Contrastive Instruction Tuning


Tianyi Lorena Yan[1]


Fei Wang[1]


James Y. Huang[1]

Wenxuan Zhou[1]


Fan Yin[2]


Aram Galstyan[1]


Wenpeng Yin[3]


Muhao Chen[4]

1 USC University of Southern California

2 UCLA

3 PennState

4 UCDAVIS UNIVERSITY OF CALIFORNIA

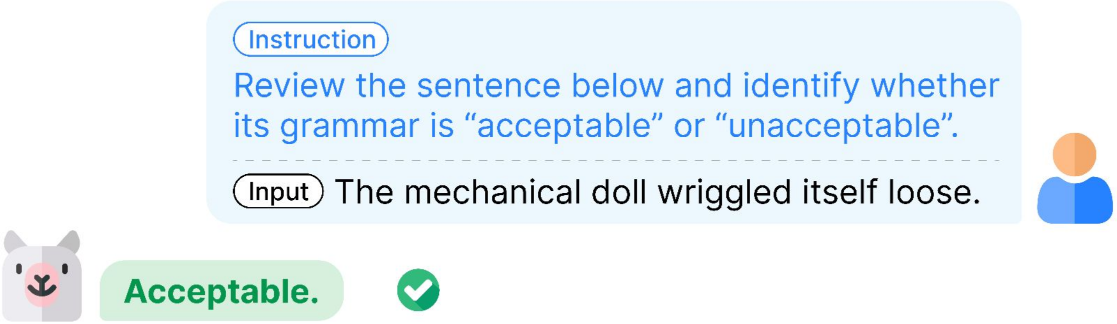
Are LLMs good human instruction followers yet?

Are LLMs good human instruction followers yet?

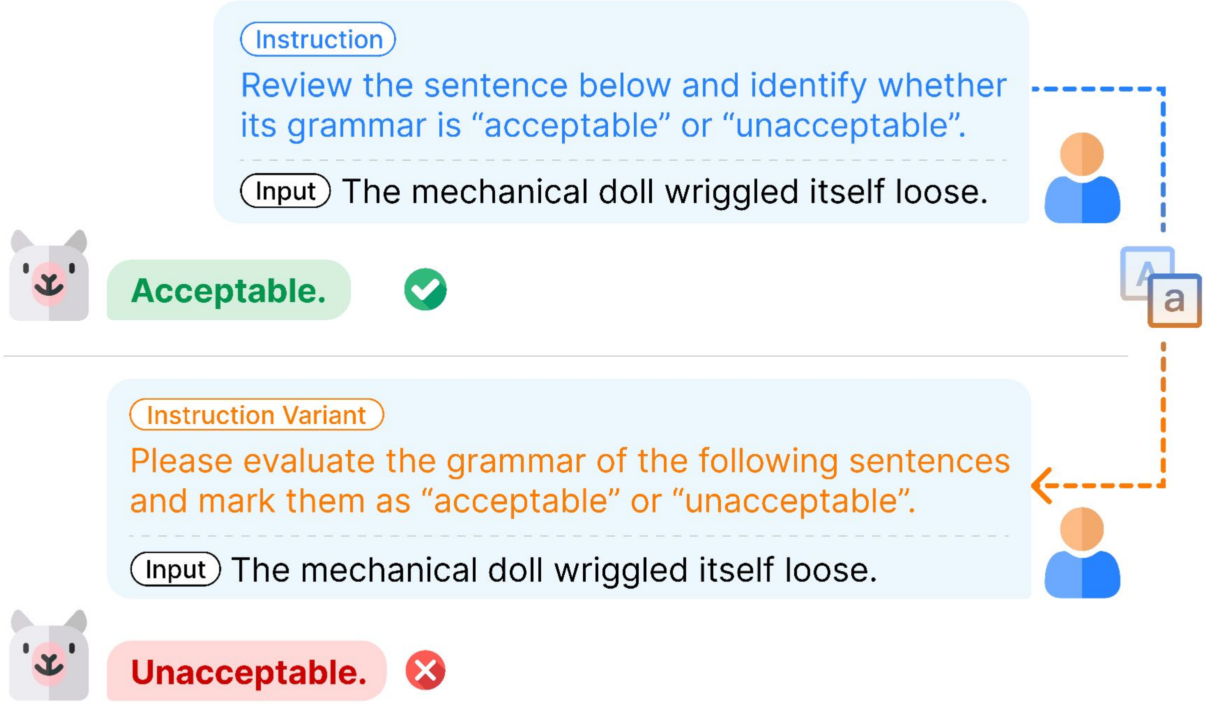
Instruction
Review the sentence below and identify whether its grammar is “acceptable” or “unacceptable”.

Input The mechanical doll wriggled itself loose.

