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

Yanda Chen[1] **Chen Zhao**[2] **Zhou Yu**[1] **Kathleen McKeown**[1] **He He**[2]
[1]Columbia University, [2]New York University

{yanda.chen, kathy}@cs.columbia.edu, cz1285@nyu.edu
zy2461@columbia.edu, hehe@cs.nyu.edu

Abstract

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

1 Introduction

Few-shot learning (FSL) refers to a system's ability to quickly learn a new task based on few labeled examples. Recently, prompting language models (LMs) for FSL achieved remarkable progress via in-context learning (ICL) (Brown et al., 2020), though a known issue is that ICL is oversensitive to the prompt (Schick and Schütze, 2021; Gao et al., 2021), which makes it less reliable and stable in practice. Despite a near-universal acknowledgement of this problem, we do not know its magnitude and effect. This paper fills that gap.

We define prediction sensitivity as how much the model output changes given small perturbations of the input. Here in the context of ICL, we focus on *prompt* perturbations — perturb the task instruction and the order of in-context examples. With thorough study, we first argue that the extent of oversensitivity is significantly underestimated because of *label bias*—LMs systematically assign a higher probability to a preferred label (Zhao et al., 2021). Our experiments on ten classification benchmarks confirm that addressing label bias significantly increases ICL sensitivity to up to **2.5x**.

After removing label bias, we observe a negative correlation between ICL sensitivity and accuracy: if a prediction is sensitive to prompt perturbations, then it is likely to be incorrect. Our intuition is that if the model is sensitive to the elicitation content (the prompt), then the model does not actually learn the task during pre-training, and is thus likely to make mistakes. Our experiments show a strong negative correlation of up to $-\mathbf{0.422}$ between ICL sensitivity and accuracy.

With our findings, a natural idea is to use sensitivity as a signal to abstain from making predictions on examples that are likely to have wrong predictions—an important mechanism to increase user trust when deploying ICL models into real-world, especially in high-stakes domains such as medical (Korngiebel and Mooney, 2021; Sezgin et al., 2022) and legal (Eliot and Lance, 2021). Unlike the fully supervised setting, training an abstention predictor is impossible in the few-shot scenario as only few labeled examples are available. Our proposed method, Sensitivity-based Selective prediction (SENSEL), uses ICL sensitivity to make decisions, where the LM abstains on examples that are most sensitive to prompt perturbations.

Experiments show that sensitivity is a stronger signal than model probabilities on selective prediction, as SENSEL consistently outperforms a baseline based on model probabilities (MAXPROB) by up to **5.1** AUC points. Our further analysis confirms SENSEL works better on tasks with high sensitivity (correlation $+\mathbf{0.285}$), where the margin over MAXPROB is larger.

Our contributions are as follows. *(i)* We find that prior work may have significantly underestimated true ICL sensitivity if label bias is not addressed beforehand. *(ii)* We observe a strong negative correlation between ICL sensitivity and accuracy. *(iii)* We propose a sensitivity-based selective prediction method SENSEL that consistently outperforms a common baseline based on model probabilities.

