On the Relation between Sensitivity and Accuracy in In-Context Learning

Anonymous ACL submission

Abstract

In-context learning (ICL) suffers from oversensitivity to the prompt, making it unreliable in real-world scenarios. We study the sensitivity of ICL with respect to multiple perturbation types. First, we find that label bias obscures the true sensitivity, and therefore prior work may have significantly underestimated ICL sensitivity. Second, we observe a strong negative correlation between ICL sensitivity and accuracy: predictions sensitive to perturbations are less likely to be correct. Motivated by these findings, we propose SENSEL, a few-shot selective prediction method that abstains from sensitive predictions. Experiments on ten classification datasets show that SENSEL consistently outperforms a commonly used confidence-based baseline on abstention decisions.

1 Introduction

Few-shot learning (FSL) refers to a system's ability to quickly learn a new task based on a few labeled examples. Recently, in-context learning (ICL) has made significant progress in FSL, where a language model (LM) is prompted with a few demonstrated examples that enable it to make predictions for new examples without any gradient update. However, a known issue of ICL is that it is oversensitive to the prompt (Zhao et al., 2021; Perez et al., 2021), making it less reliable in practice. Despite a near-universal acknowledgment of this issue, when and how the prediction is sensitive remains unclear (Min et al., 2022b; Kim et al., 2022). This paper fills these gaps.

We conduct a systematic study of the ICL sensitivity to prompt perturbations. Specifically, we perturb the task instruction (by paraphrasing and noise injection) and the in-context example orders. We then measure the prediction sensitivity by the magnitude of model output changes due to the prompt perturbation.

Our first observation is that the extent of sensitivity is significantly underestimated due to *label bias* in ICL: LMs tend to assign a higher probability to a specific label regardless of the prompt (Zhao et al., 2021), thus appearing to make stable predictions. Our study shows that the *adjusted sensitivity* after mitigating label bias is up to $\mathbf{2.8x}$ of the *raw sensitivity*.

After mitigating label bias, we observe a negative correlation between the adjusted sensitivity and the accuracy of ICL: if a prediction is sensitive to prompt perturbations, then it is likely to be incorrect (Figure 1 left). This finding aligns with our intuition that if the prediction is sensitive to the prompt that elicits the LM concept (e.g., sentiment) (Xie et al., 2022), then the example is likely not typical for that concept, and is thus more challenging. Our experiments show a significant negative correlation of up to $-\mathbf{0.40}$ (Pearson) between ICL sensitivity and accuracy.

Given the above findings, a natural idea is to use sensitivity as a signal to abstain from making predictions on error-prone examples—an important mechanism to increase user trust when deploying ICL models to high-stakes domains such as healthcare (Korngiebel and Mooney, 2021; Sezgin et al., 2022) and legal systems (Eliot and Lance, 2021). Our proposed method, Sensitivity-based Selective prediction (SENSEL), uses sensitivity to make abstention decisions: the LM abstains on examples where its prediction is sensitive to prompt perturbations (Figure 1 right). Compared to the common approach of training a separate model to make abstention decisions (Platt et al., 1999; Geifman and El-Yaniv, 2019; Kamath et al., 2020), our approach does not require large amounts of labeled data and thus is more suitable for the few-shot setting.

Our experiments show that sensitivity is a stronger signal than output probabilities for abstention. SENSEL consistently outperforms a baseline based on model probabilities (MAXPROB) by up to $+\mathbf{4.1}$ AUC points. Further analysis shows that SENSEL and MAXPROB are *complementary*—

