

SteerConf: Steering LLMs for Confidence Elicitation

Ziang Zhou, Tianyuan Jin, Jieming Shi, Qing Li

NeurIPS 2025

SteerConf: Steering LLMs for Confidence Elicitation

Ziang Zhou, Tianyuan Jin, Jieming Shi, Qing Li

NeurIPS 2025

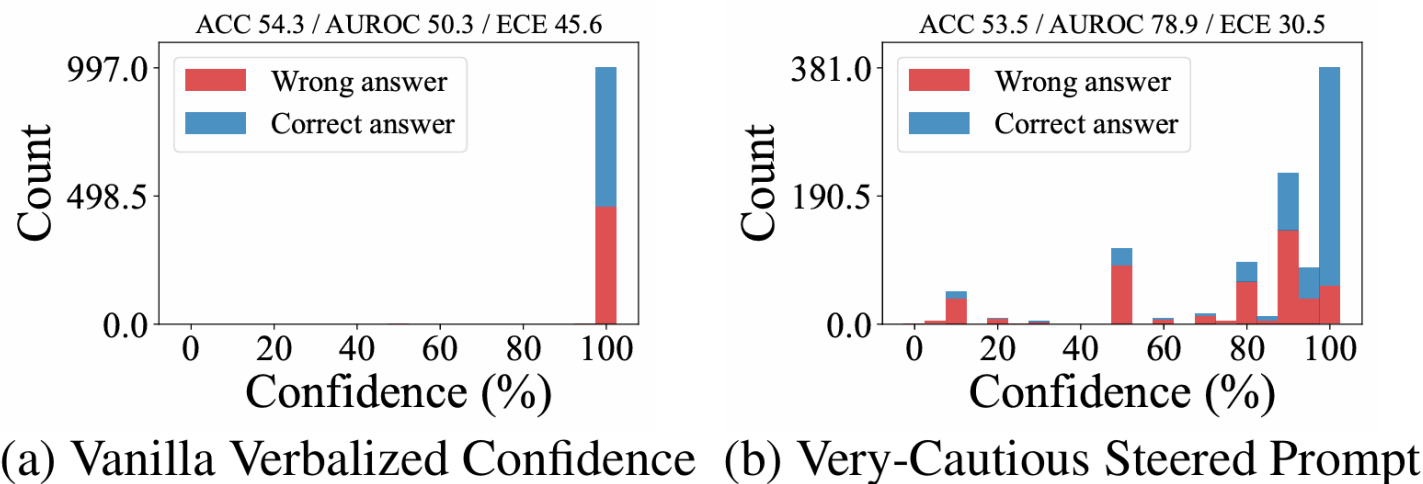
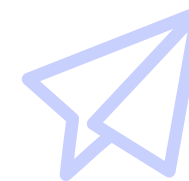