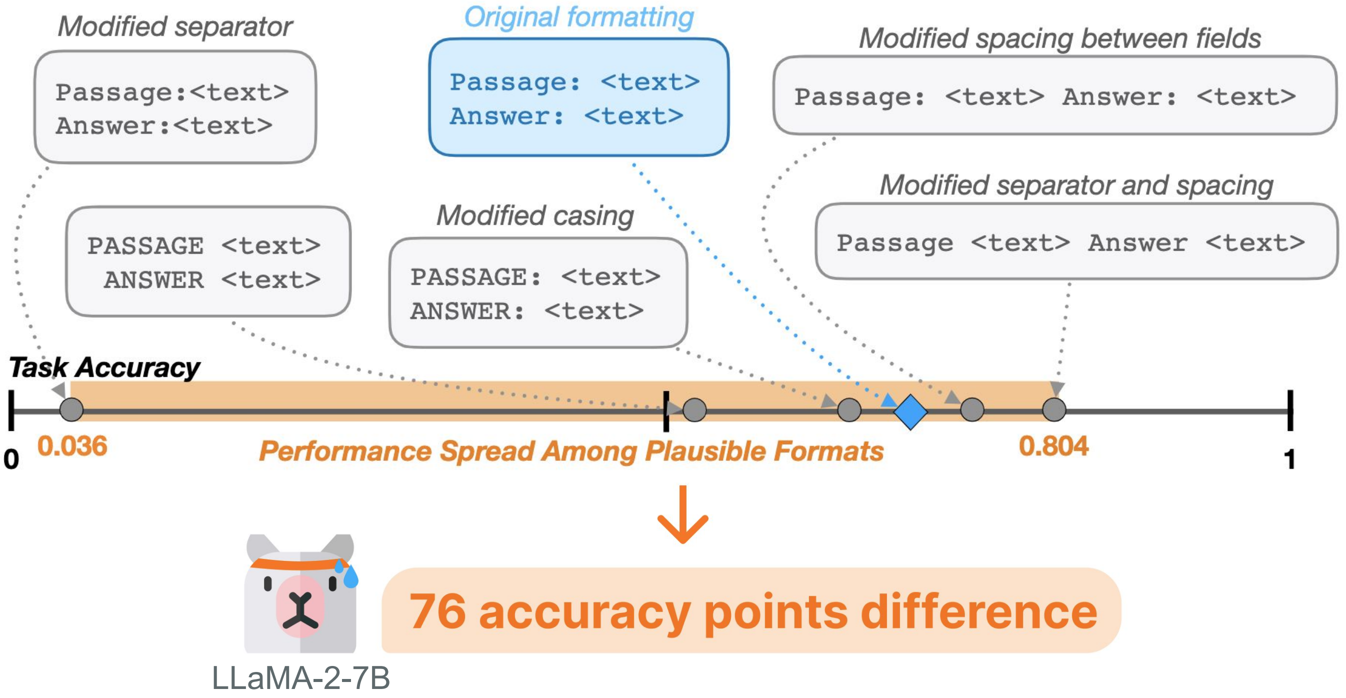
Are LLMs good human instruction followers yet?



Are LLMs good human instruction followers yet?