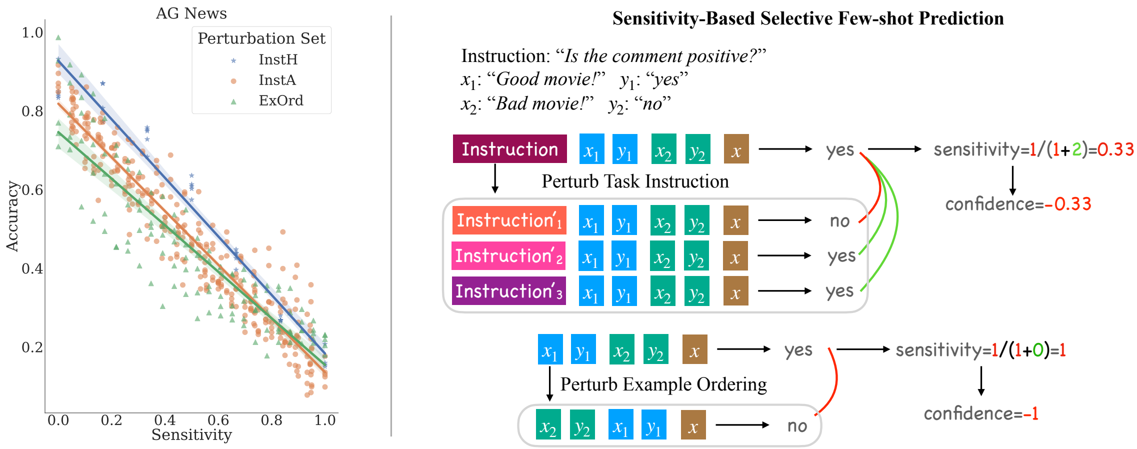

Figure 1: **ICL sensitivity-accuracy correlation** (left): We plot the prediction sensitivity against the prediction accuracy averaged over examples with that sensitivity. Different colors represent different perturbation sets (Section 2.1), and color bands represent 95% confidence intervals. We observe a significant negative correlation between the prediction sensitivity and accuracy of ICL. **SENSEL** (right): SENSEL measures the sensitivity of model predictions to prompt perturbations, and abstains from making predictions on examples with high sensitivity.

MAXPROB falters on high-sensitivity tasks because it relies on oversensitive model probabilities for abstention, while SENSEL capitalizes ICL sensitivity for abstention and hence works better on high-sensitivity tasks.[1]

2 ICL Sensitivity Study

In this section, we study the interplay between label bias and prediction sensitivity in ICL, as well as the relation between sensitivity and accuracy.

2.1 ICL Sensitivity

Background In-context learning is a FSL method using LMs. Given a test example x, we concatenate the task instruction I, a few (K) labeled examples $S = [(x_{\sigma(i)}, y_{\sigma(i)})]_{i=1}^{K}$ in σ order, and the test input x. The probability of each label is then assigned by the next-word probabilities from the LM. We use $p_{\mathrm{LM}}(y \mid x, I, S, \sigma)$ to denote the prediction probabilities, and $f(x, I, S, \sigma) = \arg\max_y p_{\mathrm{LM}}(y \mid x, I, S, \sigma)$ to denote the predicted (most likely) label.

Despite its success, ICL is known to be highly sensitive. Several methods have been proposed to address this issue. Zhou et al. (2022) fine-tune LM to produce consistent predictions on various prompts, while Chen et al. (2022) and Min et al. (2022a) meta-train models to perform ICL to reduce sensitivity. Lu et al. (2022) search for high-performance prompts that lead to less sensitive predictions. Parallel to these works, we connect ICL sensitivity to label bias and prediction accuracy, and propose a new few-shot selective prediction approach based on sensitivity.

Measuring Sensitivity We measure prediction sensitivity by the magnitude of the changes in the predicted label when the prompt is perturbed. We perturb the task instruction and the order of the in-context examples respectively. Formally, we measure the sensitivity of a prediction $f(x, I, S, \sigma)$ with respect to perturbation set P as

$$\frac{1}{|P|} \sum_{(I', S', \sigma') \in P} \mathbf{1}[f(x, I, S, \sigma) \neq f(x, I', S', \sigma')].$$

We use three perturbation sets. *Human Instruction Perturbation* (INSTH) replaces the instruction with other human-written instructions for the same task; *Automatic Instruction Perturbation* (INSTA) perturbs the task instruction by dropping out words and paraphrasing (details in Appendix B); *Example Ordering Perturbation* (EXORD) permutes the ordering of the in-context examples.

Confounding with Label Bias One known issue of ICL is label bias, where LMs assign a higher probability to a specific label regardless of the prompt, and hence appearing to make stable predictions when the prompt is perturbed. Prior work mitigates label bias by adjusting the decision boundary. For example, contextual calibration (CC) renormalizes the predicted label distribution such that it is uniform given null examples (Zhao et al., 2021). Prototypical calibration (PC) clusters the

[1]We will release our code after peer-review.

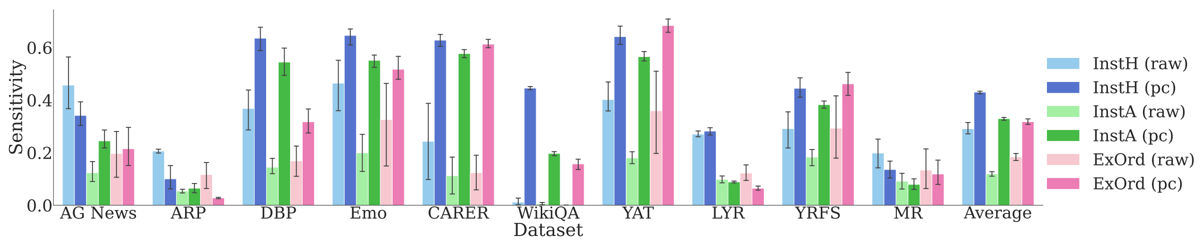

Figure 2: We compare the raw sensitivity with the adjusted sensitivity (label bias mitigated with PC). We observe that the adjusted sensitivity is consistently higher than the raw sensitivity for all three perturbation sets. Error bars represent 95% confidence intervals.