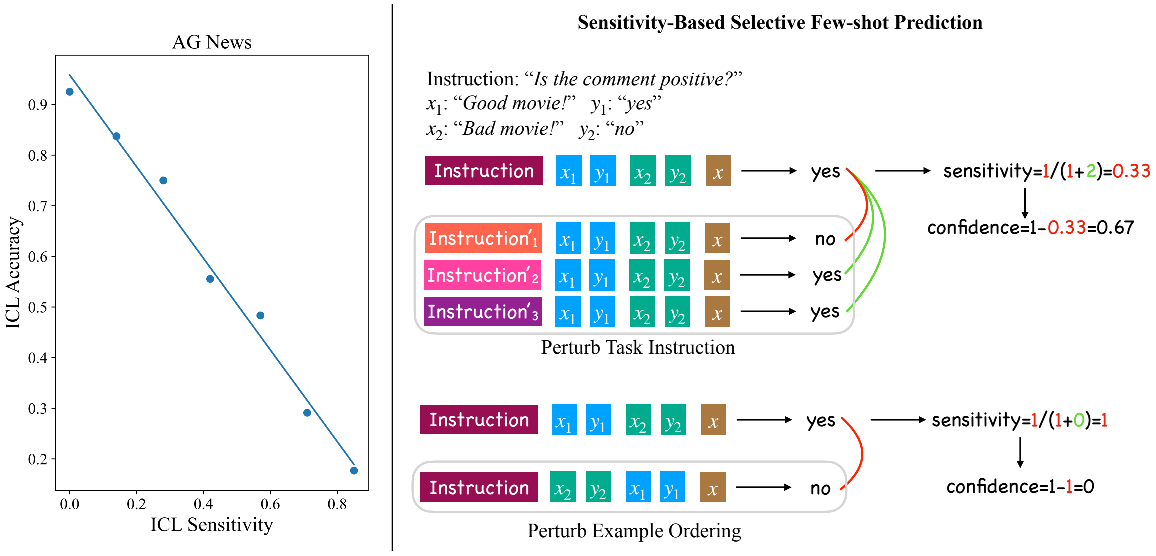

Figure 1: **ICL sensitivity-accuracy correlation** (left): We plot ICL sensitivity (with respect to human-written instruction perturbations) against ICL accuracy across test examples for the AG News benchmark, where we observe a strong negative correlation between ICL sensitivity and ICL accuracy. **SENSEL** (right): Our proposed Sensitivity-based Selective prediction method (SENSEL) measures the sensitivity of model predictions on perturbed prompts, and abstains from making predictions on examples with high sensitivity.

2 Background

We first introduce relevant concepts related to ICL over LMs, and then describe existing approaches for selective prediction and model calibration.

In-context Learning We focus on few-shot setting by prompting language models through in-context learning (ICL), where the LM is given few (K) labeled examples $S = \{(x^i, y^i)\}_{i=1}^K$ to learn a new task. Given a test example x^{target}, we concatenate the task instruction I, few-shot labeled examples S ordered by σ, and the target input to predict x^{target}. We score the probability of each label based on LM's next word predictions. We use $\hat{y}_{x^{\text{target}},I,S,\sigma}$ to denote the predicted label.

Despite its success, ICL is known to be oversensitive, and therefore the output is unreliable in practice. Several methods have been proposed to tackle this issue. Zhou et al. (2022) propose to fine-tune the LM to output consistent predictions across diverse prompts, while Chen et al. (2022) and Min et al. (2022) propose to meta-train ICL to improve consistency. This paper aims at a comprehensive understanding of ICL oversensitivity in terms of magnitudes and effects, and proposes a new few-shot selective prediction approach based on the findings.

Removing Label Bias in ICL One known issue of ICL is label bias, where LM systematically assigns higher probability to a preferred label, and hence assign the same label for most inputs. Prior work proposed to adjust the decision boundary to address label bias, including contextual calibration (Zhao et al., 2021) and prototypical calibration (Han et al., 2022). Contextual calibration (CC) re-normalizes the LM's prediction distribution so that its label prior is shifted towards uniform, and estimates the label prior of an off-the-shelf LM using its predicted distributions on a set of label-neutral examples. Prototypical calibration (PC) re-normalizes the LM's prediction distribution via clustering (Han et al., 2022). PC uses Gaussian mixture distribution to estimate the prototypical clusters for all label categories, and then assigns each cluster to the corresponding label by solving a weighted bipartite matching problem. During inference, the prediction is calibrated by the likelihood of prototypical clusters.

Selective Prediction In selective prediction, the model is allowed to abstain from making predictions on examples that it is not confident about, so as to avoid giving wrong predictions (Chow, 1957; El-Yaniv and Wiener, 2010). The goal of selective prediction approaches is to find the optimal tradeoff between accuracy and coverage. MAXPROB is a simple yet effective method, which directly uses the probability as confidence estimate (Hendrycks and Gimpel, 2017; Lakshminarayanan et al., 2017; Xin et al., 2021). Recent work also proposed to train a model to predict abstention (Platt et al., 1999; Geifman and El-Yaniv, 2017, 2019; Kamath et al., 2020), which requires large amount of labeled data.

Model Calibration A related problem to selective prediction is model calibration, where the goal is to provide accurate prediction probabilities so that it reflects the likelihood of correctness (Nguyen and O'Connor, 2015). As long as the model is well-calibrated, directly using its probability (i.e., MAXPROB) is the optimal selective prediction method. A common approach for model calibration is temperature scaling (Guo et al., 2017), which uses a single scalar parameter (optimized on the dev set) to scale the confidence. While temperature scaling is shown to be highly effective on the widely used calibration metric Expected Calibration Error (Guo et al., 2017; Desai and Durrett, 2020; Jiang et al., 2021), its calibrated probability is not distinguishable for selective prediction (Si et al., 2022).