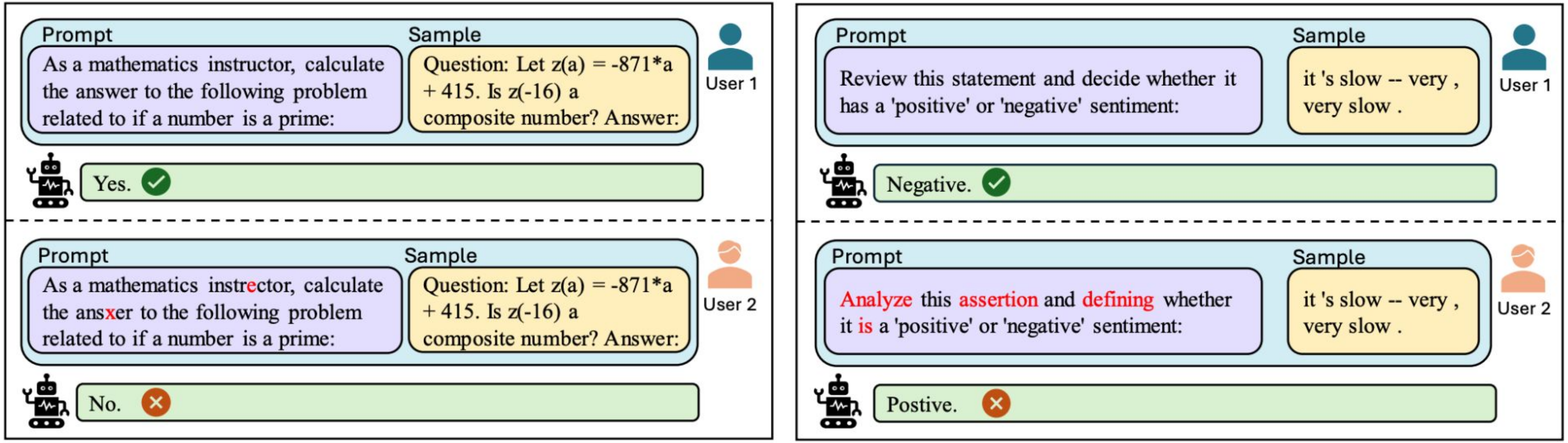


Issue: LLMs are sensitive to variations in instructions



Sclar, M., Choi, Y., Tsvetkov, Y., & Suhr, A. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. ICLR 2024

Issue: LLMs are sensitive to variations in instructions



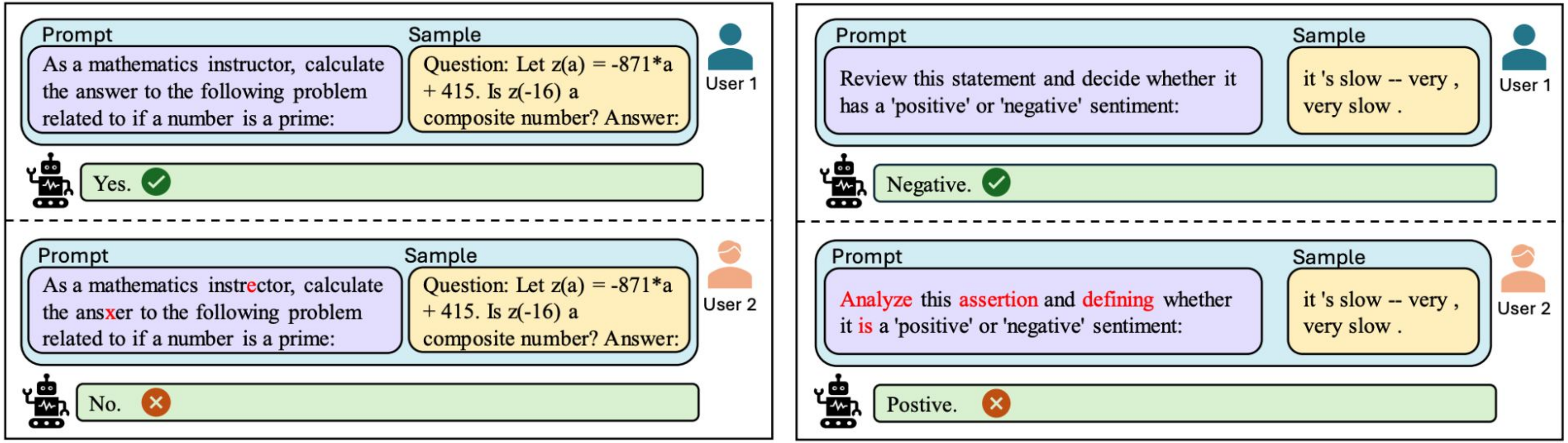
(a) Typos lead to errors in math problems.

(b) Synonyms lead to errors in sentiment analysis problems.

ChatGPT gives inconsistent answers when facing variations in instructions.

Zhu, K., Wang, J., Zhou, J., Wang, Z., Chen, H., Wang, Y., ... & Xie, X. (2023). Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. arXiv preprint arXiv:2306.04528.

Issue: LLMs are sensitive to variations in instructions



(a) Typos lead to errors in math problems.

(b) Synonyms lead to errors in sentiment analysis problems.

ChatGPT gives inconsistent answers when facing variations in instructions.

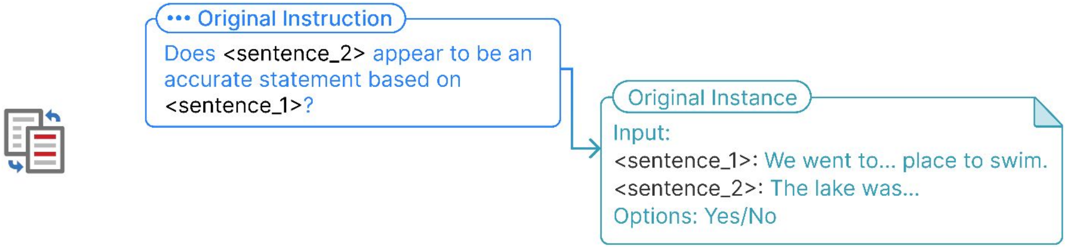
Zhu, K., Wang, J., Zhou, J., Wang, Z., Chen, H., Wang, Y., ... & Xie, X. (2023). Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. arXiv preprint arXiv:2306.04528.