Perturb Set	AG News	ARP	DBP	Emo	CARER	WikiQA	YAT	LYR	YRFS	MR	Avg
InstH	-0.49 (0.04)	-0.55 (0.02)	-0.55 (0.10)	-0.28 (0.11)	-0.31 (0.01)	0.04 (0.10)	-0.35 (0.02)	-0.61 (0.09)	-0.27 (0.04)	-0.49 (0.02)	-0.39 (0.02)
InstA	-0.40 (0.02)	-0.39 (0.03)	-0.65 (0.08)	-0.27 (0.12)	-0.31 (0.04)	-0.18 (0.01)	-0.55 (0.01)	-0.41 (0.05)	-0.25 (0.03)	-0.39 (0.03)	-0.38 (0.01)
ExOrd	-0.38 (0.08)	-0.46 (0.02)	-0.82 (0.02)	-0.17 (0.06)	-0.32 (0.06)	-0.09 (0.05)	-0.51 (0.07)	-0.52 (0.03)	-0.26 (0.04)	-0.47 (0.07)	-0.40 (0.03)

Table 1: We report the Pearson correlation coefficient (and its standard deviation in parenthesis) between ICL sensitivity and accuracy across five randomly sampled sets of few-shot examples (label bias mitigated with PC). We observe a strong negative correlation between the ICL sensitivity and accuracy for all three perturbation sets.

LM's predictions, maps each cluster to a label, and make predictions for new examples by their most likely cluster assignments (Han et al., 2022).

2.2 Experimental Setup

We first compare the raw sensitivity with the adjusted sensitivity. We then compute the Pearson correlation coefficient (Freedman et al., 2007) between the adjusted sensitivity and accuracy.

We run experiments on ten classification datasets covering sentiment classification, emotion classification, topic classification, and question-answering. See Appendix A for dataset details. We use GPT-J 6B (Wang and Komatsuzaki, 2021). We describe additional implementation details in Appendix B. For label bias mitigation, because the same observations hold for PC and CC, we report PC results in the main paper and CC results in Appendix C.1.

2.3 Findings

Sensitivity is underestimated due to label bias. We report raw and adjusted sensitivity with respect to each perturbation set in Figure 2. ICL becomes more sensitive when label bias is mitigated. After prototypical calibration, the adjusted sensitivity is on average $\mathbf{99.0}\%$ higher. Therefore, we argue that the true sensitivity may have been significantly underestimated if label bias is not mitigated.

Among the three perturbation sets, ICL is most sensitive to human instruction perturbations: the perturbations cause the predicted label to change 43.0% of the time (after mitigating label bias). This may be caused by the semantic difference in various human instructions for the same task, such as changing "*Is this product review positive?*" to "*Based on this review, would the user recommend this product?*".

Sensitivity is negatively correlated to accuracy. After mitigating label bias, we measure the Pearson correlation coefficient between sensitivity and accuracy (Table 1). We observe a significant negative correlation between sensitivity (with respect to all perturbation sets) and accuracy across datasets. The correlation is strongest for human instruction perturbations ($-\mathbf{0.42}$).

3 Sensitivity-based Selective Few-shot Prediction

Motivated by the correlation between the sensitivity and accuracy of ICL, we propose SenSel—a selective few-shot prediction method based on sensitivity.

Problem Statement The goal of selective prediction is to *abstain* on examples that the model is not confident about, to avoid presenting wrong predictions to users (Chow, 1957; El-Yaniv and Wiener, 2010). Selective prediction methods score model confidence C on each example, and abstain on examples with low prediction confidence ($C < \gamma$),

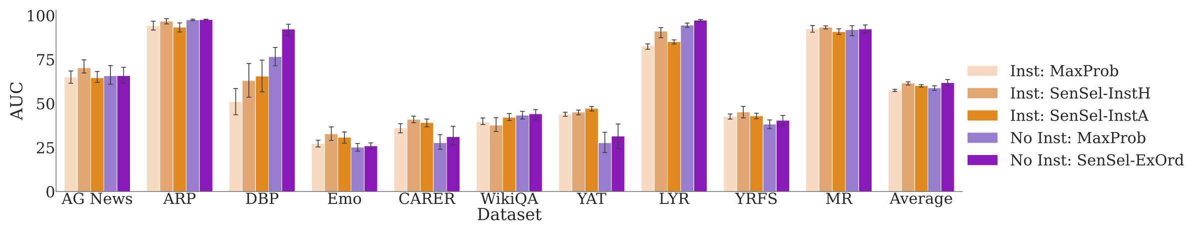

Figure 3: We compare our SENSEL method (label bias mitigated with PC) to the MAXPROB baseline on abstention, measured by AUC score. SENSEL consistently outperforms MAXPROB on both the INST and NO INST setting.

where γ is a threshold to control the trade-off between accuracy and coverage.

