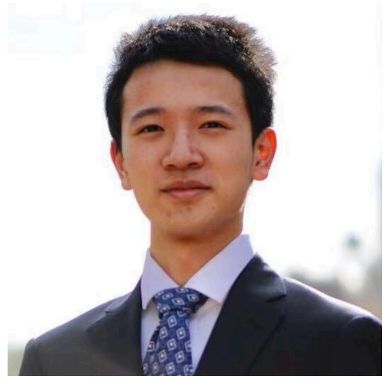
On the Relation between Sensitivity and Accuracy in In-Context Learning

Yanda Chen[1], Chen Zhao[2,3], Zhou Yu[1], Kathleen McKeown[1], He He[2]





[1]Columbia University, [2]New York University, [3]NYU Shanghai

Topic Introduction - Few-shot Learning

- Def: quickly learns a ***new*** task with ***few*** labeled examples

Adapt x_1: "I like the movie!", y_1 = Positive 😃
x_2: "Horrible movie!", y_2 = Negative 😔

Predict x^{target}: "The movie is boring.", y^{target}: ?

- Why we care?
  - save annotation efforts
  - human-like AI

In-context Learning (ICL)

$I \circ x_1 \circ y_1 \circ x_2 \circ y_2 \circ x^{\text{target}}$ ____ $\longrightarrow \hat{y}^{\text{target}}$

What is the sentiment of this review? I like the movie! Positive. Horrible movie! Negative. This movie is boring. ______ —> Negative

We study the Sensitivity of ICL

What do we measure?

Magnitude of the changes in the predicted label when the prompt is **perturbed**.

$$\frac{1}{|P|}\sum_{p' \in P} \mathbf{1}[f(x,p) = f(x,p')]$$

What do we perturb?

- instruction wording
  - human-written perturbation (InstH)
  - automatic perturbation (InstA): word dropout & paraphrase
- few-shot example ordering (ExOrd)

Data & Models?

- Models: GPT-J-6B, GPT-Neo-2.7B
- Dataset: 10 classification datasets (sentiment, emotion, topic, question-answering)

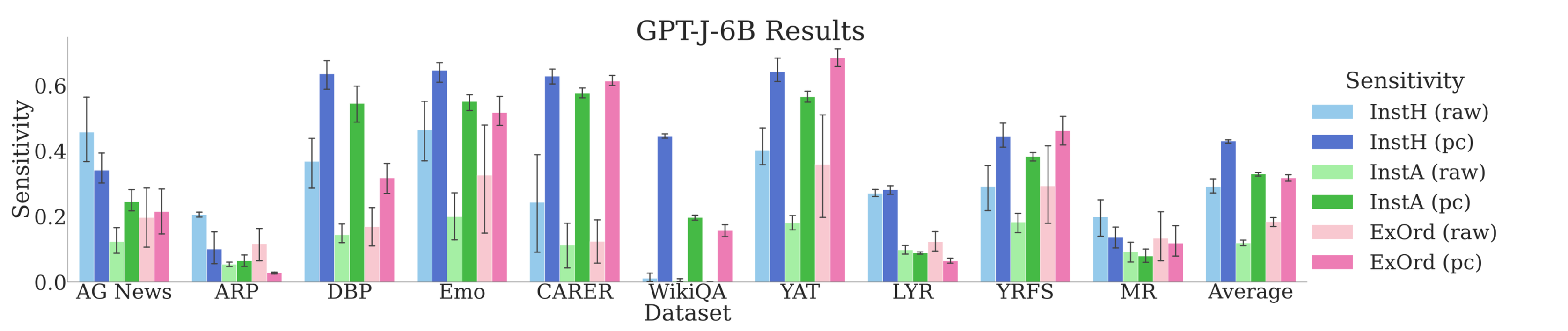
Finding #1: Sensitivity is underestimated due to label bias.