Our solution: Contrastive Instruction Tuning

Idea: Encourage **semantically equivalent** inputs to **stay close to each other** while **dissimilar** ones to **be far apart** in LLMs' hidden representation space

Our solution: Contrastive Instruction Tuning (CoIN)

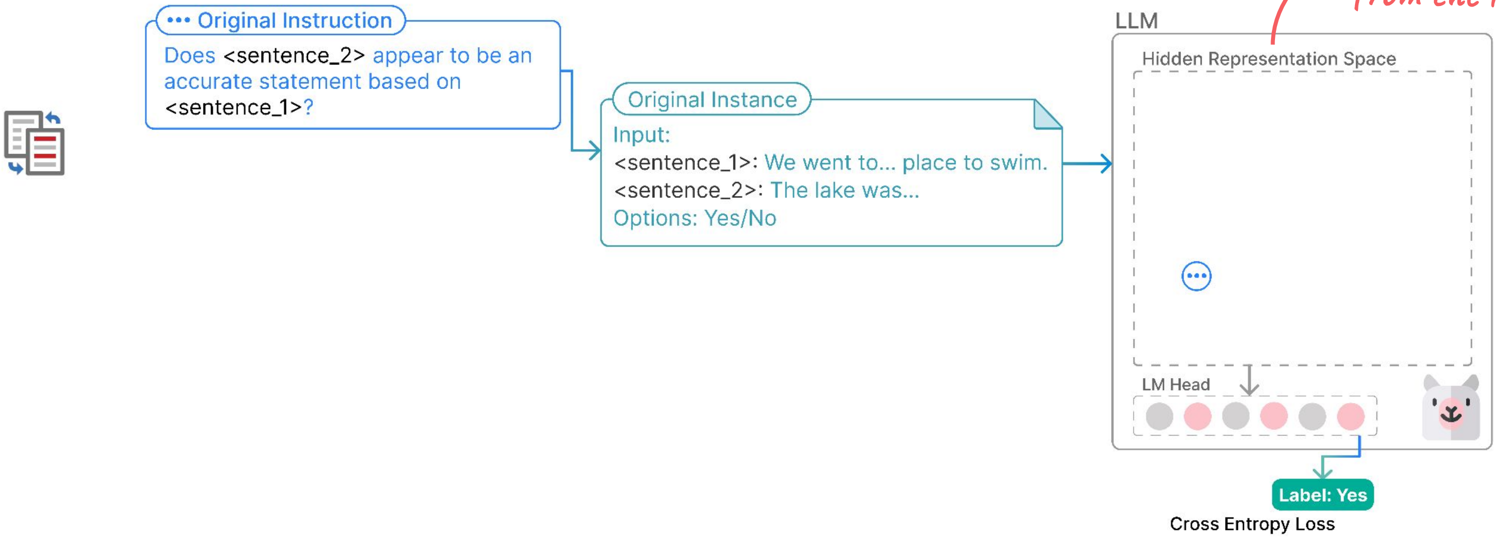
Idea: Encourage **semantically equivalent** inputs to **stay close to each other** while **dissimilar** ones to **be far apart** in LLMs' hidden representation space



Our solution: Contrastive Instruction Tuning

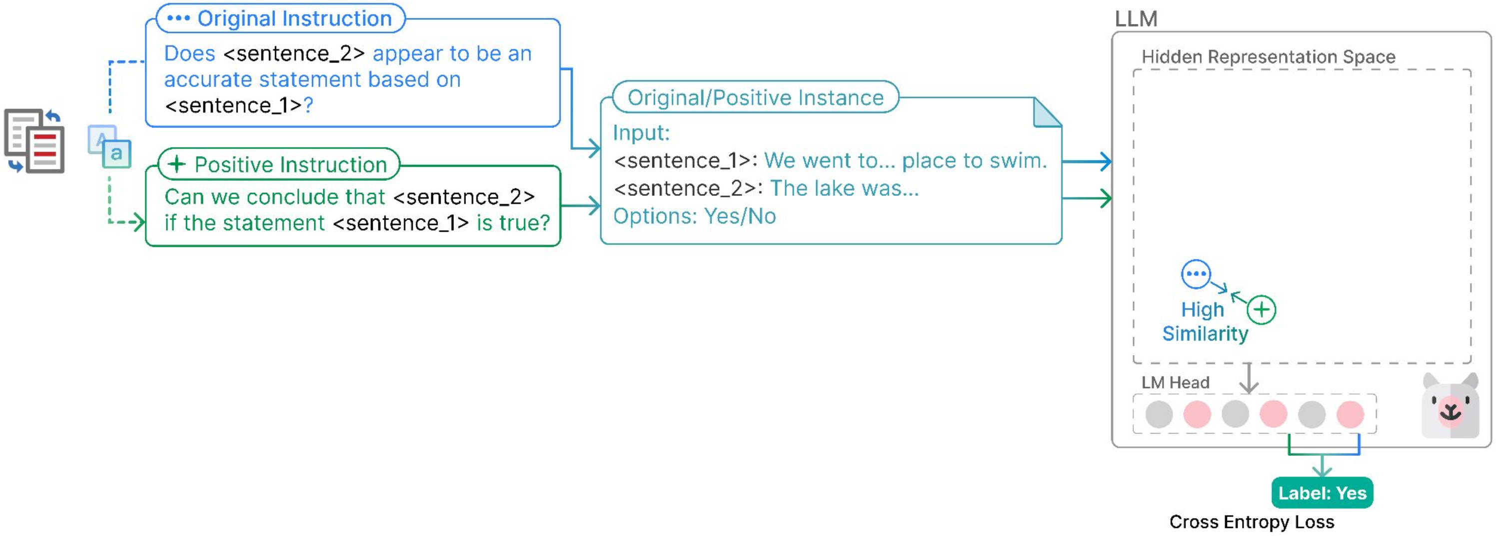
Idea: Encourage **semantically equivalent** inputs to **stay close to each other** while **dissimilar** ones to **be far apart** in LLMs' hidden representation space