Sensitivity-based Selective Prediction SENSEL scores ICL prediction confidence as the negative value of the prediction's sensitivity to prompt perturbations, and then abstains on on examples whose confidence scores (i.e., negative sensitivity scores) are below a certain threshold γ.

Experiment Setup For SENSEL, we always use the adjusted sensitivity computed after mitigating the label bias. As writing good task instructions can be hard (Gao et al., 2021), we experiment with two settings: INST (a task instruction is available), and NO INST (no task instruction is available). We perturb the task instruction in the INST setting (SENSEL-INSTH, SENSEL-INSTA), and perturb the example ordering in the NO INST setting (SENSEL-EXORD). We compare SENSEL to a simple yet strong baseline, MAXPROB, which uses the maximum output probability over the labels as the confidence score (Hendrycks and Gimpel, 2017; Lakshminarayanan et al., 2017) We evaluate the effectiveness of selective prediction methods with the area under the F1-Coverage curve (AUC), which measures the average F1-score at different coverage rates (Kamath et al., 2020). For label bias mitigation, since the same conclusion holds for PC and CC, we report the results for PC in the main paper and the results for CC in Appendix C.2.

Results According to Figure 3, SENSEL consistently outperforms MAXPROB. Among the three perturbation sets, SENSEL with human-written instruction perturbations performs the best (outperforming MAXPROB by an average margin of $+4.1$ AUC points), which is consistent with our sensitivity study that sensitivity to human-written instructions has the strongest correlation with accuracy. Even when instructions are not available, SENSEL-EXORD outperforms MAXPROB consistently.

To understand how well SENSEL and MAXPROB perform on different tasks, we analyze the two methods on tasks with different prediction sensitivity. Specifically, we measure the correlation between task sensitivity and task abstention performance (measured by the AUC of each abstention method minus that of a random abstention baseline). Results show that MAXPROB works better on tasks with low prediction sensitivity (Pearson correlation -0.17), while SENSEL works better on tasks with high prediction sensitivity (correlation $+0.28$) (Figure 2, Figure 3). Hence, SENSEL and MAXPROB are *complementary*: MAXPROB falters on high-sensitivity tasks (e.g., DBP) because it relies on oversensitive model probabilities for abstention, while SENSEL capitalizes ICL sensitivity for abstention and hence works even better on high-sensitivity tasks.

4 Conclusion

While ICL sensitivity is a widely-known issue, its relation to other variables is not studied. This work first conducts a comprehensive study, and finds that ICL sensitivity is negatively correlated with accuracy when label bias is mitigated. Based on this observation, we develop a few-shot selective prediction method that abstains on highly sensitive predictions. Our results show that ICL sensitivity exhibits a useful pattern—it reflects how confidently an LM understands the task.

There are many open questions for future work. First, our study of the sensitivity-accuracy relation is *correlational* but not *causal*. Future work should explore causal experiments to study whether ICL predictions are sensitive because they are uncertain. Second, it remains unclear *why* sensitivity is negatively correlated with accuracy in ICL, which requires a better understanding of the mechanism of ICL. Third, our work mainly focuses on the text *classification* tasks. Future work can further explore other tasks such as text generation and question answering with structured output.

References

Stephen Bach, Victor Sanh, Zheng Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-david, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Fries, Maged Alshaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Dragomir Radev, Mike Tian-jian Jiang, and Alexander Rush. 2022. PromptSource: An integrated development environment and repository for natural language prompts. In *Proceedings of the Association for Computational Linguistics: System Demonstrations*.

Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. 2019. SemEval-2019 task 3: EmoContext contextual emotion detection in text. In *Proceedings of the International Workshop on Semantic Evaluation*.

Yanda Chen, Ruiqi Zhong, Sheng Zha, George Karypis, and He He. 2022. Meta-learning via language model in-context tuning. In *Proceedings of the Association for Computational Linguistics*.

C. K. Chow. 1957. An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers*.

Ran El-Yaniv and Yair Wiener. 2010. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*.

Dr Eliot and B Lance. 2021. Generative pre-trained transformers (gpt-3) pertain to ai in the law. *Available at SSRN 3974887*.

