Contrastive Instruction Tuning (CoIN)

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

Correspondence to: tianyiy@usc.edu / @LorenaYannnnn

Paper

Code

Motivation

LLMs are sensitive to variations in instructions

Instruction
Review the sentence below and identify whether its grammar is "acceptable" or "unacceptable".
Input The mechanical doll wriggled itself loose.

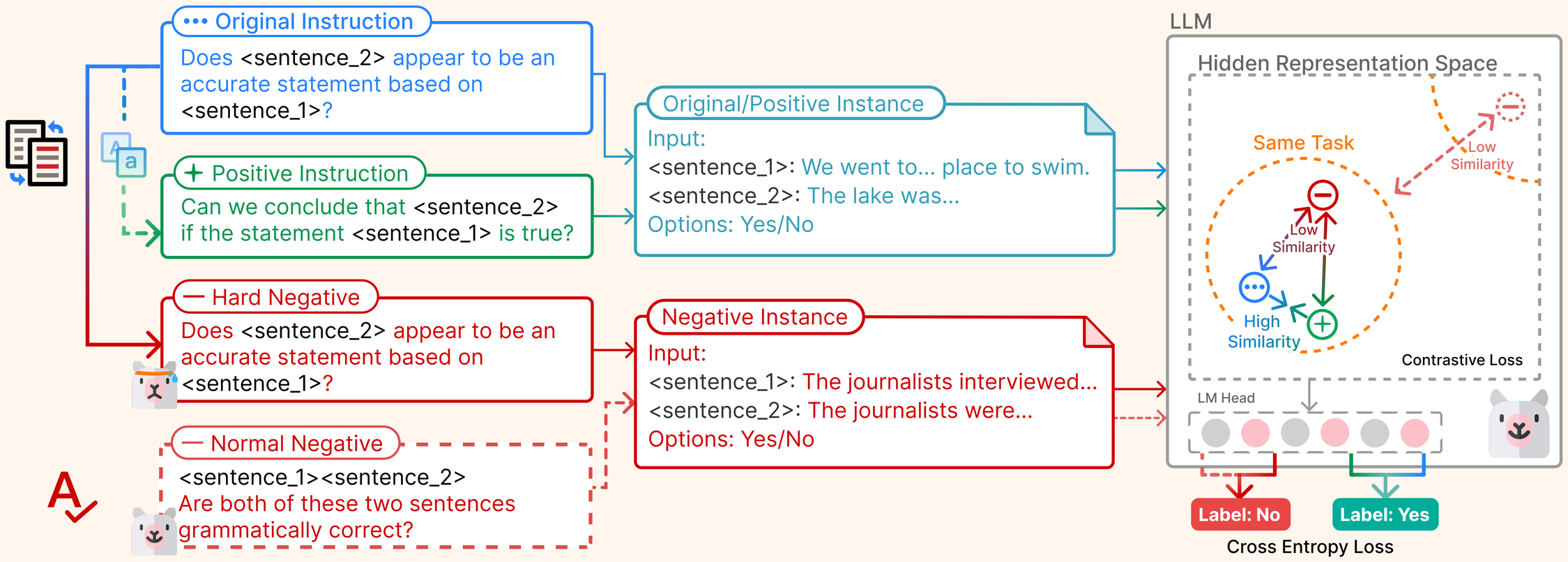
Acceptable.

Instruction Variant
Please evaluate the grammar of the following sentences and mark them as "acceptable" or "unacceptable".
Input The mechanical doll wriggled itself loose.

Unacceptable.

