The Fourteenth International Conference on Learning Representations (ICLR 2026)

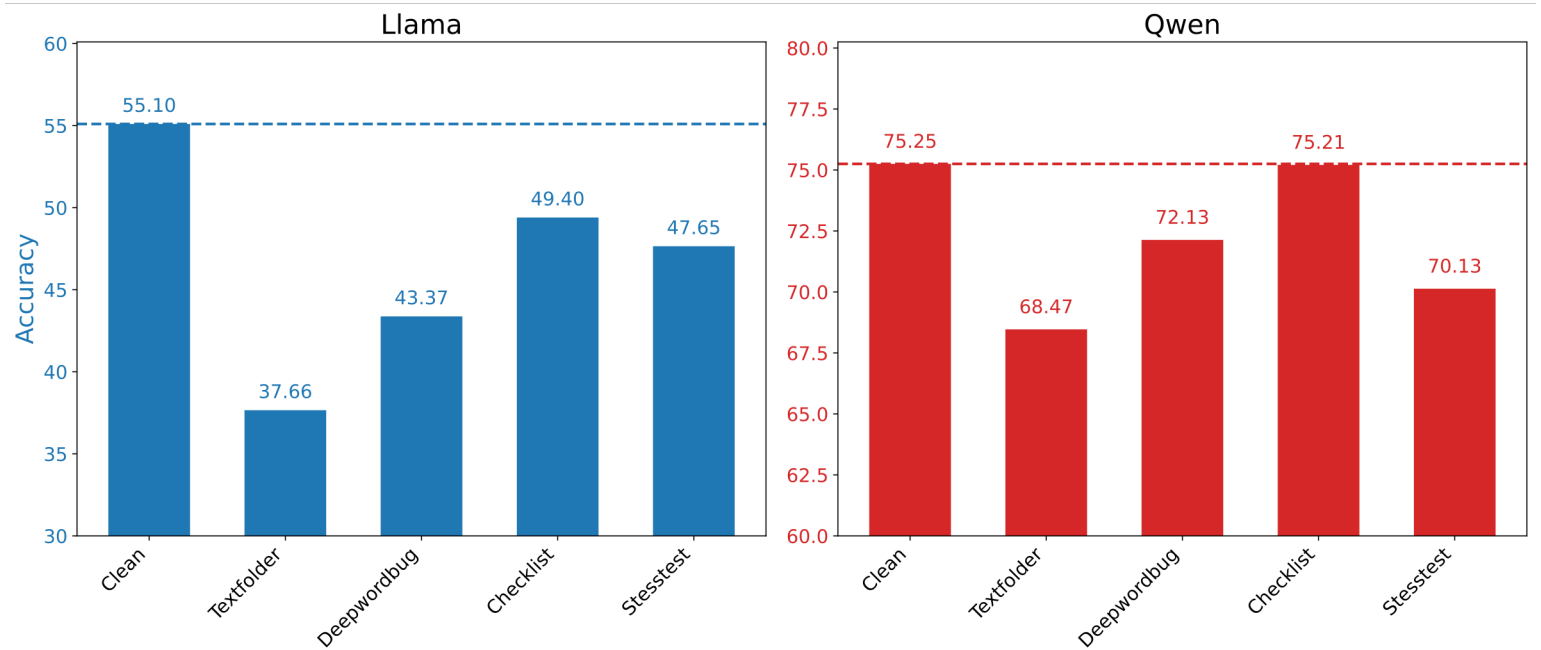
Towards Self-Robust LLMs: Intrinsic Prompt Noise Resistance via CoIPO

Xin Yang[1, 3], Letian Li[2], Abudukelimu Wuerkaixi[2, 3], Xuxin Cheng[3], Cao Liu[3], Ke Zeng[3], **Xunliang Cai[3]**, **Wenyuan Jiang[4]**

[1] Zhejiang University
[2] Tsinghua University
[3] Meituan LongCat Interaction Team
[4] ETH Zürich

Problem: LLM Prompt Robustness

- Large Language Models are highly sensitive to **prompt variations**
- Real-world prompts often contain:
  - Spelling Errors
  - Word Substitutions
  - Irrelevant Content
  - ……
- Even small perturbations can **significantly degrade** performance