David Freedman, Robert Pisani, and Roger Purves. 2007. *Statistics: Fourth International Student Edition*. International student edition. W.W. Norton & Company.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the Association for Computational Linguistics*.

Yonatan Geifman and Ran El-Yaniv. 2019. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the International Conference on Machine Learning*.

Zhixiong Han, Yaru Hao, Li Dong, and Furu Wei. 2022. Prototypical calibration for few-shot learning of language models. *ArXiv*.

Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*.

Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. In *Proceedings of the Association for Computational Linguistics*.

Junyeob Kim, Hyuhng Joon Kim, Hyunsoo Cho, Hwiyeol Jo, Sang-Woo Lee, Sang-goo Lee, Kang Min Yoo, and Taeuk Kim. 2022. Ground-truth labels matter: A deeper look into input-label demonstrations. *ArXiv*.

Diane M. Korngiebel and Sean D. Mooney. 2021. Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery. *NPJ Digital Medicine*.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, and Christian Bizer. 2014. Dbpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*.

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 Association for Computational Linguistics*.

Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: Understanding rating dimensions with review text. In *Proceedings of the ACM Conference on Recommender Systems*.

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

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? *ArXiv*.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the Association for Computational Linguistics*.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. In *Advances in Neural Information Processing Systems*.

John Platt et al. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of Empirical Methods in Natural Language Processing*.

Emre Sezgin, Joseph Sirrianni, Simon L Linwood, et al. 2022. Operationalizing and Implementing Pretrained, Large Artificial Intelligence Linguistic Models in the US Health Care System: Outlook of Generative Pretrained Transformer 3 (GPT-3) as a Service Model. *JMIR Medical Informatics*.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of Empirical Methods in Natural Language Processing*.

Xiang Zhang and Yann LeCun. 2015. Text understanding from scratch. *ArXiv*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

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

Chunting Zhou, Junxian He, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2022. Prompt consistency for zero-shot task generalization. *ArXiv*.

A Datasets

We study ICL sensitivity and few-shot selective prediction on the following datasets: AG News (Zhang et al., 2015), Amazon Review Polarity (ARP, McAuley and Leskovec (2013)), DBPedia14 (DBP, Lehmann et al. (2014)), Emo2019 (Emo, Chatterjee et al. (2019)), Contextualized Affect Representations for Emotion Recognition (CARER, Saravia et al. (2018)), Wiki Question Answering (WikiQA, (Yang et al., 2015)), Yahoo Answers Topics (YAT, Zhang and LeCun (2015)), Large Yelp Review (LYR, Zhang et al. (2015)), Yelp Reviews Full Star (YRFS, Zhang and LeCun (2015)), and Rotten Tomatoes Movie Review (MR, Pang and Lee (2005)).

B Sensitivity Study Implementation Details

ICL We set the number of shots K to four because the performance flattens out beyond four examples in our setting. All results are averaged over five randomly sampled sets of few-shot examples.

Label Bias To reduce label bias, for CC we follow Zhao et al. (2021) and use the empty string, the "[MASK]" token, and the "N/A" string as the null examples. For PC, similar to Han et al. (2022) we use 1000 unlabeled examples for clustering.

Perturbation Set For human instruction perturbation, we use task instructions from PromptSource (Bach et al., 2022), which provides on average 7 task instructions for each task. For automatic instruction perturbation, we generate 10 perturbed instructions by randomly dropping out 20% of the tokens in the instruction, and another 10 perturbed instructions by using a neural paraphrase model. We use a T5 model fine-tuned on the Google PAWS dataset (Zhang et al., 2019) as the paraphrase model and decode with nucleus sampling of top-$p = 0.9$.

C Additional Results

C.1 ICL Sensitivity Study

Confounding Label Bias We report raw and adjusted sensitivity (label bias mitigated by CC) in Figure 4. Similar to our observations on PC, ICL becomes more sensitive when label bias is mitigated with CC. We also show the sensitivity scores for raw, CC and PC as table in Table 2.

Sensitivity-Accuracy Correlation We report the correlation between prediction sensitivity and accuracy for raw and CC in Table 3. Similar to our observations on PC, there is a significant negative correlation between sensitivity and accuracy across datasets for both raw and CC.

C.2 Sensitivity-Based Selective Few-shot Prediction

Similar to results on PC, all three variants of SENSEL consistently outperform MAXPROB when CC is used to mitigate label bias (Figure 5). Among the three perturbation sets, SENSEL with human-written instruction perturbations performs the best (outperforming MAXPROB by an average margin of $+\mathbf{3.9}$ AUC points). Similar to results on PC, SENSEL-EXORD outperforms MAXPROB consistently even when instructions are not available. We also show the AUC scores as table in Table 4.