3 ICL Sensitivity Study

We study why ICL is sensitive on some examples but not others. Motivated by previous work (Hahn et al., 2021) that shows examples where the *ground truth* labels change with small input perturbations (i.e., sensitive) are more *difficult* inherently, we conjecture that if an ICL *prediction* is sensitive to the elicit prompt, then this *prediction* is likely to be *incorrect*. To experiment this assumption, we need to make sure that ICL sensitivity is not obscured by other factors, and label bias is one factor. Therefore, we first study the effect of label bias to ICL sensitivity, and then experiment the correlation between ICL sensitivity and ICL accuracy after removing label bias.

3.1 ICL Sensitivity

We define prediction sensitivity as how much the model output changes given small perturbations of the input. Here we focus on label changes, i.e., a predicted output changes if the predicted label changes. In the context of ICL, we can perturb the input in multiple ways, and here we focus on *prompt* perturbations — perturb the task instruction and perturb the in-context example ordering.

Formally, to measure the sensitivity of an ICL prediction $\hat{y}_{x,I,S,\sigma}$, we first create a perturbation set P consisting of prompts perturbed from (I, S, σ). We measure the sensitivity $s(x, I, S, \sigma)$ as

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

We measure the sensitivity with respect to three perturbation sets. **Human Instruction Perturbation (Inst-H)** that consists of all task instructions written by humans; **Automatic Instruction Perturbation (Inst-A)** that automatically perturbs task instructions in two ways – word dropout and paraphrase; **Example Ordering Perturbation (Ex Ord)** that consists of the $K! - 1$ other permutations of the same K labeled examples.

3.2 Experimental setup

Datasets Our sensitivity study is on ten classification benchmarks covering different task categories, including sentiment classification, emotion classification, topic classification, and question-answer classification: 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)).

Implementation For ICL, we use GPT-J 6B (Wang and Komatsuzaki, 2021). We set the number of shots K to be four because ICL performance flattens out beyond four examples. All results are averaged across five randomly sampled sets of few-shot examples. For contextual calibration, we follow Zhao et al. (2021) to use the empty string, the "[MASK]" token, and the "N/A" string as the label-neutral examples. 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-$p = 0.9$.

Adjustment	Perturb Set	AGNews	ARP	DBP	Emo	CARER	WikiQA	YAT	LYR	YRFS	MR	Avg
None	Inst - H	**0.415**	**0.194**	0.389	0.460	0.405	0.011	0.541	0.203	0.307	0.191	0.312
	Inst - A	0.132	0.050	0.144	0.199	0.239	0.006	0.180	0.071	0.217	0.060	0.130
	Ex. Ord.	0.197	0.117	0.169	0.326	0.124	0.000	0.359	0.122	0.293	0.134	0.184
CC	Inst - H	0.407	0.164	0.435	0.571	0.439	**0.518**	0.517	0.187	0.347	**0.199**	0.378
	Inst - A	0.123	0.070	0.168	0.270	0.303	0.201	0.145	0.058	0.231	0.125	0.170
	Ex. Ord.	0.121	0.034	0.105	0.387	0.291	0.050	0.350	0.084	0.302	0.191	0.191
PC	Inst - H	0.322	0.093	**0.619**	**0.645**	**0.691**	0.445	**0.694**	**0.209**	0.445	0.135	**0.430**
	Inst - A	0.245	0.046	0.539	0.543	0.634	0.198	0.569	0.071	0.396	0.085	0.333
	Ex. Ord.	0.205	0.028	0.318	0.514	0.613	0.156	0.681	0.064	**0.456**	0.119	0.315

Table 1: ICL sensitivity with respect to different perturbation sets under different decision boundary adjustment methods. We observe that ICL is consistently sensitive with respect to all three perturbation sets, especially when decision boundary adjustment methods are applied to reduce label bias.

