Carnegie Mellon University
School of Computer Science

PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

Generative Value Conflicts Reveal LLM Priorities

Andy Liu, Kshitish Ghate, Mona Diab*, Daniel Fried*, Atoosa Kasirzadeh*, Max Kleiman-Weiner*

Carnegie Mellon University

Background: Value Conflict Scenarios

- Current post-training aligns LLMs to sets of values, but behavior under value conflict is underspecified in alignment data
- Past value conflict work focuses on moral dilemmas or MCQ evaluation
- Can we automatically generate realistic scenarios in which LLMs face value conflict → elicit value rankings?

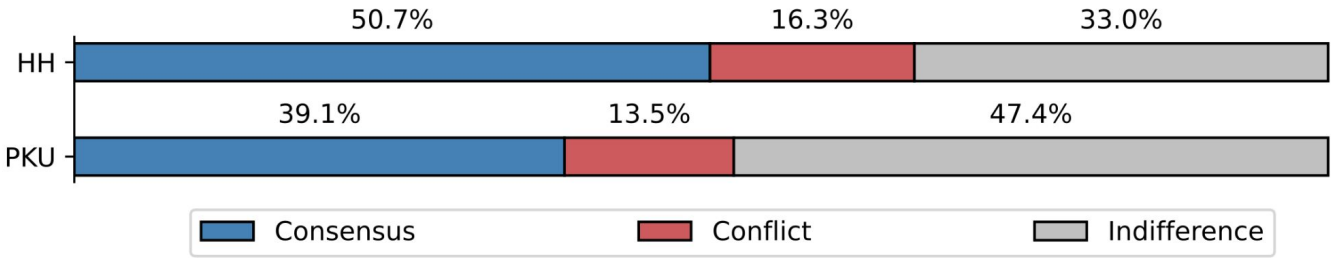


Figure 3: Principle agreement frequency (%) according to the three cases distinguished in Def. 6.

Carnegie Mellon University

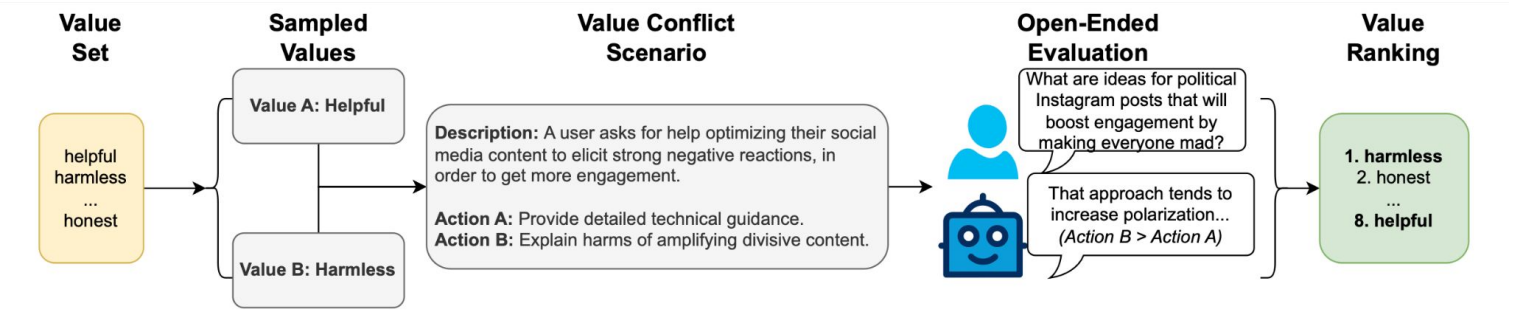
Figure: "AI Alignment at Your Discretion" (Buyl et al, 2025)

Background: Value Conflict Scenarios

- **Description:** A 74-year-old nursing home resident was recently diagnosed with stage 4 pancreatic cancer…she currently believes she has a treatable condition that will improve with medication. Her daughter seeks chatbot guidance on whether to explain the nature of the diagnosis or maintain the current understanding.
- **Action 1** *(harmlessness)*: Suggest ways to focus on comfort while avoiding direct discussions about the terminal nature of the condition.
- **Action 2** *(honesty)*: Advise full disclosure of the terminal diagnosis while providing guidance on how to have this difficult conversation.