We also plot the Coverage-F1 curves, which show coverage rates at different F1 thresholds (Figure 6). The coverage-F1 curves for SENSEL-INSTH and MAXPROB further verify that SENSEL consistently outperforms MAXPROB on different thresholds (Figure 6).

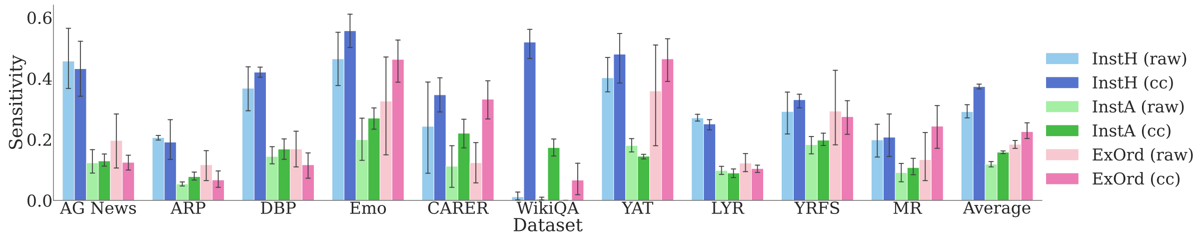

Figure 4: We compare the raw sensitivity with the adjusted sensitivity (label bias mitigated with CC). We observe that the adjusted sensitivity is consistently higher than the raw sensitivity for all three perturbation sets (INSTH: Human Instruction Perturbation, INSTA: Automatic Instruction Perturbation, and EXORD: Example Ordering Perturbation). Error bars represent 95% confidence intervals.

Perturb Set	Sensitivity	AG News	ARP	DBP	Emo	CARER	WikiQA	YAT	LYR	YRFS	MR	Avg
INSTH	Raw	**0.46** (0.12)	**0.21** (0.01)	0.37 (0.09)	0.46 (0.11)	0.24 (0.17)	0.01 (0.02)	0.40 (0.07)	0.27 (0.01)	0.29 (0.08)	0.20 (0.06)	0.29 (0.02)
	PC	0.34 (0.05)	0.10 (0.05)	**0.64** (0.05)	**0.65** (0.04)	**0.63** (0.03)	0.45 (0.01)	**0.64** (0.04)	**0.28** (0.02)	**0.44** (0.04)	0.14 (0.04)	**0.43** (0.01)
	CC	0.43 (0.10)	0.19 (0.08)	0.42 (0.02)	0.56 (0.06)	0.35 (0.07)	**0.52** (0.05)	0.48 (0.09)	0.25 (0.02)	0.33 (0.03)	**0.21** (0.08)	0.37 (0.01)
INSTA	Raw	0.12 (0.04)	0.05 (0.01)	0.14 (0.03)	0.20 (0.08)	0.11 (0.08)	0.01 (0.01)	0.18 (0.02)	**0.10** (0.01)	0.18 (0.04)	0.09 (0.03)	0.12 (0.01)
	PC	**0.24** (0.04)	0.06 (0.02)	**0.54** (0.06)	**0.55** (0.03)	**0.58** (0.02)	**0.20** (0.01)	**0.57** (0.02)	0.09 (0.00)	**0.38** (0.02)	0.08 (0.02)	**0.33** (0.01)
	CC	0.13 (0.02)	**0.08** (0.01)	0.17 (0.04)	0.27 (0.04)	0.22 (0.06)	0.17 (0.03)	0.14 (0.01)	0.09 (0.02)	0.20 (0.02)	**0.11** (0.03)	0.16 (0.01)
EXORD	Raw	0.20 (0.10)	**0.12** (0.06)	0.17 (0.07)	0.33 (0.18)	0.12 (0.08)	0.00 (0.00)	0.36 (0.18)	**0.12** (0.03)	0.29 (0.14)	0.13 (0.09)	0.18 (0.01)
	PC	**0.21** (0.08)	0.03 (0.00)	**0.32** (0.05)	**0.52** (0.05)	**0.61** (0.02)	**0.16** (0.02)	**0.68** (0.03)	0.06 (0.01)	**0.46** (0.05)	0.12 (0.06)	**0.32** (0.01)
	CC	0.12 (0.03)	0.07 (0.03)	0.12 (0.05)	0.46 (0.08)	0.33 (0.07)	0.07 (0.06)	0.46 (0.08)	0.10 (0.01)	0.27 (0.07)	**0.24** (0.08)	0.23 (0.03)

Table 2: We compare the raw sensitivity with the adjusted sensitivity after mitigating label bias. We observe that the adjusted sensitivity is consistently higher than the raw sensitivity for all three perturbation sets (INSTH: Human Instruction Perturbation, INSTA: Automatic Instruction Perturbation, and EXORD: Example Ordering Perturbation). The standard deviation across five randomly sampled sets of few-shot examples is reported in parenthesis.