Adj.	Perturb	AG News	ARP	DBP	Emo	CARER	WikiQA	YAT	LYR	YRFS	MR	Avg
None	Inst - H	-0.438†	-0.440†	-0.203†	-0.228†	-0.156	-0.113	-0.149	-0.424†	-0.183†	-0.328†	-0.266
	Inst - A	-0.272†	-0.327†	-0.167	-0.089	-0.111	-0.302	-0.192	-0.343†	-0.084†	-0.157	-0.204
	Ex Ord	-0.145†	-0.358†	-0.159†	**-0.295**†	-0.080	/	-0.132	-0.590†	-0.215†	-0.322†	-0.255
CC	Inst - H	**-0.501**†	**-0.577**†	-0.571†	-0.097†	**-0.359**†	**-0.515**†	-0.385†	-0.472†	-0.253†	-0.336†	-0.407
	Inst - A	-0.228	-0.269	-0.382†	0.000	-0.136	-0.391†	-0.280†	-0.328†	-0.161†	-0.370†	-0.254
	Ex Ord	-0.191†	-0.484†	-0.516†	-0.282†	-0.240†	-0.310†	-0.301†	-0.567†	-0.194†	-0.419†	-0.350
PC	Inst - H	-0.491†	-0.541†	-0.688†	-0.274†	-0.351†	-0.065†	-0.393†	**-0.624**†	-0.256†	**-0.542**†	**-0.422**
	Inst - A	-0.427†	-0.445†	-0.646†	-0.269†	-0.315†	-0.185†	**-0.540**†	-0.507†	-0.238†	-0.337†	-0.391
	Ex Ord	-0.381†	-0.459†	**-0.822**†	-0.159	-0.333†	-0.089	-0.503†	-0.518†	**-0.258**†	-0.471†	-0.399

Table 2: The correlation between the sensitivity and accuracy of ICL predictions. We observe that sensitivity is negatively correlated with accuracy, with stronger correlation when decision boundary adjustment methods are applied to address label bias. † denotes correlations with p-value < 0.05.

3.3 Findings

Sensitivity is underestimated with label bias. According to Table 1, we first observe that among the three perturbation sets, ICL is most sensitive to human instruction perturbations (in which case the perturbations cause the label predicted by the LM to change in 31.2% of the times). Human instruction perturbations often change the semantics of the original instruction, such as perturbing "*Is this product review positive?*" to "*Based on this review, would the user recommend this product?*".

ICL is more sensitive when decision boundary adjustment methods are applied. After prototypical calibration, the sensitivity is higher by **37.8%** on human instruction perturbations, **156.2%** on automatic instruction perturbations, and **71.2%** on example ordering perturbations. Hence, prior work may have significantly underestimated the true ICL sensitivity if label bias is not addressed beforehand.

Sensitivity is negatively correlated to Accuracy. Table 2 reports the correlation between the sensitivity and accuracy of ICL predictions. Results verify that a sensitive ICL prediction is likely to be incorrect, as sensitivity correlates negatively with accuracy on all tasks and perturbation sets. The correlation is strongest for sensitivity with respect to human instruction perturbations ($-\mathbf{0.422}$ when PC is applied), since human instruction perturbations are the most diverse ones.

4 Sensitivity-based Selective Few-shot Prediction

Motivated by our findings on the correlation between ICL sensitivity and accuracy, we propose a sensitivity-based method (SENSEL) to improve few-shot prediction algorithms.

Adj.	Method	AG News	ARP	DBP	Emo	CARER	WikiQA	YAT	LYR	YRFS	MR	Avg
CC	F1@Cov100	55.5	85.8	47.6	15.5	31.6	34.7	42.5	86.3	50.3	79.8	53.0
	MAXPROB	57.7	87.7	52.3	21.2	40.0	35.4	45.2	95.4	50.5	86.9	57.2
	SENSEL-Inst-H	65.9	93.2	62.2	21.4	**40.8**	41.9	49.5	94.1	**56.4**	83.0	60.8
	SENSEL-Inst-A	57.0	86.8	53.9	18.3	37.3	39.6	**49.6**	90.8	54.1	89.7	57.7
PC	F1@Cov100	65.7	91.6	47.9	25.2	28.6	36.0	38.5	87.1	44.4	84.0	54.9
	MAXPROB	65.4	94.6	50.6	27.4	33.3	38.0	43.5	91.3	43.2	92.0	57.9
	SENSEL-Inst-H	**70.3**	**96.8**	**69.0**	**33.0**	39.1	38.7	46.1	**96.5**	46.9	**94.1**	**63.1**
	SENSEL-Inst-A	64.9	94.7	65.4	30.8	37.1	40.6	47.1	95.7	44.0	90.4	61.1

Table 3: We compare our sensitivity-based selective prediction method based on instruction perturbations to the MAXPROB baseline. F1@Cov100: F1 scores when no abstention is applied. All scores except F1@Cov100 are AUC scores under the F1-coverage curves (Section 4.3). Our sensitivity-based SENSEL-Inst method consistently outperforms the MAXPROB baseline when decision boundary adjustment methods are applied.