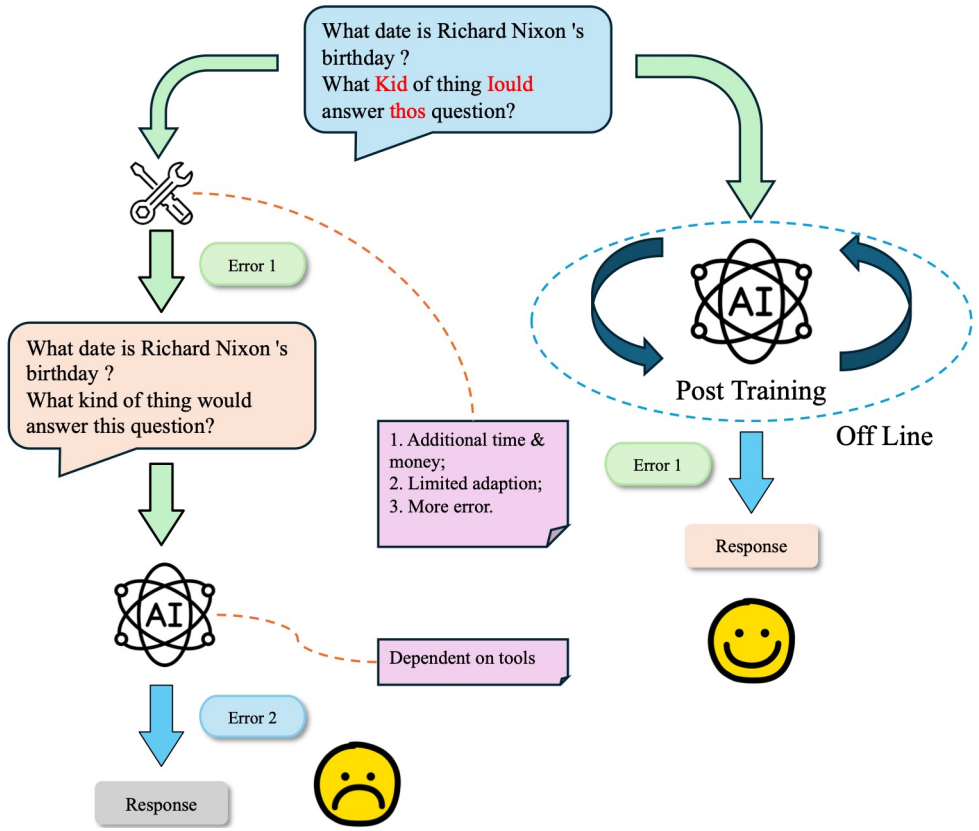
Existing Solutions and Limitations

Current methods

- Grammar correction tools
- Prompt rewriting
- LLM-based prompt optimization

Limitations

- Extra computational overhead
- Limited adaption
- Pipeline cascading errors



Problem Formulation

Definition:

- Ideal prompt: $\tilde{P}$
- Clean prompt: $\hat{P}$
- Noisy prompt: P'

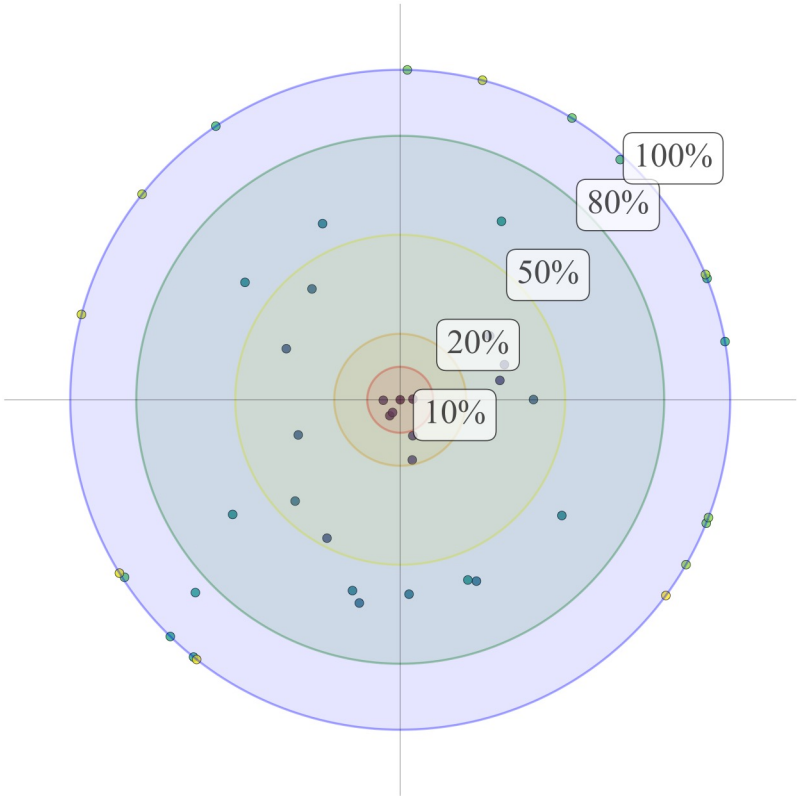
$\tilde{P}$ is too difficult to get, So we define noisy prompt as:

$P' = \hat{P} + N$

N: Perturbation operation

To measure the effect of noise, we define:

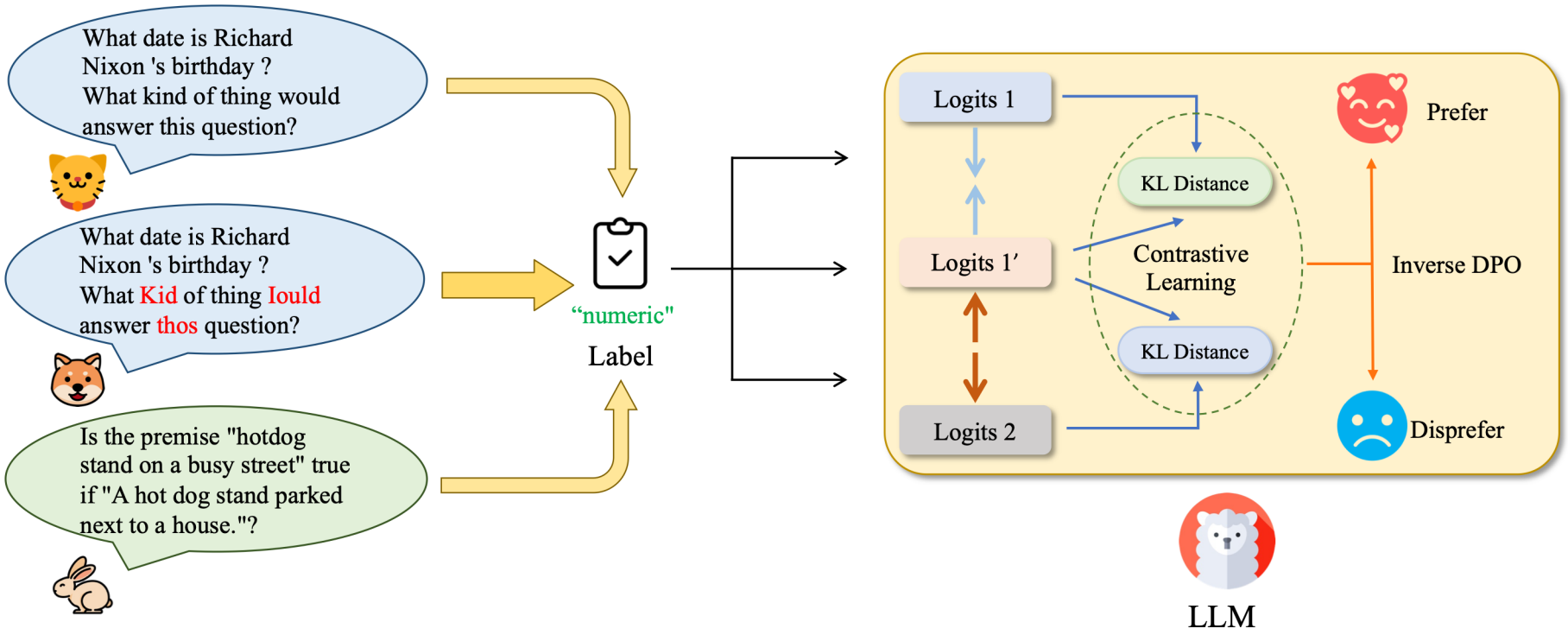
$r = d(P', \hat{P})$, where d denotes the performance gap.