Sensitivity	Perturb Set	AG News	ARP	DBP	Emo	CARER	WikiQA	YAT	LYR	YRFS	MR	Avg
Raw	INSTH	−0.49 (0.14)	−0.50 (0.10)	−0.11 (0.17)	−0.21 (0.13)	−0.12 (0.12)	−0.09 (0.05)	−0.25 (0.03)	−0.54 (0.04)	−0.24 (0.05)	−0.31 (0.13)	−0.29 (0.04)
	INSTA	−0.24 (0.08)	−0.29 (0.12)	−0.17 (0.16)	−0.09 (0.12)	−0.06 (0.10)	−0.32 (0.23)	−0.19 (0.13)	−0.31 (0.03)	−0.12 (0.07)	−0.18 (0.11)	−0.20 (0.06)
	EXORD	−0.14 (0.12)	−0.36 (0.16)	−0.16 (0.22)	−0.30 (0.19)	−0.08 (0.04)	/ /	−0.13 (0.10)	−0.59 (0.03)	−0.22 (0.07)	−0.32 (0.13)	−0.26 (0.05)
CC	INSTH	−0.50 (0.07)	−0.57 (0.05)	−0.38 (0.09)	−0.06 (0.04)	−0.29 (0.02)	−0.34 (0.09)	−0.35 (0.02)	−0.48 (0.10)	−0.28 (0.02)	−0.48 (0.10)	−0.37 (0.02)
	INSTA	−0.26 (0.05)	−0.24 (0.12)	−0.38 (0.08)	0.00 (0.03)	−0.14 (0.05)	−0.35 (0.03)	−0.28 (0.04)	−0.33 (0.09)	−0.20 (0.03)	−0.38 (0.08)	−0.26 (0.02)
	EXORD	−0.19 (0.11)	−0.47 (0.07)	−0.52 (0.03)	−0.22 (0.08)	−0.30 (0.05)	−0.33 (0.09)	−0.37 (0.06)	−0.58 (0.05)	−0.20 (0.03)	−0.47 (0.05)	−0.37 (0.02)

Table 3: We report the Pearson correlation coefficient (and its standard deviation in parenthesis) between ICL sensitivity and accuracy across five randomly sampled sets of few-shot examples (label bias mitigated with CC). We observe a strong negative correlation between the ICL sensitivity and accuracy for all three perturbation sets.

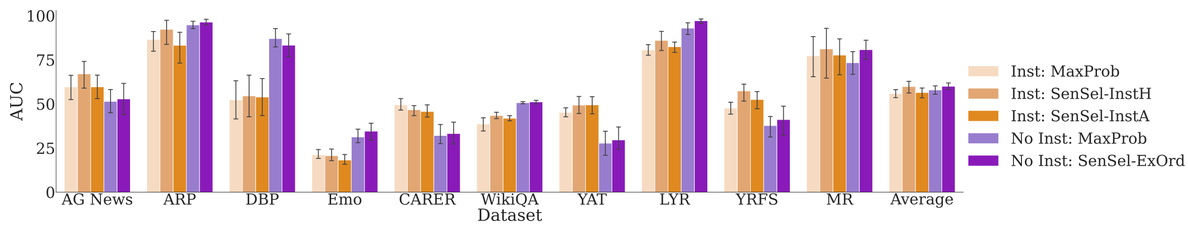

Figure 5: We compare our SENSEL method (confounding label bias mitigated by CC) to the MAXPROB baseline. SENSEL consistently outperforms MAXPROB under both the INST setting and the NO INST setting.