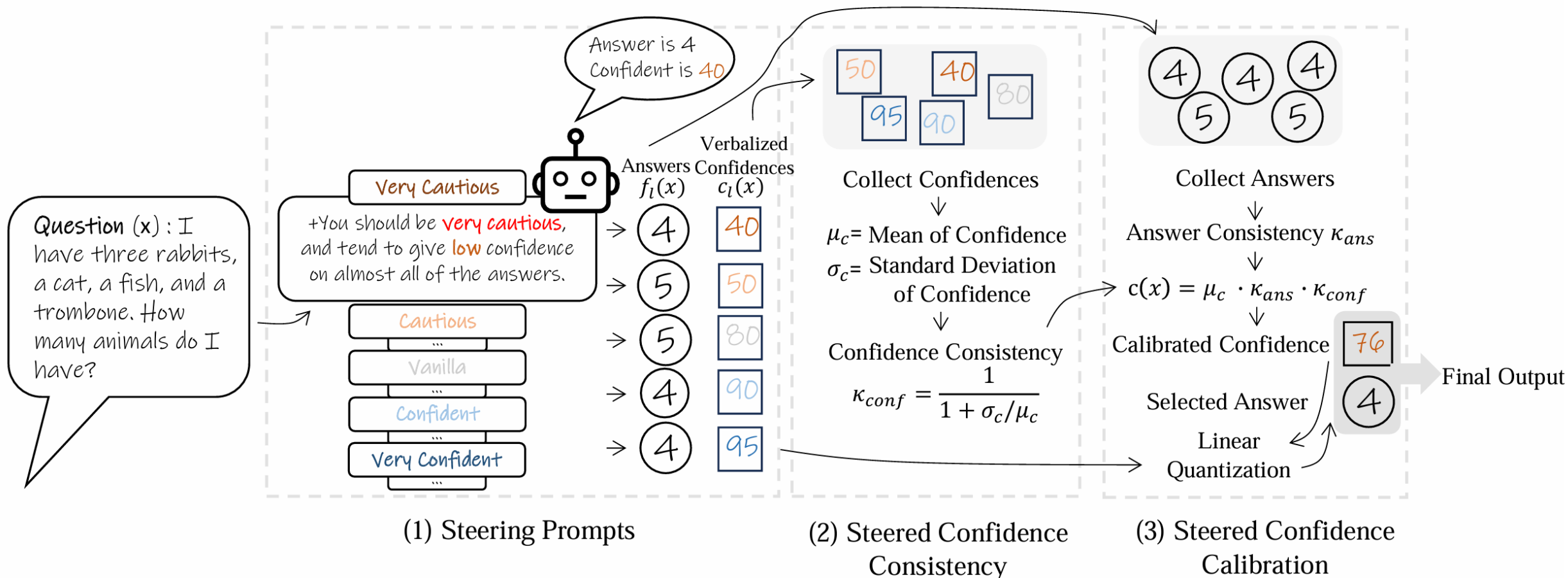
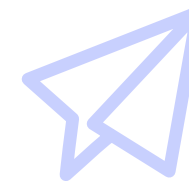
Figure 1: Vanilla and *very cautious* Prompt Steered Confidence on Object Counting data with GPT-3.5.

- *Vanilla Prompt*: “Read the question carefully, analyze step by step, and provide your answer along with your confidence. Confidence reflects how likely you believe your answer is correct.”
- *Very Cautious Prompt*: The Vanilla Prompt with an additional instruction: “You should adopt a very cautious approach and assign low confidence to most of your answers.”

SteerConf: Steering LLMs for Confidence Elicitation

Ziang Zhou, Tianyuan Jin, Jieming Shi, Qing Li

NeurIPS 2025



SteerConf: Steering LLMs for Confidence Elicitation

Ziang Zhou, Tianyuan Jin, Jieming Shi, Qing Li

NeurIPS 2025



Model	GSM8K	DateUnd	ObjCnt	StrategyQA	Sport	Law	Ethics	Avg
LLaMA3(CoT)	5.0	13.7	8.7	11.8	7.7	22.8	7.8	11.1
+SteerConf	7.8	1.0	6.5	8.6	5.4	20.9	5.0	7.9
GPT-3.5(CoT)	20.3	30.8	41.8	26.0	20.5	44.3	24.8	29.8
+SteerConf	6.5	6.7	19.7	14.5	10.8	16.0	15.1	12.7
GPT-4(CoT)	6.5	6.6	4.9	18.5	9.2	23.0	6.1	10.7
+SteerConf	4.3	8.4	1.6	11.5	5.6	9.9	9.3	7.2

(a) ECE ↓ (Lower is Better)

Model	GSM8K	DateUnd	ObjCnt	StrategyQA	Sport	Law	Ethics	Avg
LLaMA3(CoT)	12.4	13.4	8.7	29.7	38.3	38.2	26.1	23.8
+SteerConf	41.3	24.0	49.0	43.2	47.5	50.6	51.7	43.9
GPT-3.5(CoT)	26.5	28.0	42.0	36.8	46.9	55.5	45.7	40.2
+SteerConf	70.9	49.7	74.7	48.9	52.2	60.9	75.0	61.8
GPT-4(CoT)	10.5	47.6	4.9	48.7	27.7	46.2	83.2	38.4
+SteerConf	59.3	61.8	44.9	38.9	48.7	55.4	74.9	54.9

(c) PR-N ↑ (Higher is Better)

Model	GSM8K	DateUnd	ObjCnt	StrategyQA	Sport	Law	Ethics	Avg
LLaMA3(CoT)	55.1	54.3	50.0	64.6	74.1	54.3	54.2	58.1
+SteerConf	81.2	70.7	87.6	74.5	79.4	64.7	81.2	77.0
GPT-3.5(CoT)	56.2	49.8	50.1	56.4	62.7	53.0	65.2	56.2
+SteerConf	85.6	76.0	82.5	67.5	66.4	61.5	86.7	75.2
GPT-4(CoT)	52.1	75.1	50.0	68.8	65.0	59.5	87.6	65.5
+SteerConf	86.0	81.5	93.3	70.3	75.2	68.4	93.0	81.1