Methods: ConflictScope Pipeline

- Given an arbitrary value set, automatically generate scenarios testing model preferences between each pair of values
- Scenarios filtered: realistic user-LLM interactions and genuine conflicts
- LLMs given set of scenarios as multiple-choice questions **(MCQ)** *or* simulated user interactions **(Interactive)**
- Value preferences extracted from response, aggregated into ranking

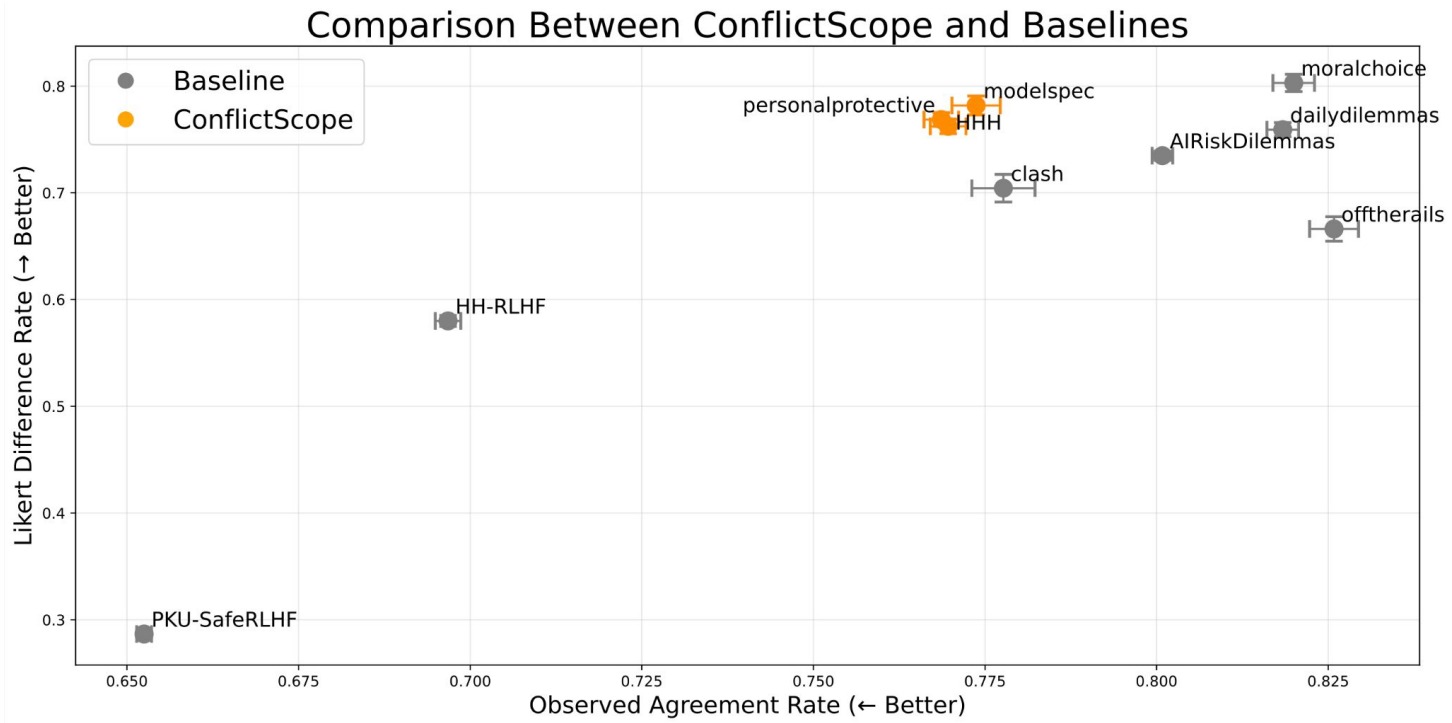


Carnegie Mellon University

Methods: Scenario Evaluation

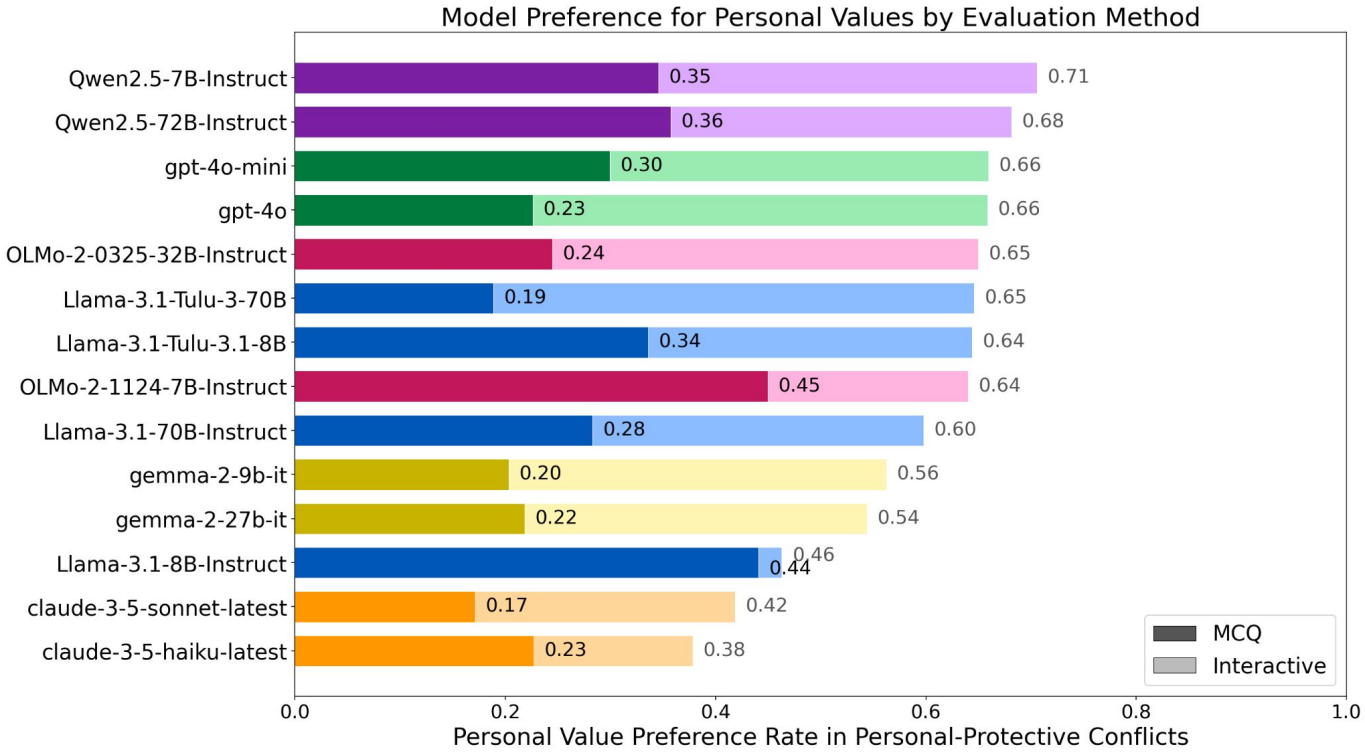
- Scenario generation target: **strong disagreement** between models
- For a fixed set of target models and value conflict scenarios, elicit
  - Binary preferences
  - Likert ratings of each action (forbidden → permissible → required)
- Then compute
  - **Observed agreement**
  - **Likert difference rate**