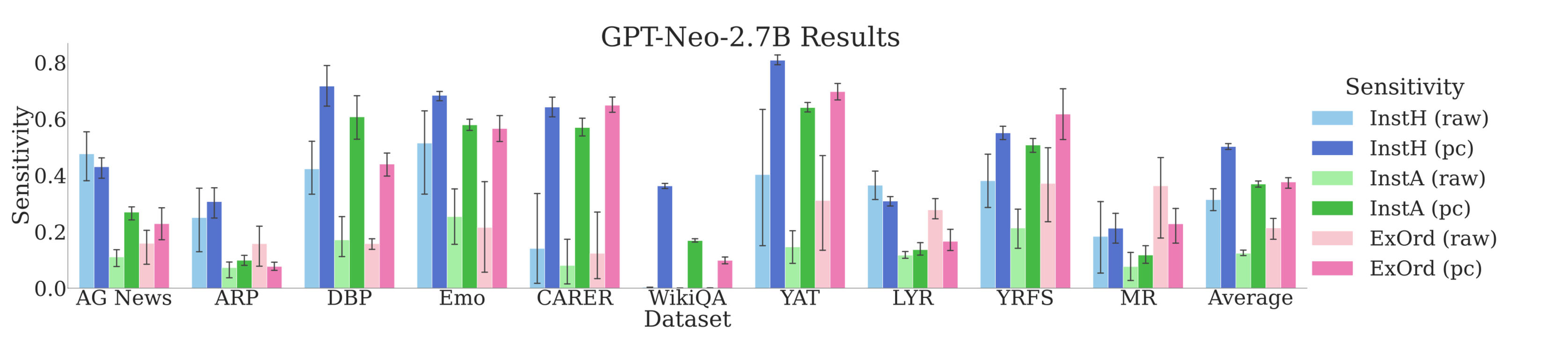
Label bias: LMs assign a higher probability to a specific label [Zhao et al., 2021]

=> appear to make stable predictions under prompt perturbations

How does ICL sensitivity change after addressing label bias?

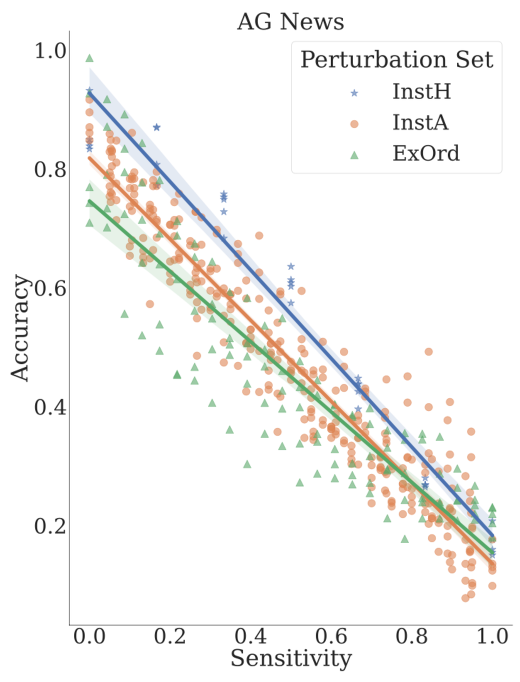
Prototypical Calibration [Han et al. 2022]: cluster predictions into gaussian mixtures.





ICL sensitivity is 99.0% higher after addressing label bias.

Finding #2: Sensitivity is negatively correlated to accuracy.



Model	Perturb Set	AG News	ARP	DBP	Emo	CARER	WikiQA	YAT	LYR	YRFS	MR	Avg
GPT-J-6B	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)
GPT-Neo-2.7B	InstH	−0.49 (0.04)	−0.57 (0.04)	−0.53 (0.14)	−0.09 (0.12)	−0.36 (0.04)	−0.36 (0.03)	−0.25 (0.02)	−0.54 (0.09)	−0.21 (0.07)	−0.48 (0.03)	−0.39 (0.02)
	InstA	−0.39 (0.03)	−0.22 (0.02)	−0.61 (0.13)	−0.09 (0.06)	−0.36 (0.05)	−0.10 (0.03)	−0.41 (0.03)	−0.19 (0.07)	−0.17 (0.07)	−0.28 (0.05)	−0.28 (0.02)
	ExOrd	−0.27 (0.06)	−0.48 (0.06)	−0.76 (0.04)	−0.21 (0.12)	−0.29 (0.06)	−0.33 (0.04)	−0.29 (0.14)	−0.46 (0.08)	−0.14 (0.04)	−0.28 (0.07)	−0.35 (0.04)