(b) AUROC ↑ (Higher is Better)

Model	GSM8K	DateUnd	ObjCnt	StrategyQA	Sport	Law	Ethics	Avg
LLaMA(CoT)	95.4	89.3	91.3	85.2	89.2	66.5	83.0	85.7
+SteerConf	97.8	92.6	98.1	90.0	92.8	73.6	93.1	91.1
GPT-3.5(CoT)	80.5	71.5	58.2	71.5	70.2	48.4	75.8	68.0
+SteerConf	92.1	88.4	83.2	77.7	75.0	57.3	91.8	80.8
GPT-4(CoT)	93.7	94.2	95.1	81.4	88.0	68.3	92.8	87.6
+SteerConf	97.3	94.0	99.2	87.6	91.1	75.6	97.9	91.8

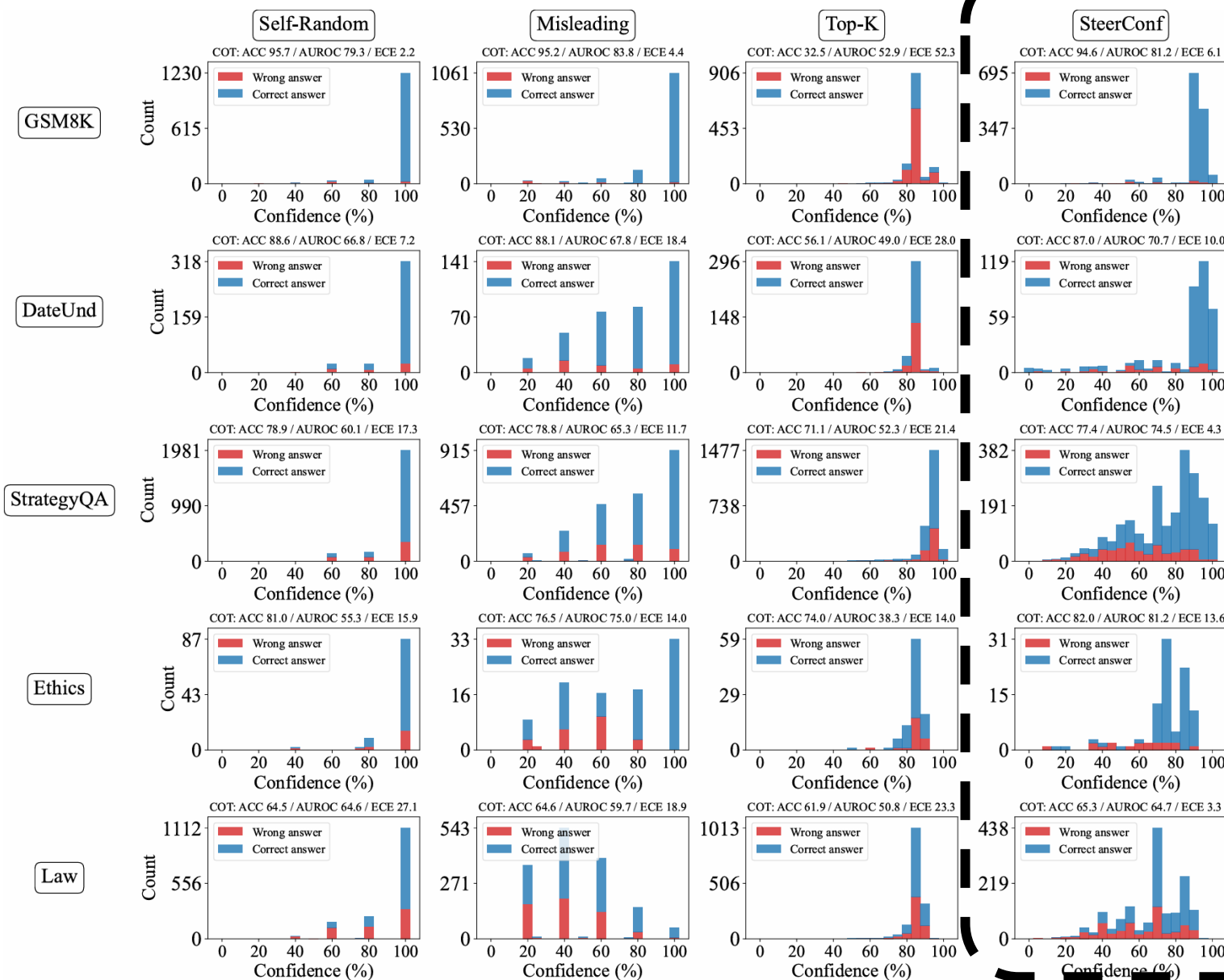
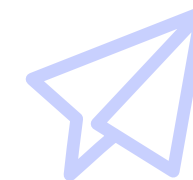
(d) PR-P ↑ (Higher is Better)