Last token's hidden state from the last layer



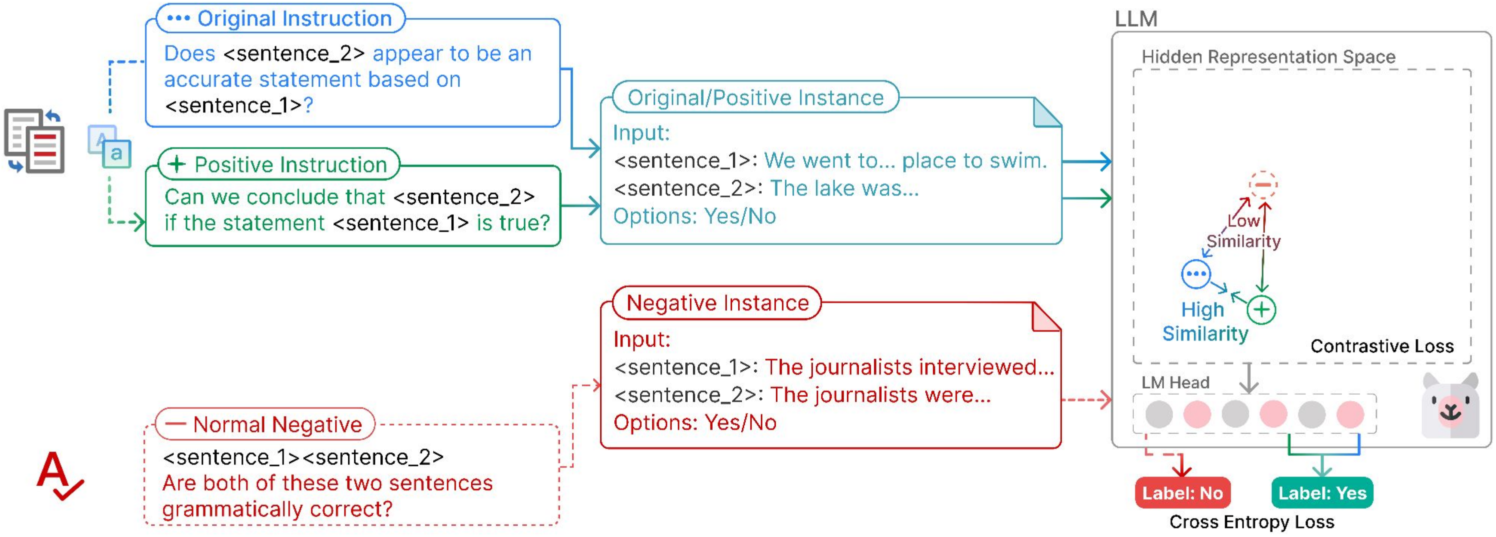
Our solution: Contrastive Instruction Tuning

Idea: Encourage **semantically equivalent** inputs to **stay close to each other** while **dissimilar** ones to **be far apart** in LLMs' hidden representation space



