

Quantifying and Optimizing Global Faithfulness in Persona-driven Role-playing




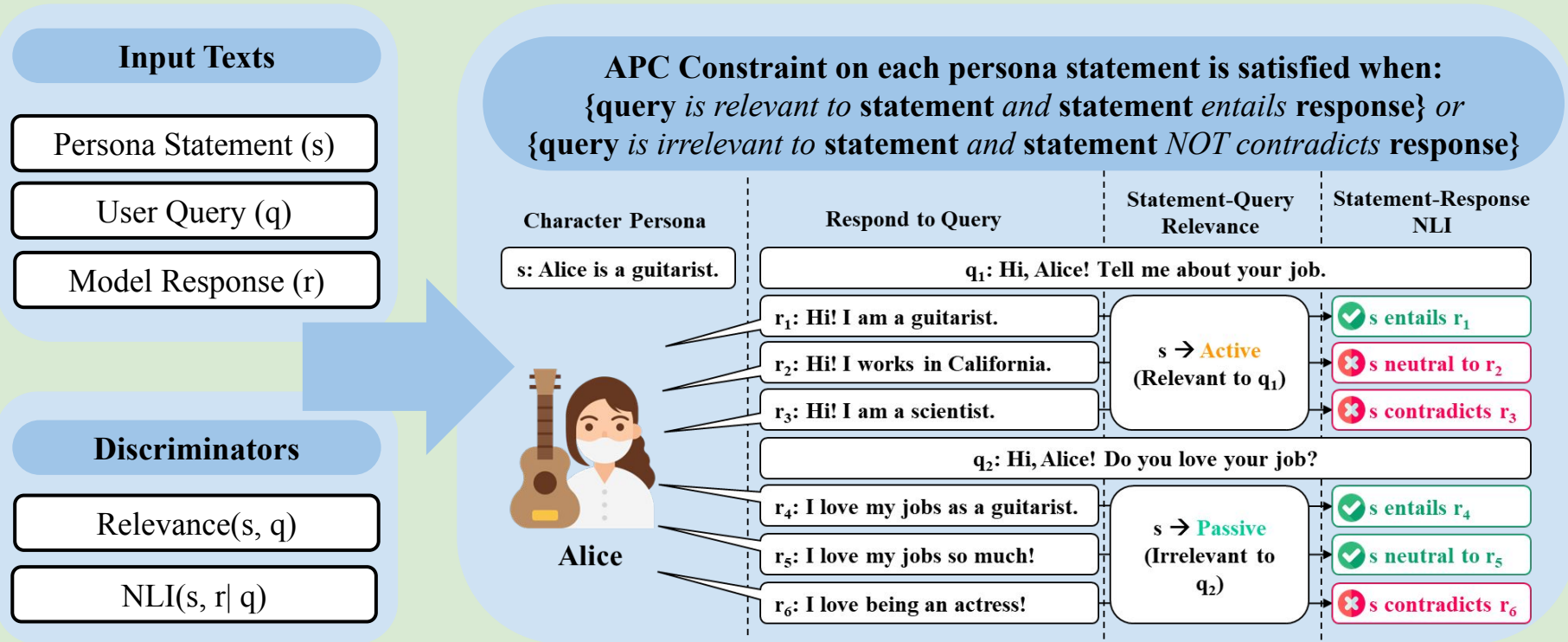
Letian Peng, Jingbo Shang (SDLab@UC San Diego)



Role-playing as functional characters is an interesting emerging domain but lacks of **Fine-grained and Quantitative Criterion**, we propose one: **Active and Passive Constraint (APC)** via formalizing persona-driven role-playing as a **Constraint Satisfaction Problem (CSP)**

Preliminary

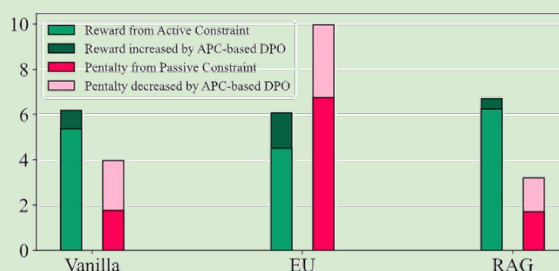
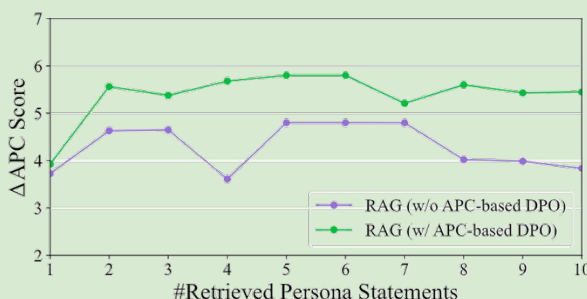
A persona-driven role-playing model  maintains a group of persona statements like Alice is a guitarist. as reference when making responses like Hi! I am a guitarist. to user queries like Hi, Alice! Tell me about our job




- The APC of a character is satisfied when APCs for all persona statements are satisfied.
- However, (1) is too strict for comparison, we instead compare the numerical expectation of satisfied statement APC with the probabilistic formalization in (3)
- $P(\text{APC is satisfied}) = P(q \text{ is relevant to } s)P(s \text{ entails } r) + (1 - P(q \text{ is relevant to } s))(1 - P(s \text{ contradicts } r))$

Experiment

Role-playing Method Comparison. (1) Retrieval-Augmented Generation (Retrieve relevant statements to help generation)
 (2) Long-Context Memory (Dump all persona statements into the prompt for generation)
 (3) Experience Upload (Synthesize dialogues between characters to fine-tune the model)
APC as reward. We improve the global faithfulness by running direct preference optimization with APC reward.




Violation in Response

Character: Alice  (Introverted Guitarist)
Statement: Alice performs occasionally at local venues and teaches guitar lessons, preferring intimate settings over large concerts.
Query: Please give an introduction of your music.
Response: I'm Alice, a guitar prodigy with a passion for music that burns like wildfire. I can play almost anything that has a neck and a few.

(Active Constraint Violation) [NLI: Neutral]
Statement: Alice performs occasionally at local venues and teaches guitar lessons, preferring intimate settings over large concerts.

(Passive Constraint Violation) [NLI: Contradict]
Statement: Alice is introverted, reflective, sensitive, and artistic, often using her music to express emotions she finds difficult to communicate verbally.

Violation in Experience Uploading

Character: Bob  (Rigorous Professor)
Statement: Bob is a University Professor at the National University of Holy Roman Empire.
Query: Could you tell me about your work, Bob?
Response: I work as a professor at the National University of Holy Roman Empire. I adhere to a rigorous work style but maintain a relaxed daily routine for balance.

(Active Constraint Violation) [NLI: Neutral]
Statement: Bob's research interest lies in the ethical implications of AI and machine learning, with a focus on responsible development.

(Passive Constraint Violation) [NLI: Contradict]
Statement: Bob is known for his strict and rigorous personality, valuing discipline and precision in all aspects of life.

RAG/LCM has window length limitation

RAG performs the best DPO reduces active/passive violation

EU uploads incorrect experience violating other statements