4.1 Problem Statement

The goal of selective prediction is to *abstain* on examples that the LM is not confident about, so as to avoid giving wrong predictions (Chow, 1957). Selective prediction methods score the confidence C of each model prediction, and abstain on examples with low prediction confidence ($C < \gamma$). γ is a threshold to control the trade-off between accuracy and coverage.

4.2 Proposed Approach: SENSEL

Motivated by sensitivity analysis, we propose a sensitivity-based selective few-shot prediction method called SENSEL. SENSEL uses sensitivity to predict abstention — the LM abstains on examples where ICL predictions are highly sensitive to prompt perturbations. More formally, to score the confidence of an ICL prediction $\hat{y}_{x,I,S,\sigma}$, we create a perturbation set P consisting of prompts perturbed from (I, S, σ), and score the confidence of prediction $\hat{y}_{x,I,S,\sigma}$ as

$$\begin{aligned} C &= -s(x, I, S, \sigma) \\ &= -\frac{1}{|P|} \sum_{(I',S',\sigma') \in P} \mathbf{1}[\hat{y}_{x,I,S,\sigma} \neq \hat{y}_{x,I',S',\sigma'}]. \end{aligned}$$

We abstain the prediction when its confidence C is below the threshold γ.

4.3 Evaluation Setup

We compare our SENSEL to a selective prediction baseline MAXPROB (Hendrycks and Gimpel, 2017). We follow the setup where decision boundary adjustment is always applied. We experiment SENSEL with three different perturbation sets: SENSEL-Inst-H, SENSEL-Inst-A, and SENSEL-ExOrd. 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), and Coverage@F1 scores, which measure coverage rates at different F1 thresholds.

4.4 Results

According to Table 3 and 4, SENSEL consistently outperforms MAXPROB over three perturbation sets. Among these perturbation sets, SENSEL with human-written instruction perturbations performs the best (e.g., outperform MAXPROB by an average margin of **5.1** AUC points), which is consistent with our sensitivity study that human-written instruction perturbations have strongest correlation between sensitivity and accuracy. Using a different metric, SENSEL still achieves higher Coverage@F1 scores compared to MAXPROB (Table 5).

We conduct an error analysis to investigate SENSEL results over different tasks. We observe that MAXPROB works better (measured by AUC − F1@Cov100) on tasks with low sensitivity (correlation -0.130), while our best-performing SENSEL-Inst-H method works better on tasks with high sensitivity (correlation $+0.285$). This observation aligns with our intuition that MAXPROB does not work well under the FSL setting for tasks with high sensitivity because it relies on oversensitive model probabilities to provide stable signals for abstention. In contrast to MAXPROB, SENSEL takes advantage of ICL sensitivity, and performs well on corresponding tasks.

Adj.	Method	AG News	ARP	DBP	Emo	CARER	WikiQA	YAT	LYR	YRFS	MR	Avg
CC	F1@Cov100	54.2	94.2	78.0	29.3	29.0	50.8	29.2	90.5	38.9	70.9	56.5
	MAXPROB	52.5	**98.5**	87.8	34.0	32.6	50.5	31.0	96.2	37.1	74.9	59.5
	SENSEL-ExOrd	54.0	97.7	85.1	**35.5**	**33.4**	**51.1**	**34.6**	**97.3**	**40.3**	76.8	60.6
PC	F1@Cov100	64.5	94.5	73.3	20.8	25.3	41.9	23.8	90.7	39.6	84.3	55.9
	MAXPROB	66.5	97.5	76.7	25.3	27.7	43.5	27.4	94.4	38.5	91.9	58.9
	SENSEL-ExOrd	**66.6**	97.7	**91.9**	25.9	31.5	44.2	30.9	**97.3**	**40.3**	**92.4**	**61.9**

Table 4: We compare our sensitivity-based selective prediction method based on example ordering perturbations to the MAXPROB baseline. F1@Cov100: F1 scores when no abstention is applied. All scores except F1@Cov100 are AUC scores under the F1-coverage curves (Section 4.3). Our sensitivity-based SENSEL-ExOrd method consistently outperforms MAXPROB when decision boundary adjustment methods are applied.