Our solution: Contrastive Instruction Tuning

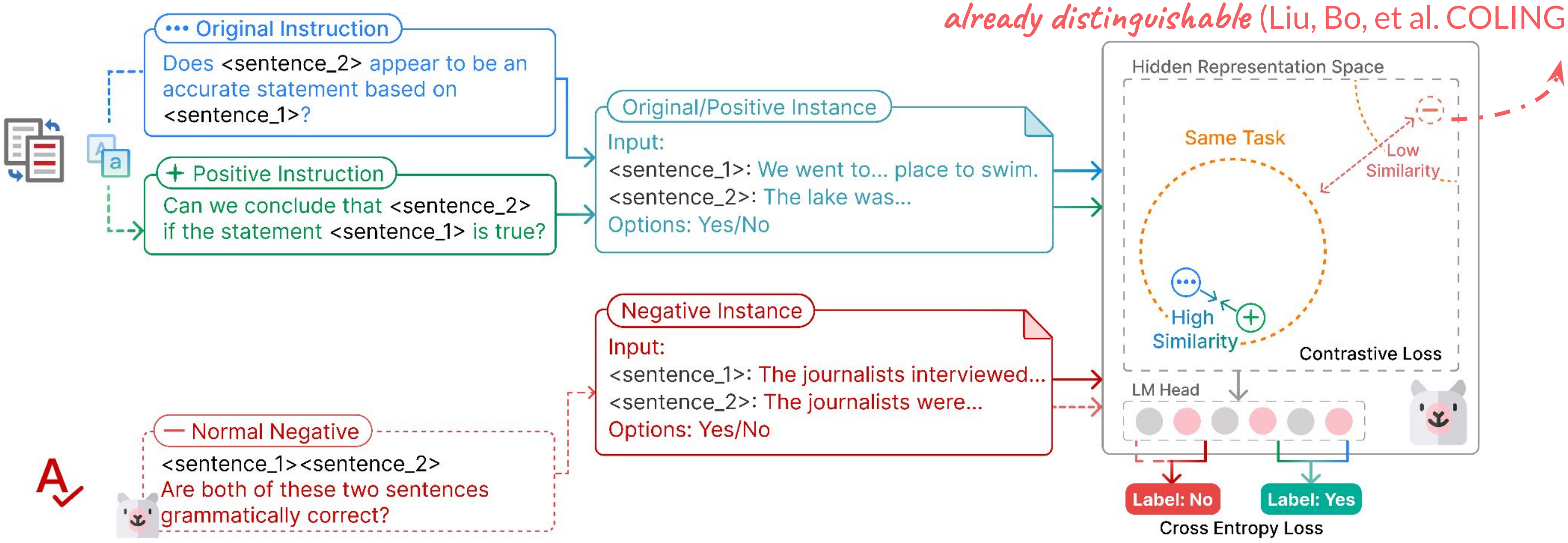
Idea: Encourage **semantically equivalent** inputs to **stay close to each other** while **dissimilar** ones to **be far apart** in LLMs' hidden representation space



Our solution: Contrastive Instruction Tuning

Idea: Encourage **semantically equivalent** inputs to **stay close to each other** while **dissimilar** ones to **be far apart** in LLMs’ hidden representation space

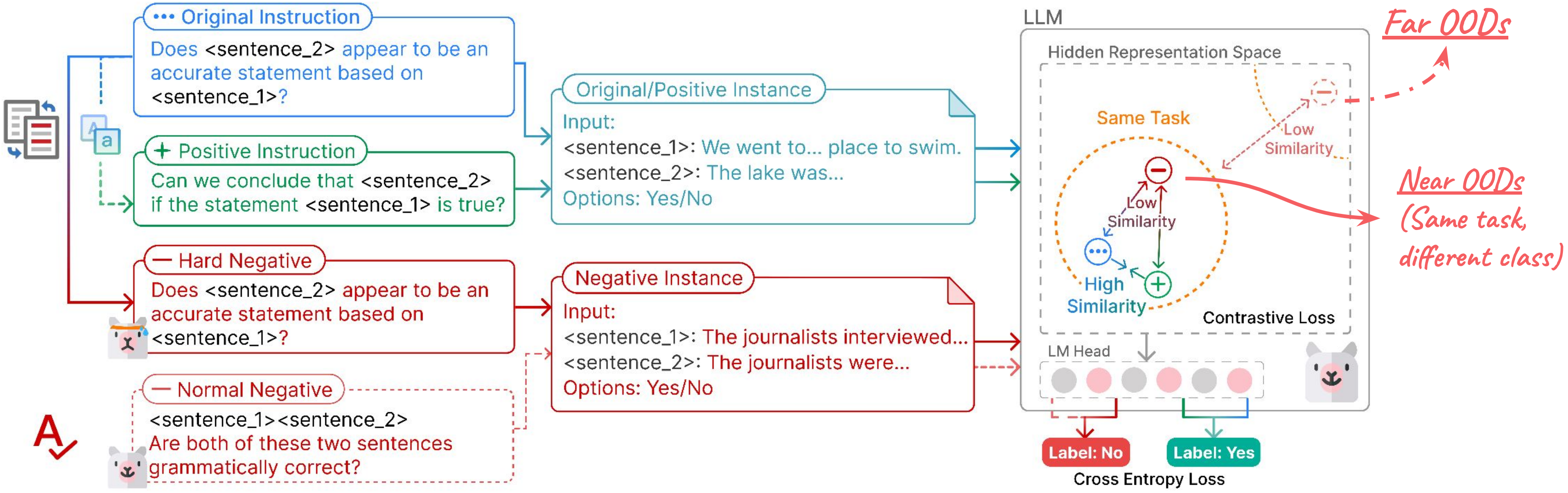
Data w/ instructions of different tasks (Far OODs) are already distinguishable (Liu, Bo, et al. COLING 2024)



