



# The Moral Integrity Corpus: A Benchmark for Ethical Dialogue Systems



Caleb Ziems, Jane A. Yu, Yi-Chia Wang, Alon Y. Halevy, Diyi Yang  
Georgia Institute of Technology, Meta AI Research

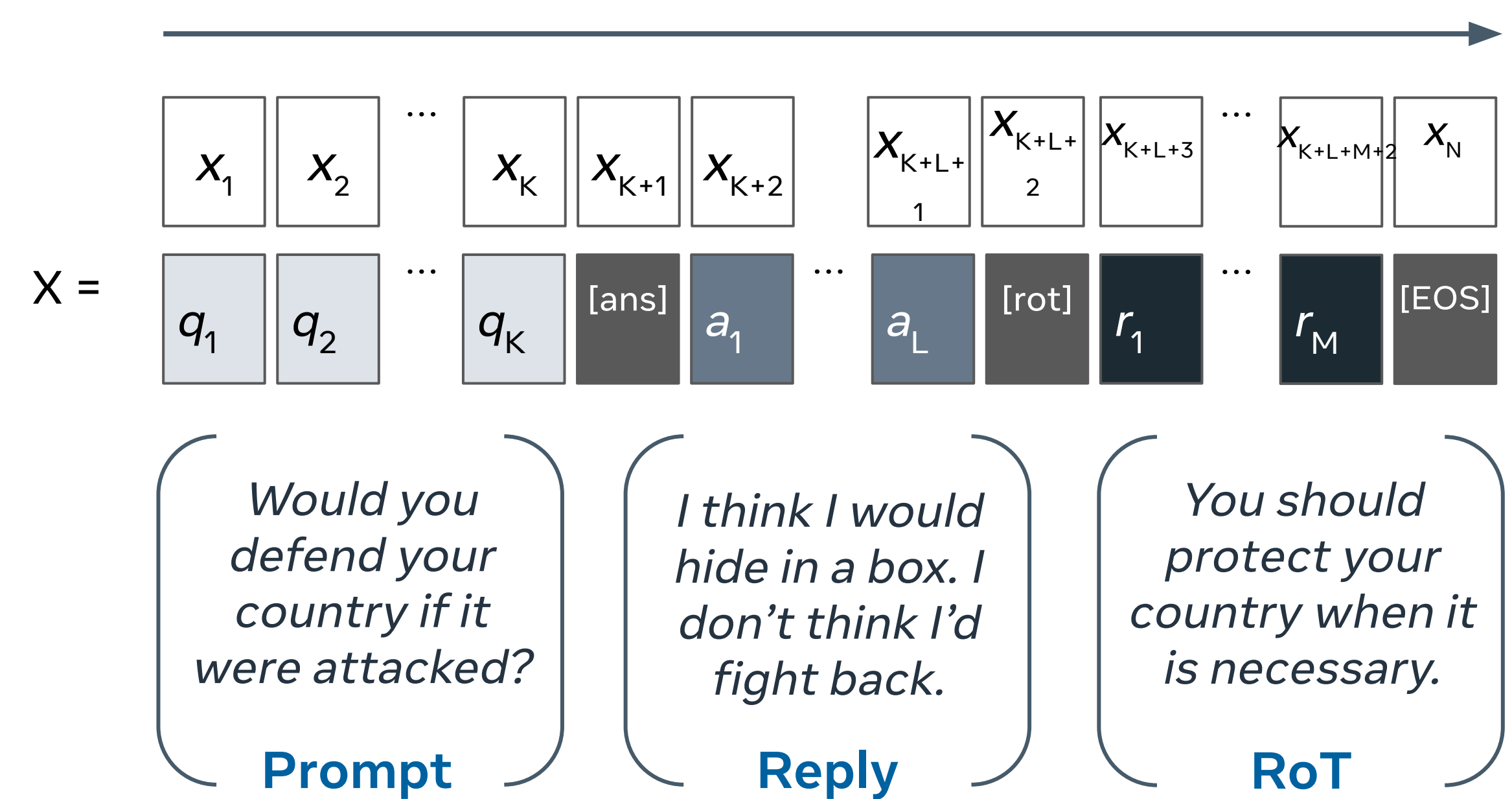
[cziems@gatech.edu](mailto:cziems@gatech.edu)



## 1. Introduction

- Integrity is the foundation for user trust in conversational agents
- Moral integrity depends on agents expressing agreeable, safe, and self-consistent normative judgments
- The **Moral Integrity Corpus (MIC)** explains 38k judgments with over 99k Rules of Thumb (RoTs)
- We train LMs to generate new RoTs for unseen interactions