Method

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



Experiment Setup

Training

- 25 datasets from FLAN collection (52k instruction-instance pairs)
- \+ Positive sample: Instruction paraphrases
  (Avoid making assumptions about types of variation in instructions)
- — Negative sample: Randomly select one instance from the remaining dataset

Evaluation

PromptBench + GLUE

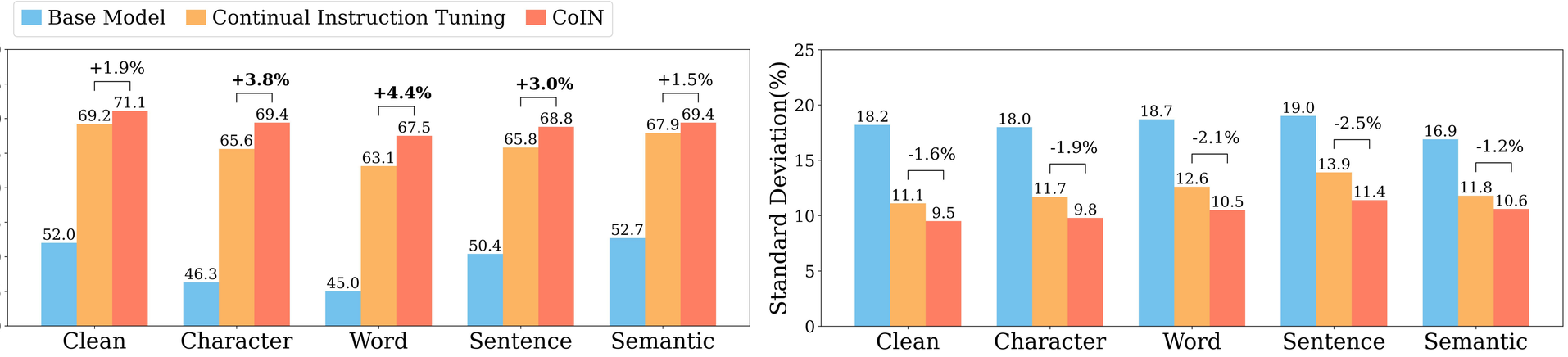
Six clean instructions + perturbed versions

```
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':
```

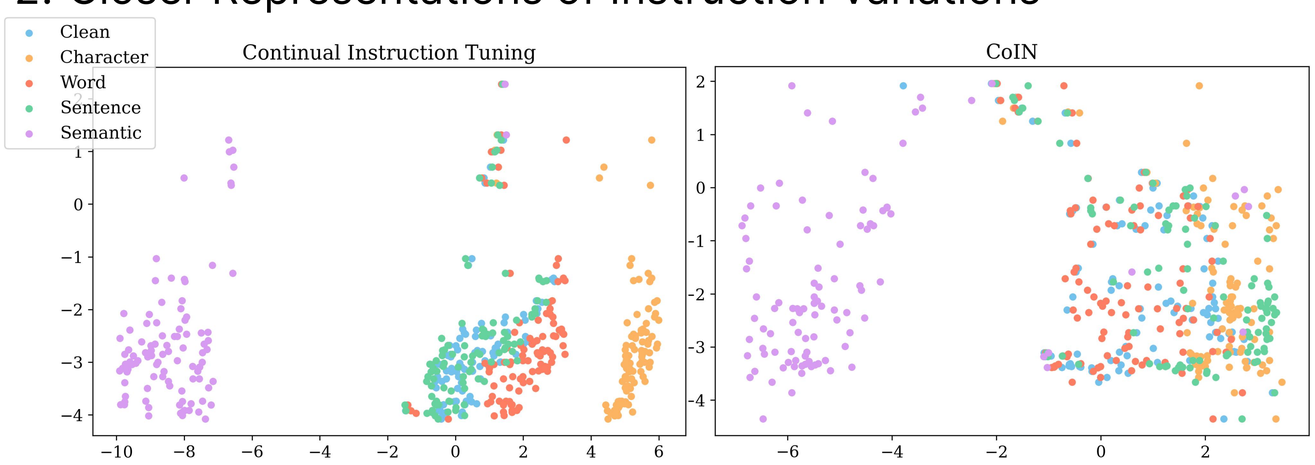
* All instructions are **unseen** during training

Results & Analyses

1. Consistent improvement in performance w/o introducing any new data and training steps
   + Generalizes from paraphrases to all four types of variations in instructions



2. Closer Representations of Instruction Variations



3. Positive Impacts 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**

Future Applications

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