Pearson correlations (std is shown in parenthesis)

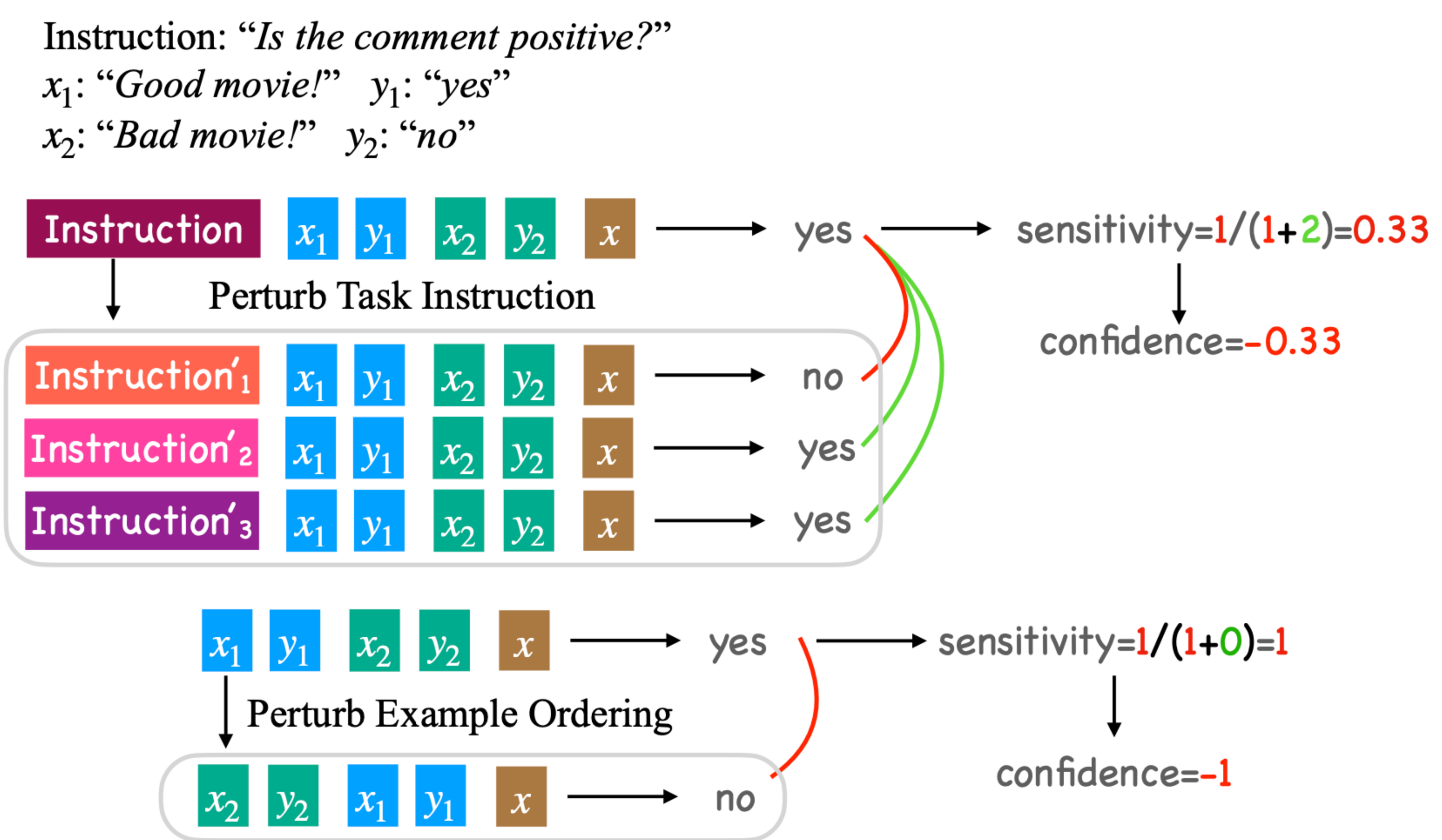
Application: Sensitivity-based Selective Few-shot Prediction (SenSel)

Goal: Abstain on examples that the model is *not* confident about => avoid presenting wrong predictions to users

Score model confidence C on each example, abstain on examples where $C < \gamma$. γ controls the trade-off between coverage and accuracy.