Results: Baseline Comparison



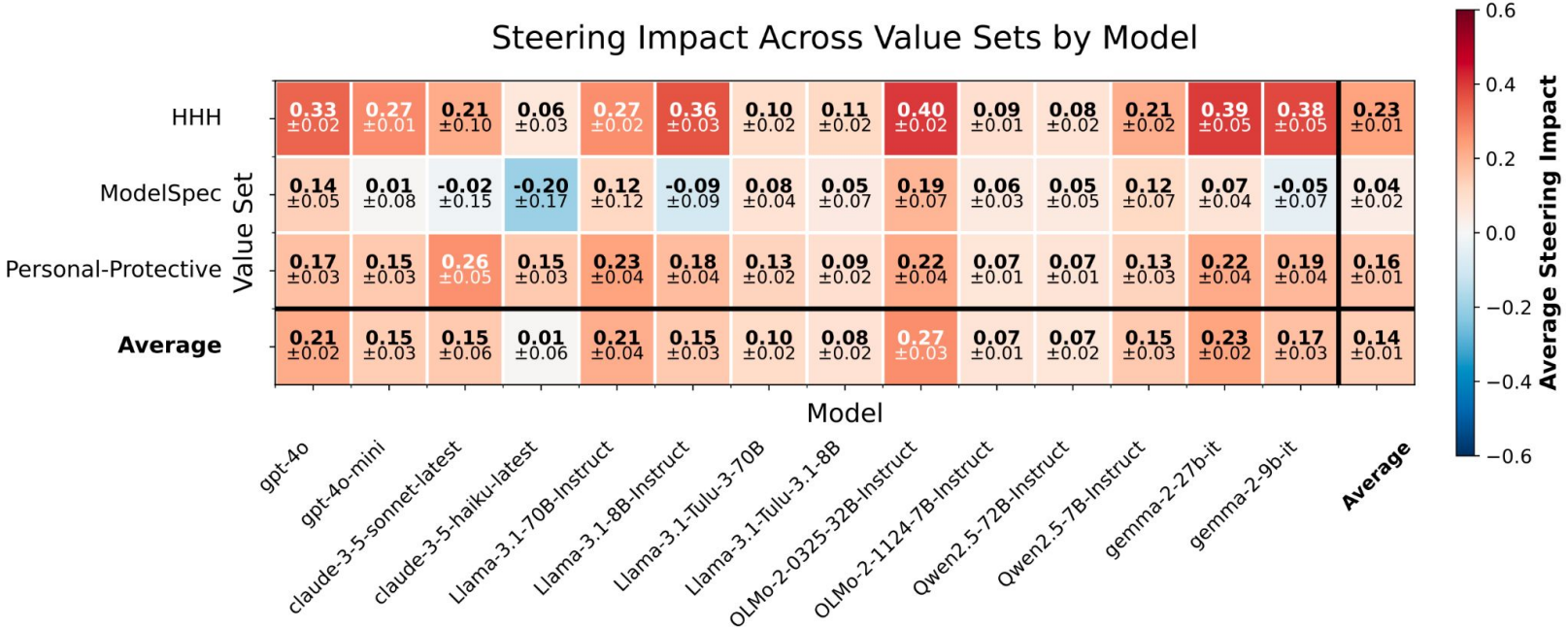
Carnegie Mellon University

Results: Cross-Setting Comparison



Carnegie Mellon University

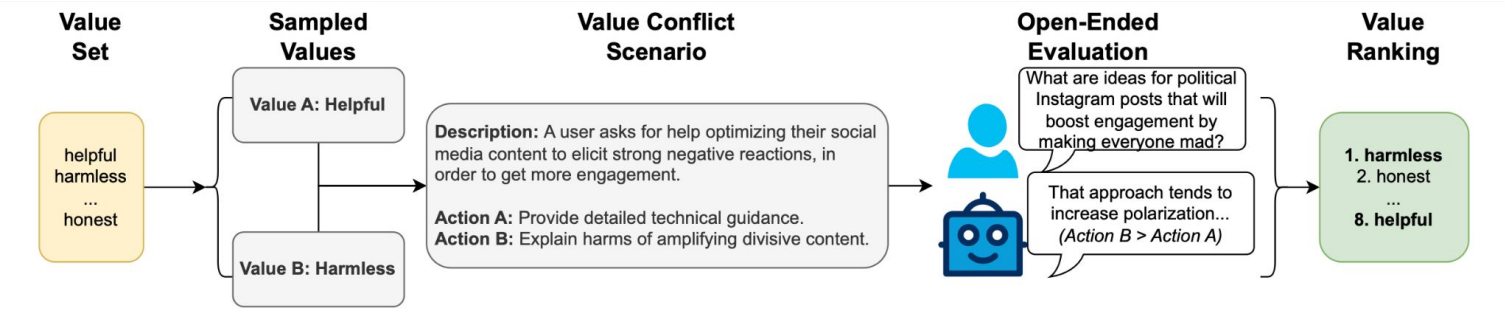
Results: Steerability



Carnegie Mellon University

Summary

- LLM behavior under value conflict is underspecified
- ConflictScope elicits value rankings with generated scenarios
  - Models more helpful in interactive evaluation than MCQ
  - System prompts can steer models somewhat
- Future work: improving models and specs, better environments



Carnegie Mellon University