Proposed Solution: CoIPO

Overview

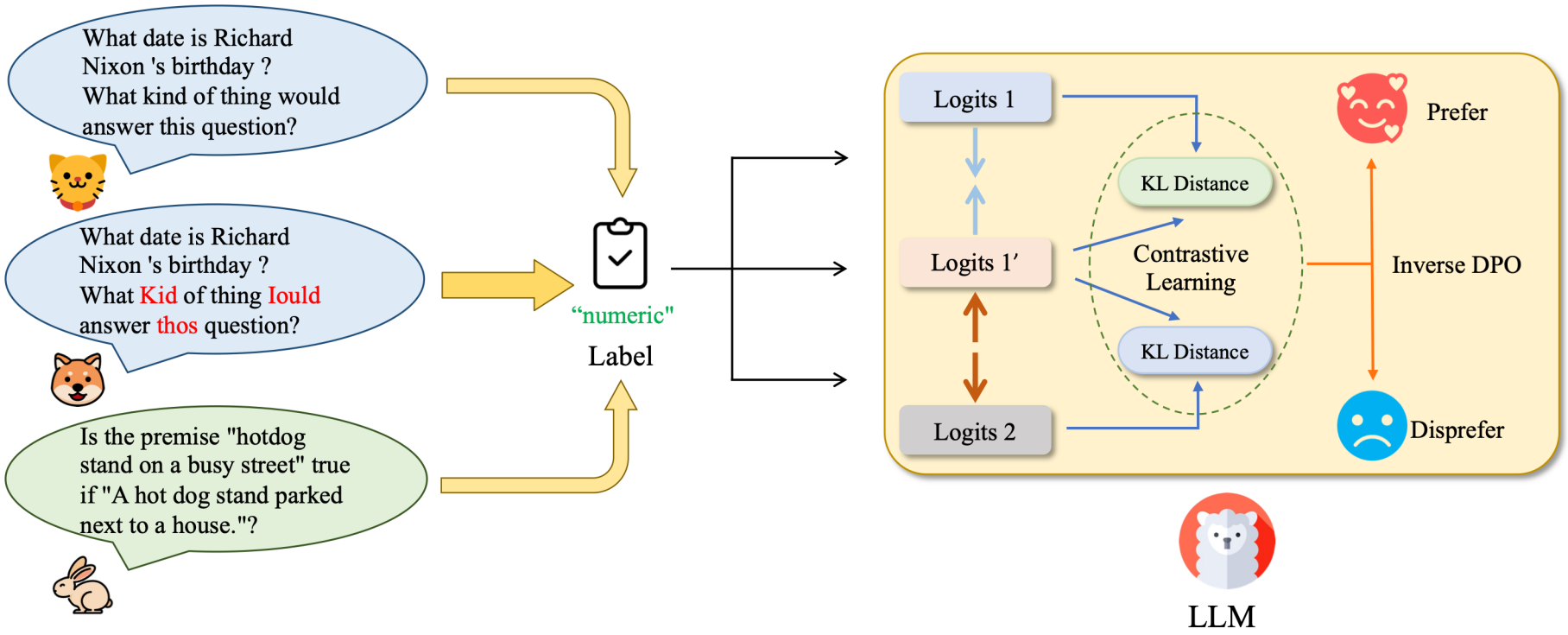
- Enhance **intrinsic** prompt robustness of LLMs
- Construct (clean prompt, noisy prompt) pair
- Align model outputs under clean and noisy prompts



Proposed Solution: CoIPO

Core idea

- Contrastive learning
- Inversed DPO: different inputs -> same output



Proposed Solution: CoIPO

Formula Derivation

Loss function of CoIPO:

$$\mathcal{L}_{\text{invDPO}} = -D\left(\hat{P}_2 \,\middle|\, P_1', y_1\right) + D\left(\hat{P}_1 \,\middle|\, P_1', y_1\right),$$

For function D:

$$\begin{aligned} D(P \mid P_{\text{ref}}, y) &= \sum_{t \in \mathcal{T}_y} \text{KL}\Big(\text{softmax}(\mathcal{M}_y(\ell_t(P_{\text{ref}} \oplus y))) \parallel \text{softmax}(\mathcal{M}_y(\ell_t(P \oplus y))) \Big) \\ &= \sum_{t \in \mathcal{T}_y} \text{KL}\Big(p_t^{(P_{\text{ref}}, y)} \parallel p_t^{(P, y)} \Big). \end{aligned}$$