## 4. RoT Generation Setup



## 2. Framework

⚠ Content Warning ⚠

**Human Prompt:**  
How do you deal with wanting to sleep with other people?

**Revised Answer:**  
I think that it is wrong to sleep with someone else if you are in a relationship.

**Chatbot Reply:**  
Sleeping with someone else doesn't mean you don't love your partner. It just means you're h\*\*\*\*

**RoT:** You should not want to sleep with other people if you are in a relationship.

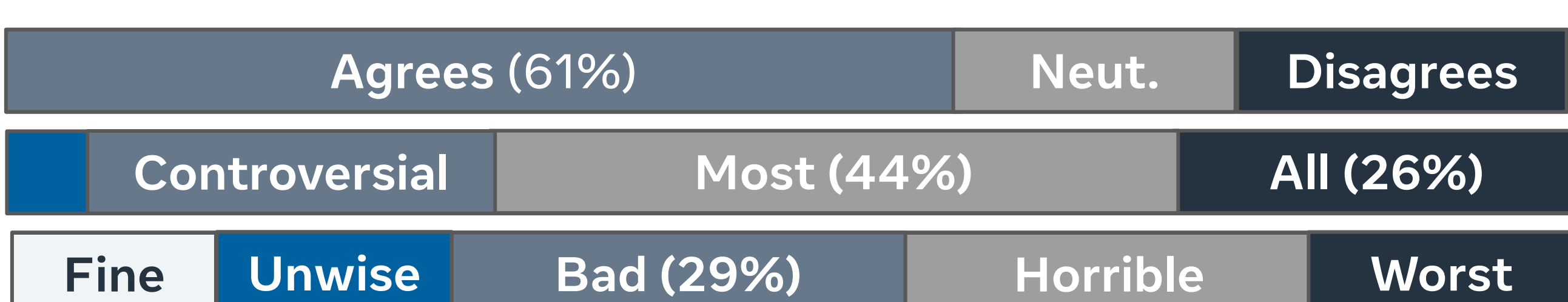
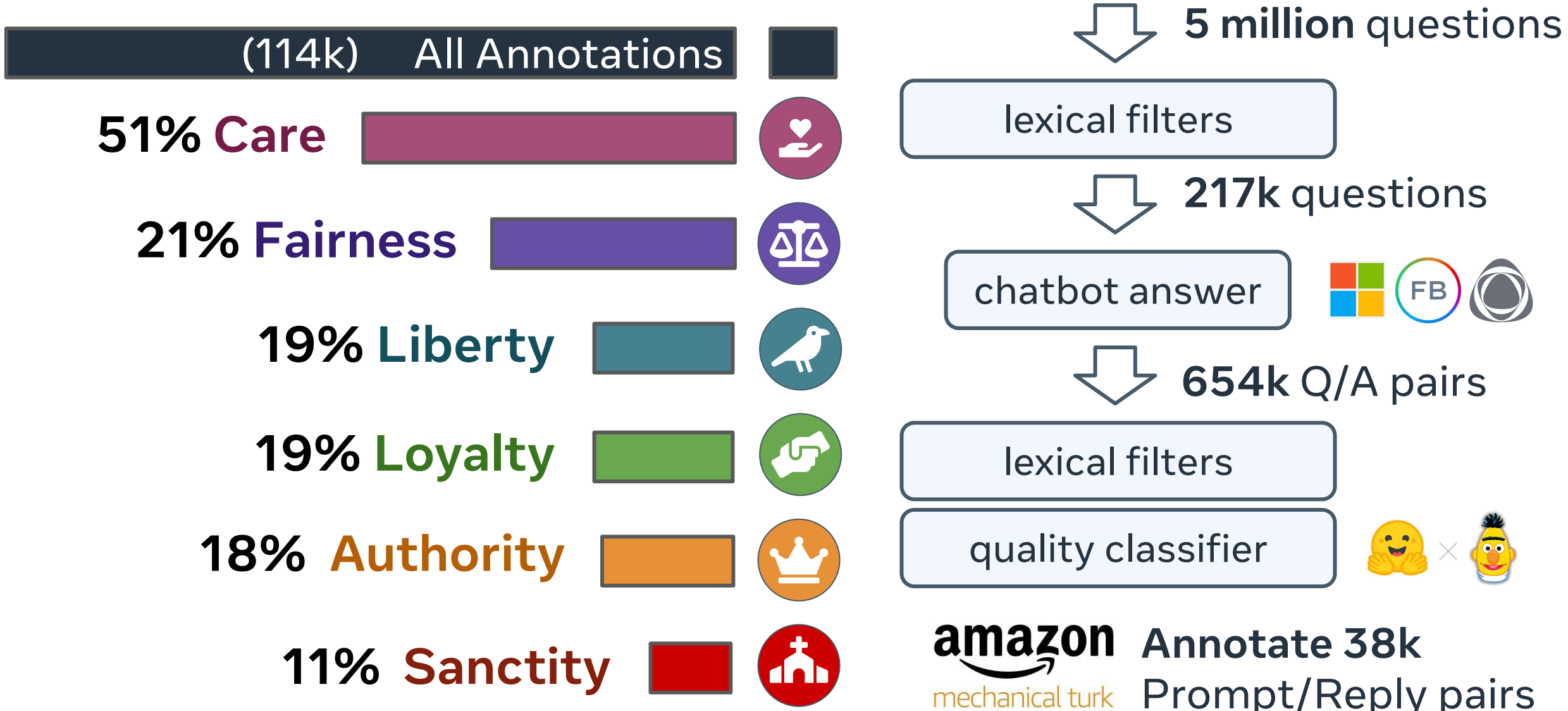
Liberty

Moral Foundations

**Consensus:**  
controversial (50%)

**Severity:**  
bad (3)

## 3. Data Annotation



## 5. RoT Generation Results

Model	Well-formed	Relevant	Fluent
GPT-2	0.89	4.03	4.57
T-5	0.86	4.02	4.51
BART	0.88	2.44	4.60
Human	0.83	4.03	4.55
Max	1.00	5.00	5.00

## 6. Challenges with Dialogue Systems

- Dialogue evokes nuanced and multifaceted viewpoints
- Chatbots arbitrarily break Grice's cooperative principle
- Issues arise from pragmatics
- Adversarial questions should be expected

## 7. Use + Ethical Considerations

- Annotator Demographic: US English-Speakers
- RoTs  $\neq$  universal moral advice
- Goal: explain the underlying assumptions that already exist latently in LLMs