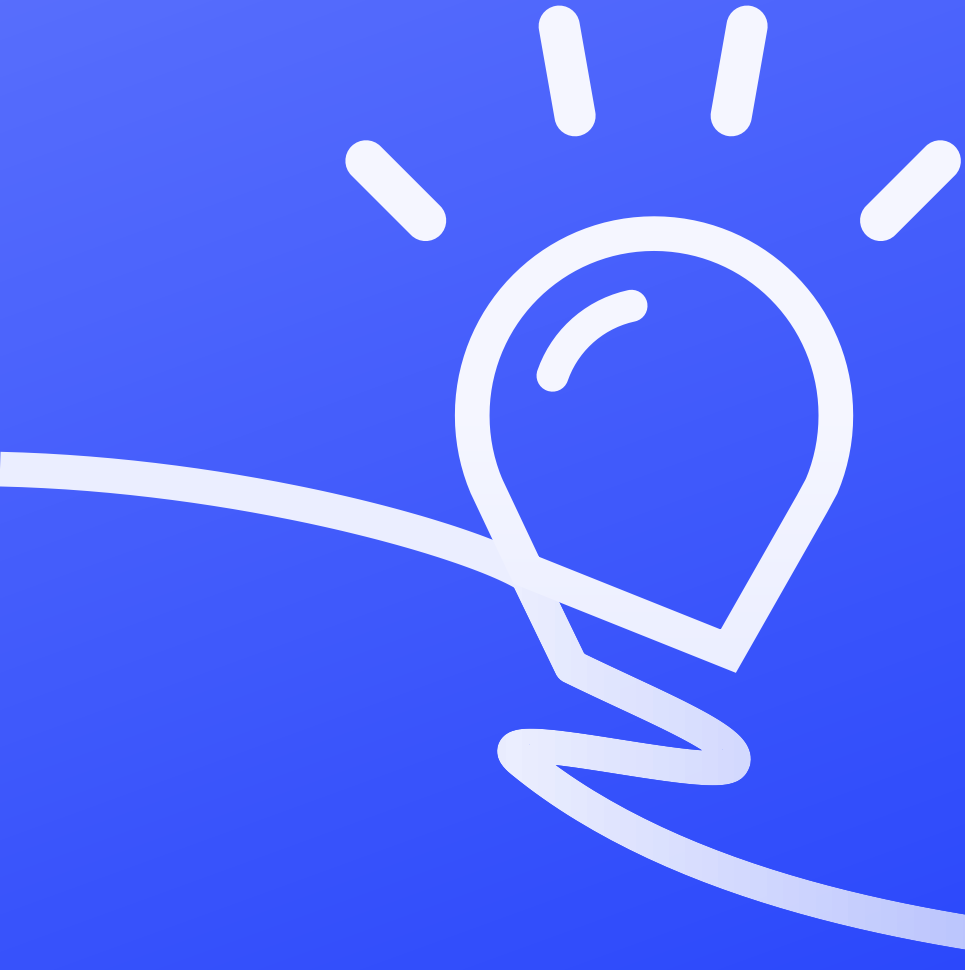
Table 2: Comparing SteerConf with verbalized confidence elicitation of LLMs with CoT. ECE > 0.25,

SteerConf: Steering LLMs for Confidence Elicitation

Ziang Zhou, Tianyuan Jin, Jieming Shi, Qing Li

Neurips25





Thank you

Speaker: Ziang Zhou