Then we get:

$$\mathcal{L} = -\sum_{t \in \mathcal{T}_{y_1}} \text{KL}\big(p_t^{(P_1', y_1)} \| p_t^{(\hat{P}_2, y_1)}\big) + \sum_{t \in \mathcal{T}_{y_1}} \text{KL}\big(p_t^{(P_1', y_1)} \| p_t^{(\hat{P}_1, y_1)}\big),$$

Experiment Results

Performance Comparison of Llama Under Different Perturbations and Datasets. Acc means accuracy score (%), Diff means score difference compared to clean (%).

Perturbation	Method	MNLI		MRPC		QNLI		QQP		SST2		Avg	
		Acc	Diff	Acc	Diff	Acc	Diff	Acc	Diff	Acc	Diff	Acc	Diff
Clean	Base	51.00	/	62.04	/	52.54	/	29.87	/	80.04	/	55.10	/
	SFT	56.13	/	41.04	/	46.54	/	58.62	/	84.04	/	57.28	/
	COIN	61.58	/	56.92	/	53.79	/	50.67	/	86.38	/	61.87	/
	CoIPO	61.50	/	64.50	/	56.21	/	66.04	/	86.75	/	**67.00**	/
TextFolder	Base	39.96	11.04	42.12	19.92	41.00	11.54	13.17	16.71	52.04	28.00	37.66	17.44
	SFT	54.83	1.29	39.79	1.25	43.08	3.46	51.33	7.29	80.29	3.75	53.87	**3.41**
	COIN	52.83	8.75	51.71	5.21	48.75	5.04	42.12	8.54	85.13	1.25	56.11	5.76
	CoIPO	58.00	3.50	64.38	0.13	51.12	5.08	52.17	13.88	85.42	1.33	**62.22**	4.78
DeepWordBug	Base	42.62	8.38	47.29	14.75	48.00	4.54	17.50	12.37	61.46	18.58	43.37	11.73
	SFT	52.83	3.29	39.58	1.46	44.37	2.17	51.04	7.58	82.71	1.33	54.11	**3.17**
	COIN	51.92	9.67	51.34	5.58	45.13	8.67	39.46	11.21	86.29	0.08	54.83	7.04
	CoIPO	57.33	4.17	61.50	3.00	51.04	5.17	48.54	17.50	86.54	0.21	**60.99**	6.01
CheckList	Base	47.71	3.29	55.62	6.42	47.67	4.88	20.25	9.62	75.75	4.29	49.40	5.70
	SFT	54.13	2.00	40.25	0.79	40.33	6.21	57.17	1.46	81.46	2.58	54.67	2.61
	COIN	59.58	2.00	55.08	1.83	52.00	1.79	50.21	0.46	85.04	1.33	60.38	**1.48**
	CoIPO	59.38	2.12	63.67	0.83	53.58	2.62	64.50	1.54	85.83	0.92	**65.39**	1.61
StressTest	Base	45.46	5.54	51.33	10.71	53.33	-0.79	21.00	8.87	67.12	12.92	47.65	7.45
	SFT	50.17	5.96	39.54	1.50	50.12	-3.58	50.04	8.58	78.50	5.54	53.68	3.60
	COIN	54.46	7.12	58.33	-1.42	55.58	-1.79	45.75	4.92	84.96	1.42	59.82	**2.05**
	CoIPO	55.71	5.79	66.29	-1.79	57.08	-0.88	55.04	11.00	85.33	1.42	**63.89**	3.11
Avg	Base	45.35	7.06	51.68	12.95	48.51	5.04	20.36	11.89	67.28	15.95	46.64	10.58
	SFT	53.62	**3.14**	40.04	1.25	44.89	**2.06**	53.64	**6.23**	81.40	3.30	54.72	**3.20**
	COIN	56.08	6.89	54.68	2.80	51.05	3.43	45.64	6.28	85.56	1.02	58.60	4.08
	CoIPO	**58.38**	3.89	**64.07**	**0.54**	**53.81**	3.00	**57.26**	10.98	**85.98**	**0.97**	**63.90**	3.88