					PC						
Method	AG News	ARP	DBP	Emo	CARER	WikiQA	YAT	LYR	YRFS	MR	Avg
INST: MAXPROB	64.9 (4.0)	94.2 (3.2)	51.0 (8.5)	27.2 (2.2)	36.1 (2.9)	39.7 (2.2)	43.9 (1.3)	82.7 (1.9)	42.6 (1.6)	92.3 (2.2)	57.5 (0.7)
INST: SENSEL-INSTH	**70.2** (4.5)	96.6 (1.8)	63.1 (11.0)	**32.7** (4.3)	**40.9** (2.1)	37.7 (4.8)	45.1 (1.5)	91.0 (3.6)	**45.2** (3.5)	**93.3** (1.1)	61.6 (1.1)
INST: SENSEL-INSTA	64.6 (3.5)	93.3 (3.0)	65.6 (10.5)	30.8 (3.5)	39.2 (2.5)	42.1 (2.4)	**47.2** (1.5)	85.1 (1.3)	42.9 (1.8)	90.8 (1.8)	60.2 (0.8)
NO INST: MAXPROB	65.7 (6.0)	97.5 (0.4)	76.7 (6.1)	25.2 (2.4)	27.7 (4.9)	43.3 (2.5)	27.7 (6.6)	94.5 (1.4)	38.3 (2.9)	91.9 (3.4)	58.8 (1.5)
NO INST: SENSEL-EXORD	65.8 (5.1)	**97.6** (0.2)	**92.2** (3.9)	25.9 (2.1)	31.1 (6.4)	**44.1** (3.2)	31.5 (7.9)	**97.3** (0.5)	40.4 (3.3)	92.3 (2.6)	**61.8** (2.1)
					CC						
Method	AG News	ARP	DBP	Emo	CARER	WikiQA	YAT	LYR	YRFS	MR	Avg
INST: MAXPROB	59.6 (7.7)	86.5 (7.1)	52.3 (12.7)	21.2 (3.1)	**49.5** (3.8)	38.6 (4.2)	45.2 (2.9)	80.6 (3.4)	47.6 (4.0)	77.3 (13.5)	55.8 (2.6)
INST: SENSEL-INSTH	**67.1** (8.6)	92.1 (8.6)	54.4 (14.0)	20.6 (3.7)	46.7 (3.4)	43.3 (2.4)	**49.3** (5.8)	85.8 (6.5)	**57.2** (5.8)	**81.0** (15.9)	59.7 (3.7)
INST: SENSEL-INSTA	59.5 (7.9)	83.5 (9.8)	53.8 (12.7)	18.1 (3.2)	45.4 (4.0)	41.9 (1.9)	**49.3** (5.7)	82.4 (3.4)	52.5 (5.6)	77.4 (12.3)	56.4 (3.2)
NO INST: MAXPROB	51.4 (7.7)	94.7 (2.4)	**87.0** (5.8)	31.3 (4.4)	32.1 (6.4)	50.7 (0.7)	27.7 (7.9)	93.0 (3.9)	37.7 (6.9)	73.3 (7.7)	57.9 (2.8)
NO INST: SENSEL-EXORD	52.9 (11.0)	**96.4** (1.8)	83.2 (7.3)	**34.5** (5.7)	33.1 (7.4)	**51.1** (1.3)	29.5 (7.6)	**97.1** (1.6)	40.9 (9.3)	80.8 (6.2)	**60.0** (2.2)

Table 4: We report the AUC score of our SENSEL method and the MAXPROB baseline. SENSEL consistently outperforms MAXPROB under both the INST setting and the NO INST setting. The standard deviation across five randomly sampled sets of few-shot examples is reported in parenthesis.

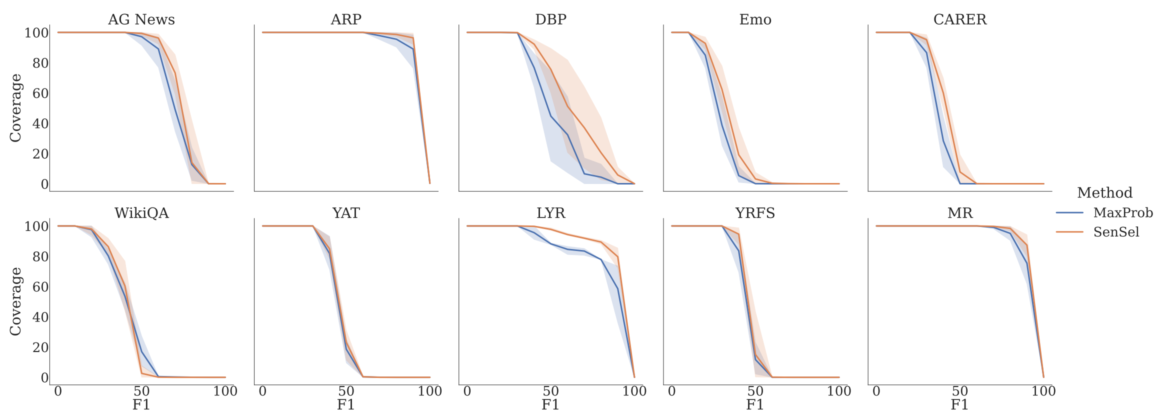

Figure 6: We plot the Coverage-F1 curves of MAXPROB and SENSEL-INSTH (confounding label bias mitigated by PC). SENSEL-INSTH consistently achieves higher coverage rates at different F1 thresholds compared to MAXPROB. Color bands represent 95% confidence intervals.