Liu, B., Zhan, L., Lu, Z., Feng, Y., Xue, L., & Wu, X. M. How Good Are Large Language Models at Out-of-Distribution Detection?. COLING 2024

Our solution: Contrastive Instruction Tuning

Idea: Encourage **semantically equivalent** inputs to **stay close to each other** while **dissimilar** ones to **be far apart** in LLMs' hidden representation space



Experiment Setup: Training

- Data:
  - Datasets from the FLAN collection (52k instruction-instance pairs)
  - For every pair from a dataset
    - Positive sample: randomly select a predefined instruction template as paraphrases (Avoid making assumptions about specific types of variations in instructions)
    - Negative sample: randomly select another pair from the remaining dataset

- Model: Alpaca LoRA

(Refer to paper for more experiment details)

Experiment Setup: Evaluation

- Sample 300 instruction-instance pairs from each of the 10 GLUE tasks
- Select six clean instructions predefined for each task & add perturbations at four levels following PromptBench

Zhu, K., Wang, J., Zhou, J., Wang, Z., Chen, H., Wang, Y., ... & Xie, X. (2023). Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. arXiv preprint arXiv:2306.04528.

Experiment Setup: Evaluation

- Sample 300 instruction-instance pairs from each of the 10 GLUE tasks
- Select six clean instructions predefined for each task & add perturbations at four levels following PromptBench

Clean: Review the sentence below and identify whether its grammar is 'Acceptable' or 'Unacceptable':

Character: Reiew the seVntence below and identifpy wheoher its gVammar is 'Acceptable' or 'Unacceptable':

Word: Analyzed the assertion below and ascertain whether its grammar is 'Acceptable' or 'Unacceptable':

Sentence: Review the sentence below and identify whether its grammar is 'Acceptable' or 'Unacceptable' LGOZMPXsPd:

Semantic: Evaluate the sentence below and determine if its grammar is 'Acceptable' or 'Unacceptable':

* All instructions are unseen during training

Zhu, K., Wang, J., Zhou, J., Wang, Z., Chen, H., Wang, Y., ... & Xie, X. (2023). Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. arXiv preprint arXiv:2306.04528.

Experiment Setup: Evaluation

- Sample 300 instruction-instance pairs from each of the 10 GLUE tasks
- Select six clean instructions predefined for each task & add perturbations at four levels following PromptBench

Clean: Review the sentence below and identify whether its grammar is 'Acceptable' or 'Unacceptable':

Character: Reiew the seVntence below and identifpy wheoher its gVammar is 'Acceptable' or 'Unacceptable':

Word: Analyzed the assertion below and ascertain whether its grammar is 'Acceptable' or 'Unacceptable':

Sentence: Review the sentence below and identify whether its grammar is 'Acceptable' or 'Unacceptable' LGOZMPXsPd:

Semantic: Evaluate the sentence below and determine if its grammar is 'Acceptable' or 'Unacceptable':

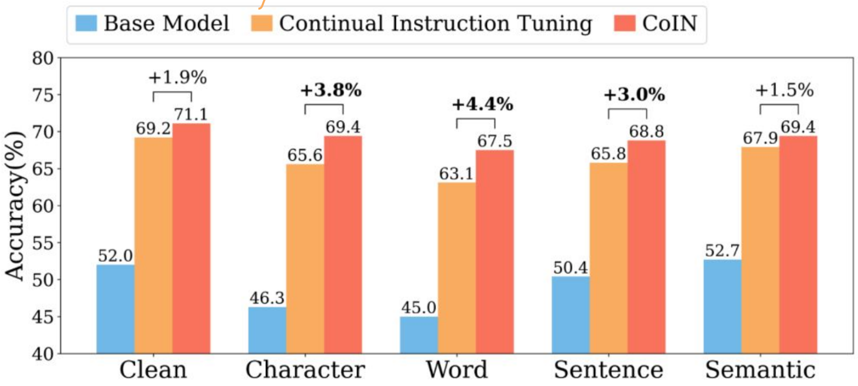
* All instructions are unseen during training