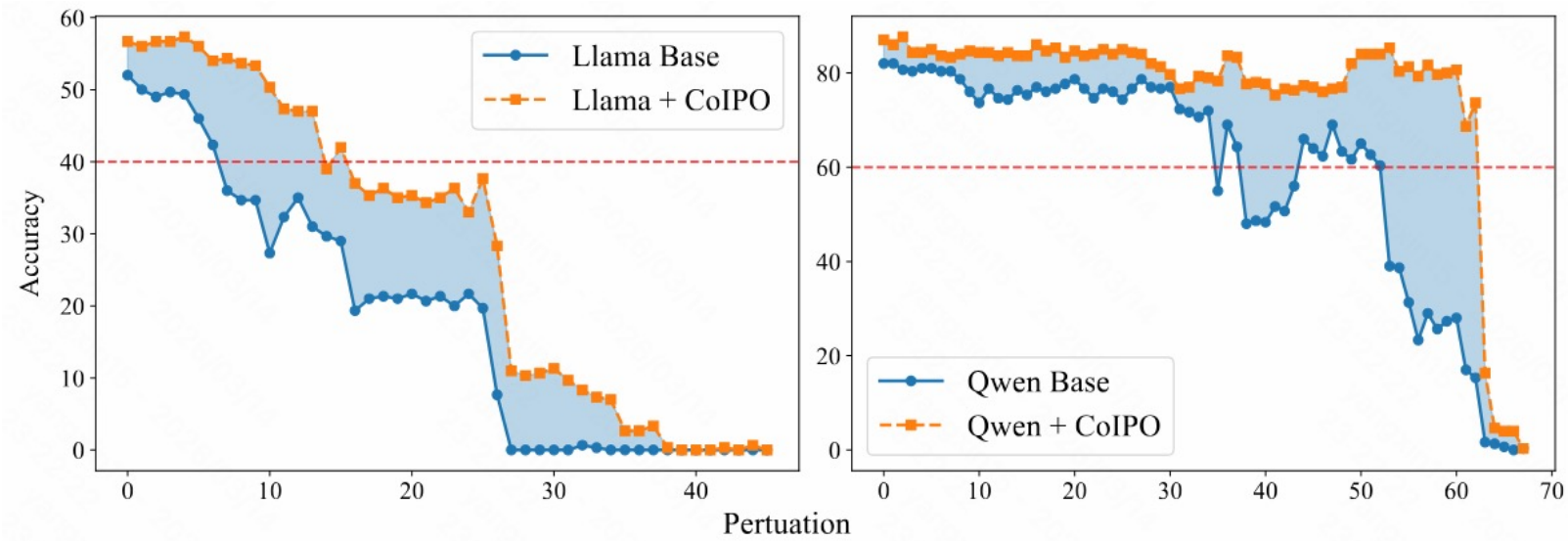
Experiment Results

Performance Comparison of Qwen Under Different Perturbations and Datasets. Acc means accuracy score (%), Diff means score difference compared to clean (%).

