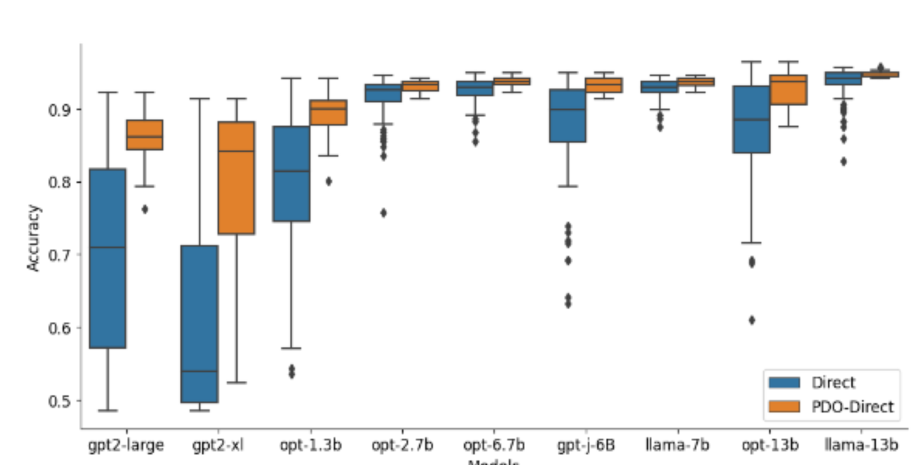
Consider we have unlabeled examples X, and the prior probability distribution Q over label space $\mathcal{Y}$. We can define the observed label distribution $\hat{P}$ as:

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

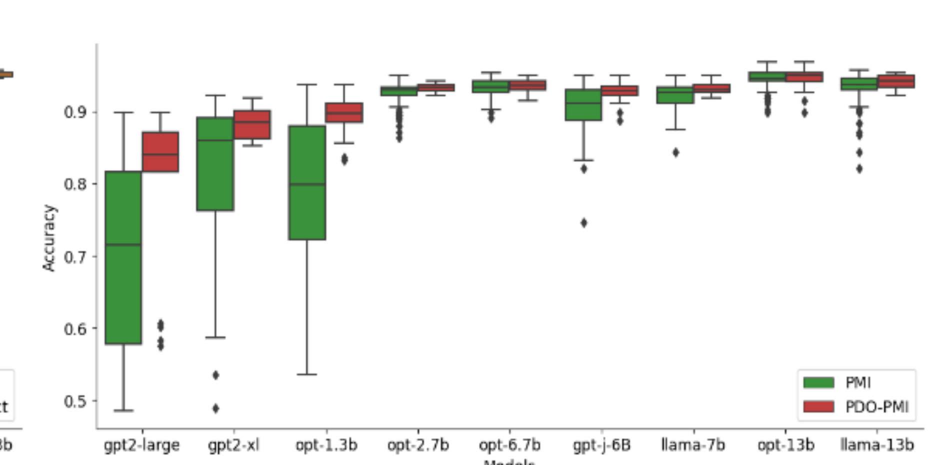
Principle II: Given an unlabeled set of examples and the prior label distribution, well-ordered in-context examples should produce an observed label distribution that matches the prior probability distribution

$$\mathcal{L}(\pi) = D_{KL}\big(\hat{P}(\mathcal{Y} \mid \pi\ ) \mid \mathcal{Q}(\mathcal{Y})\big) \qquad (2)$$

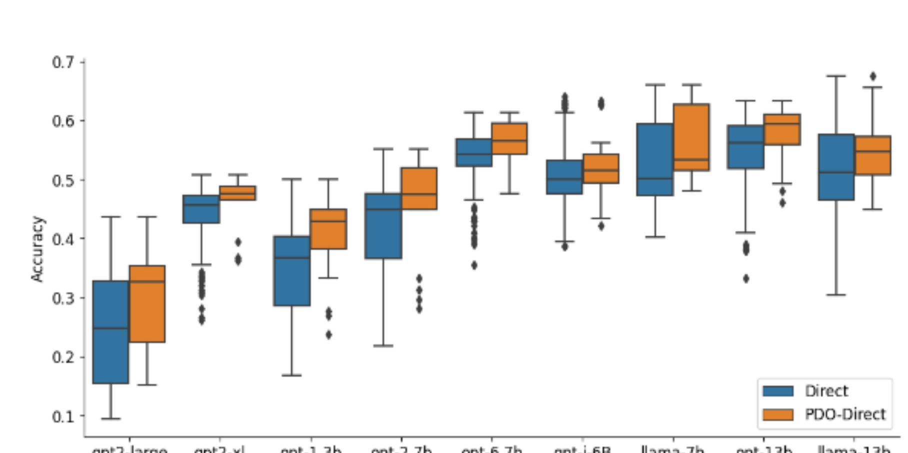
$$\pi^* = \arg\min \mathcal{L}\,(\pi) \qquad (3)$$



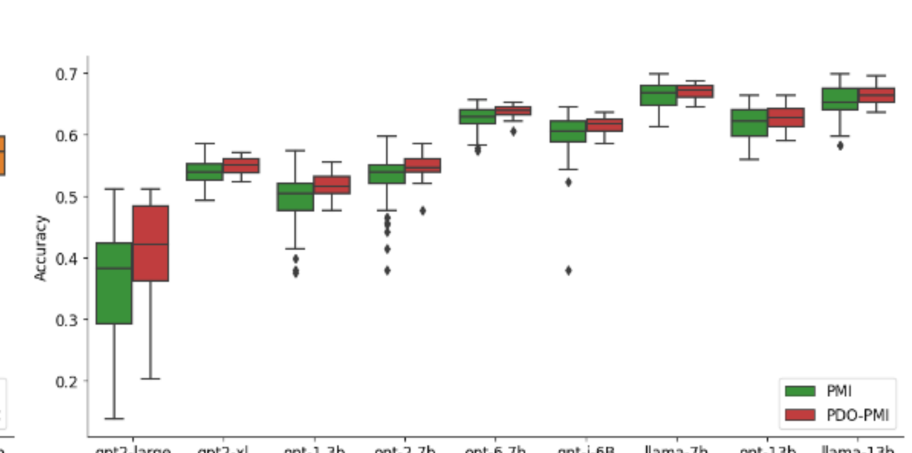
(a) SST-2 results with **Direct** and **PDO-Direct**.



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



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



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

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).

THE UNIVERSITY OF UTAH

Dataminr

www.sci.utah.edu