SenSel: $C = -$ prediction's sensitivity to prompt perturbations.

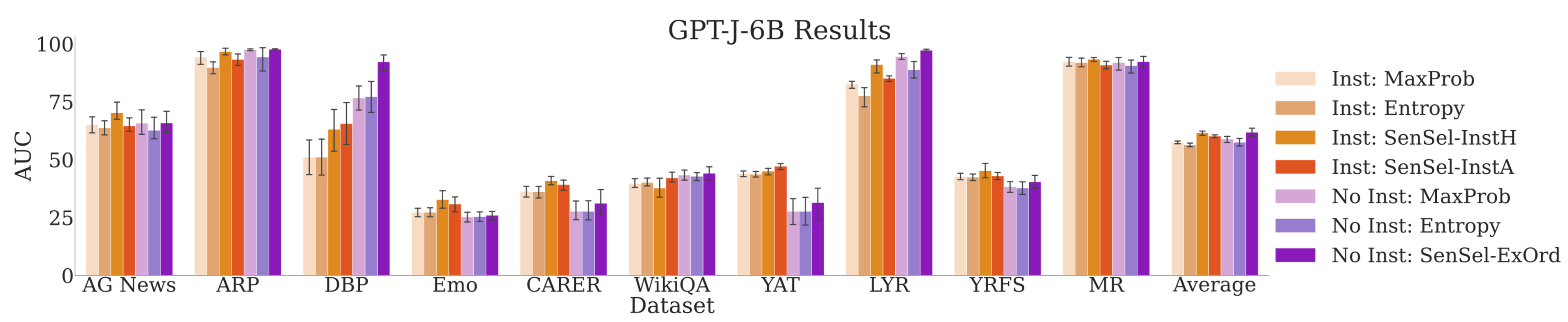
Sensitivity-Based Selective Few-shot Prediction

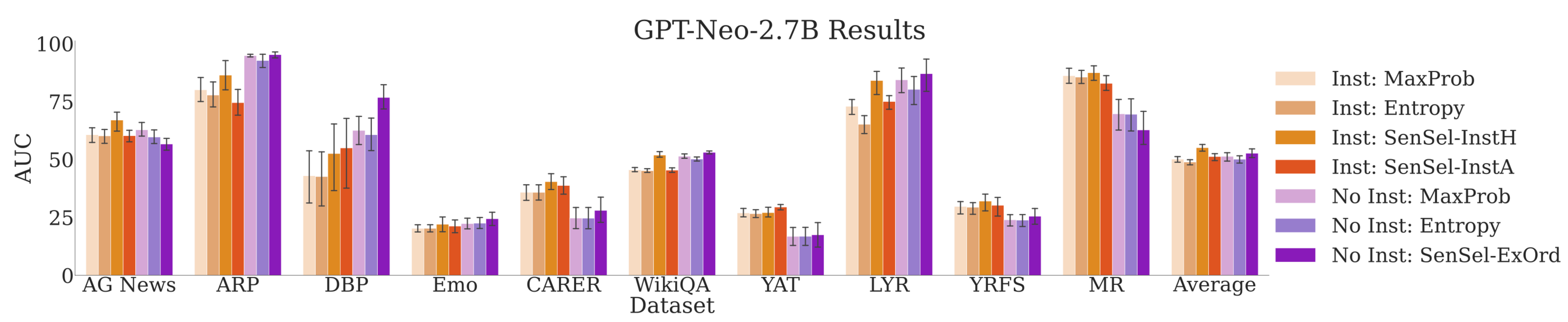


Metric: Area under the F1-Coverage curve (AUC) - measures the average F1-score at different coverage rates.

Baselines:

- MaxProb: $C =$ maximum output probability over the labels
- Entropy: $C = -$entropy of the output probabilities over the labels





SenSel outperforms baselines (by +3.5 AUC points on GPT-J-6B and +3.2 AUC points on GPT-Neo-2.7B)