	C@10	C@20	C@30	C@40	C@50	C@60	C@70	C@80	C@90
AG News	100 (100)	100 (100)	100 (100)	100 (99)	99 (96)	93 (83)	74 (62)	19 (22)	6 (4)
ARP	100 (100)	100 (100)	100 (100)	100 (100)	100 (100)	100 (99)	99 (97)	98 (96)	97 (93)
DBP	100 (100)	100 (100)	99 (98)	91 (73)	80 (49)	66 (30)	50 (16)	35 (5)	19 (1)
Emo	100 (100)	88 (78)	64 (47)	29 (11)	7 (2)	1 (0)	1 (0)	0 (0)	0 (0)
CARER	100 (100)	100 (100)	88 (74)	46 (16)	16 (3)	1 (0)	0 (0)	0 (0)	0 (0)
WikiQA	100 (100)	98 (94)	84 (76)	61 (50)	9 (17)	0 (1)	0 (0)	0 (0)	0 (0)
YAT	100 (100)	100 (100)	100 (99)	81 (74)	35 (25)	4 (4)	1 (0)	0 (0)	0 (0)
LYR	100 (100)	100 (100)	100 (100)	100 (100)	100 (100)	100 (100)	100 (100)	100 (100)	95 (74)
YRFS	100 (100)	100 (100)	100 (100)	95 (88)	45 (34)	4 (2)	1 (0)	0 (0)	0 (0)
MR	100 (100)	100 (100)	100 (100)	100 (100)	100 (100)	100 (100)	100 (100)	99 (98)	89 (71)

Table 5: We report the Coverage@F1 scores for our SELSEN method under the prototypical calibration setting, and compare to the MAXPROB baseline (scores shown in parentheses). C@X: coverage rate at F1 score X. Our method consistently gets higher Coverage@F1 scores compared to the MAXPROB baseline on all ten tasks across different F1 thresholds. We color the scores green if our SENSEL method outperforms the MAXPROB baseline by at least 10 coverage points.

5 Conclusion

ICL oversensitivity is a widely-known issue, but it has been treated as an independent artifact. We argue that ICL sensitivity is not merely an artifact, but reflects how well and confidently a LM understands the task (whether learned during pre-training or from the few-shot examples). Our study shows that ICL sensitivity is negatively correlated with accuracy. Based on this observation, we propose a new few-shot selective prediction method based on sensitivity signal, which outperforms a baseline based on model probabilities.

Limitations and Future Directions While our work takes a step towards understanding ICL sensitivity and few-shot selective prediction, there are many questions left unanswered. First, it's unclear if there are systematic patterns for prompts that achieve better ICL performance, and if so, why the pattern occurs. Second, while SENSEL significantly outperforms MAXPROB, there is still large room for improvement as SENSEL still frequently makes wrong abstention decisions (i.e., predict on incorrect predictions or abstain on correct predictions). As ICL is gradually deployed into real-world applications (with the help of larger LMs), future research should investigate on better selective prediction methods. Third, in this work we explore SENSEL for few-shot selective prediction, and leave future work to explore SENSEL as a confidence estimator for model calibration and confidence-based methods (e.g., uncertainty-based active learning).

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*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*.

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*.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of Empirical Methods in Natural Language Processing*.

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*.

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. 2017. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*.

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*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the International Conference on Machine Learning*.

Michael Hahn, Dan Jurafsky, and Richard Futrell. 2021. Sensitivity as a Complexity Measure for Sequence Classification Tasks. *Transactions of the Association for Computational Linguistics*.

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*.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering. *Transactions of the Association for Computational Linguistics*.

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

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*.

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. 2022. MetaICL: Learning to learn in context. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Khanh Nguyen and Brendan O'Connor. 2015. Posterior calibration and exploratory analysis for natural language processing models. In *Proceedings of Empirical Methods in Natural Language Processing*.

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*.

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*.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the European Chapter of the Association for Computational Linguistics*.

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*.

Chenglei Si, Chen Zhao, Sewon Min, and Jordan Boyd-Graber. 2022. Revisiting calibration for question answering. *ArXiv*.

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

Ji Xin, Raphael Tang, Yaoliang Yu, and Jimmy Lin. 2021. The art of abstention: Selective prediction and error regularization for natural language processing. In *Proceedings of the Association for Computational Linguistics*.

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*.