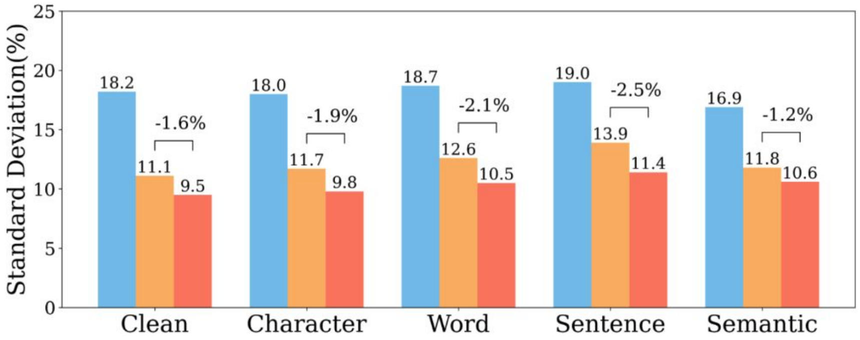
- Metric: Average accuracy (exact match) and standard deviation

Zhu, K., Wang, J., Zhou, J., Wang, Z., Chen, H., Wang, Y., ... & Xie, X. (2023). Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. arXiv preprint arXiv:2306.04528.

Main Results

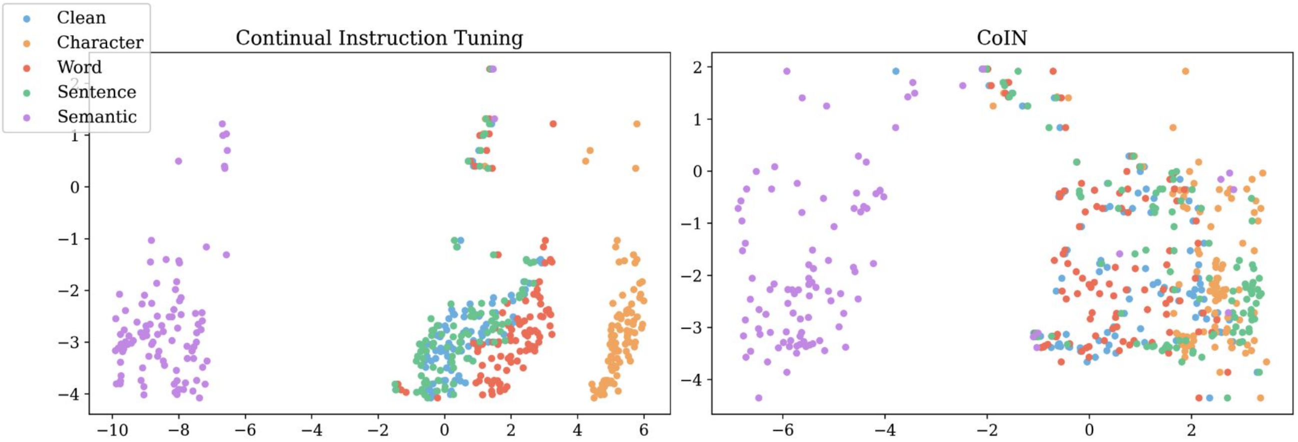
Fine-tuned on the same data as CoIN w/o contrastive loss





- 🤗 Consistent improvement in accuracy & decrease in standard deviation w/o introducing any new data & training steps
- 🤗 Able to generalize from paraphrases to all types of variations in instructions

Analyses: Closer Representations of Instruction Variations



UMAP (McInnes et al., 2020) visualization of the hidden representations of decoder's last output token (300 datapoints from CoLA dataset)

- 🤕 Continual instruction tuning:
  Instructions with different variations are clustered into distinct groups → Higher sensitivity

- 🤗 CoIN:
  Larger overlap between clean & perturbed instructions →More robust to instruction variations

Analyses: Impact on Different Tasks

(%)	Continual Instruction Tuning		COIN		△	
Task	Accuracy	Std	Accuracy	Std	Accuracy	Std
Sentiment Analysis	89.0	4.1	90.4	3.1	**+1.4**	**-1.1**
Natural Language Inference	64.4	3.7	66.1	3.5	**+1.7**	-0.2
Paraphrase Identification	63.0	11.0	68.5	5.9	**+5.4**	**-5.1**
Grammar Correctness	62.0	9.2	68.4	3.9	**+6.3**	**-5.3**

- More evident improvement in paraphrase identification and grammar correctness
- Directly benefit from model's more refined ability to group textual inputs with similar semantic meanings

Conclusions

- We propose <u>Co</u>ntrastive <u>In</u>struction Tuning (CoIN) that aligns hidden representations of semantically equivalent instruction-instance pairs

- Evaluation results on PromptBench w/ instruction variations at character, word, sentence, and semantic level demonstrate CoIN's effectiveness of enhancing LLMs' robustness to instruction variations

- CoIN can be applied to enhance models' robustness on other prompt component (e.g. system prompts, few-shot demonstration) and other modalities

QR Codes

Paper

Code

tianyiy@usc.edu / @LorenaYannnnn

Special thanks to all my amazing collaborators!