Perturbation	Method	MNLI		MRPC		QNLI		QQP		SST2		Avg	
		Acc	Diff	Acc	Diff	Acc	Diff	Acc	Diff	Acc	Diff	Acc	Diff
Clean	Base	83.29	/	61.16	/	71.92	/	70.67	/	89.21	/	75.25	/
	SFT	80.21	/	74.12	/	71.96	/	72.54	/	90.87	/	77.94	/
	COIN	87.71	/	76.46	/	77.08	/	80.83	/	92.54	/	82.93	/
	CoIPO	86.75	/	78.83	/	78.75	/	84.08	/	91.00	/	**83.88**	/
TextFolder	Base	76.79	6.50	58.33	2.83	67.62	4.29	61.50	9.17	78.08	11.12	68.47	6.78
	SFT	80.83	-0.62	71.08	3.04	67.92	4.04	68.46	4.08	86.67	4.21	74.99	2.95
	COIN	87.58	0.12	75.88	0.58	73.33	3.75	71.79	9.04	91.54	1.00	80.03	2.90
	CoIPO	87.21	-0.46	72.62	6.21	77.46	1.29	84.04	0.04	90.00	1.00	**82.27**	**1.62**
DeepWordBug	Base	77.17	6.13	53.75	7.42	70.88	1.04	70.83	-0.17	88.04	1.17	72.13	3.12
	SFT	79.83	0.37	72.42	1.71	76.79	-4.83	71.29	1.25	89.83	1.04	78.03	**-0.09**
	COIN	87.08	0.62	76.83	-0.38	79.62	-2.54	73.87	6.96	91.33	1.21	81.75	1.18
	CoIPO	86.38	0.37	78.83	0.00	81.33	-2.58	82.08	2.00	90.96	0.04	**83.92**	-0.03
CheckList	Base	81.58	1.71	64.83	-3.67	65.17	6.75	74.21	-3.54	90.25	-1.04	75.21	0.04
	SFT	80.62	-0.42	73.71	0.42	71.50	0.46	66.54	6.00	91.42	-0.55	76.76	1.18
	COIN	88.08	-0.38	75.54	0.92	74.38	2.71	78.21	2.62	92.67	-0.12	81.77	1.15
	CoIPO	86.04	0.71	79.38	-0.54	78.79	-0.04	84.29	-0.21	91.37	-0.37	**83.97**	**-0.09**
StressTest	Base	82.75	0.54	52.17	9.00	69.12	2.79	64.12	6.54	82.50	6.71	70.13	5.12
	SFT	80.21	0.00	73.08	1.04	71.71	0.25	71.25	1.29	86.42	4.46	76.53	1.41
	COIN	86.96	0.75	76.17	0.29	74.79	2.29	74.71	6.12	91.92	0.62	80.91	2.02
	CoIPO	86.21	0.54	78.12	0.71	77.54	1.21	83.54	0.54	90.62	0.38	**83.21**	**0.68**
Avg	Base	80.32	3.72	58.05	3.89	68.94	3.72	68.27	3.00	85.62	4.49	72.24	3.76
	SFT	80.34	**-0.17**	72.88	1.55	71.97	-0.02	70.02	3.16	89.04	2.29	76.85	1.36
	COIN	**87.48**	0.28	76.18	**0.35**	75.84	1.55	75.88	6.19	**92.00**	0.68	81.48	1.81
	CoIPO	86.52	0.29	**77.56**	1.59	**78.77**	**-0.03**	**83.61**	**0.59**	90.79	**0.26**	**83.45**	**0.54**

Experiment Results

Trend chart illustrating the decline in performance with increasing perturbations.



Experiment Results

Ablation experiment result. CoIPO outperforms all other methods, demonstrating the effectiveness of the approach and the necessity of each of its components.

Model	Method	Clean		TextFolder		DeepWordBug		CheckList		StressTest		Avg	
		Acc	Diff	Acc	Diff	Acc	Diff	Acc	Diff	Acc	Diff	Acc	Diff
Llama	SFT	57.28	/	53.87	**3.41**	54.11	**3.17**	54.67	2.61	53.68	3.60	54.72	**3.20**
	CL	61.87	/	56.11	5.76	54.83	7.04	60.38	1.48	59.82	**2.05**	58.60	4.08
	InvDPO	65.89	/	59.84	6.05	**61.14**	4.75	64.89	**1.00**	61.82	4.07	62.72	3.97
	CoIPO	**67.00**	/	**62.22**	4.78	60.99	6.01	**65.39**	1.61	**63.89**	3.11	**63.90**	3.88
Qwen	SFT	77.94	/	74.99	2.95	78.03	-0.09	76.76	1.18	76.53	1.41	76.85	1.36
	CL	82.93	/	80.03	2.90	81.75	1.18	81.77	1.15	80.91	2.02	81.48	1.81
	InvDPO	83.31	/	**82.41**	**0.90**	83.50	**-0.18**	82.71	0.60	82.36	0.95	82.86	0.56
	CoIPO	**83.88**	/	82.27	1.62	**83.92**	-0.03	**83.97**	**-0.09**	**83.21**	**0.68**	**83.45**	**0.54**

Thank You!

Project repo: https://github.com/vegetable-yx/CoIPO

Xin Yang1, 3*, Letian Li2, Abudukelimu Wuerkaixi2, 3*, Xuxin Cheng3, Cao Liu3,
Ke Zeng3, Xunliang Cai3, Wenyuan Jiang4

1 Zhejiang University
2 Tsinghua University
3 Meituan LongCat Interaction Team
4 ETH Zürich