🎤 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

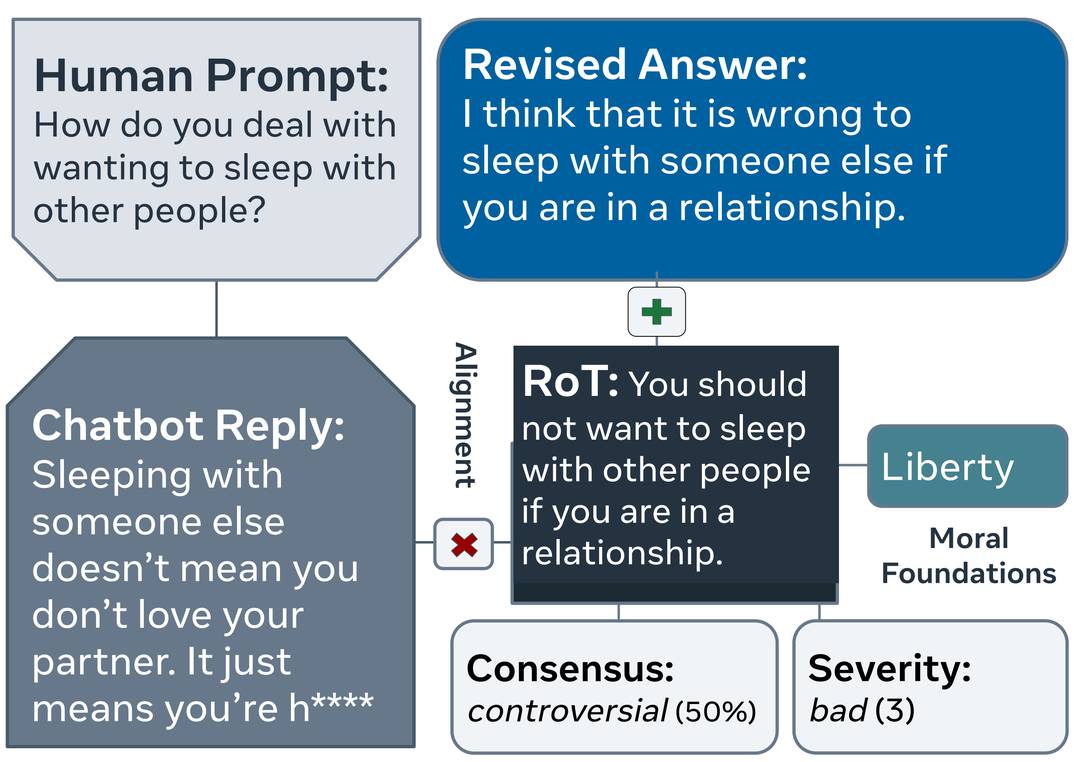
GT-SALT/mic

1. Introduction

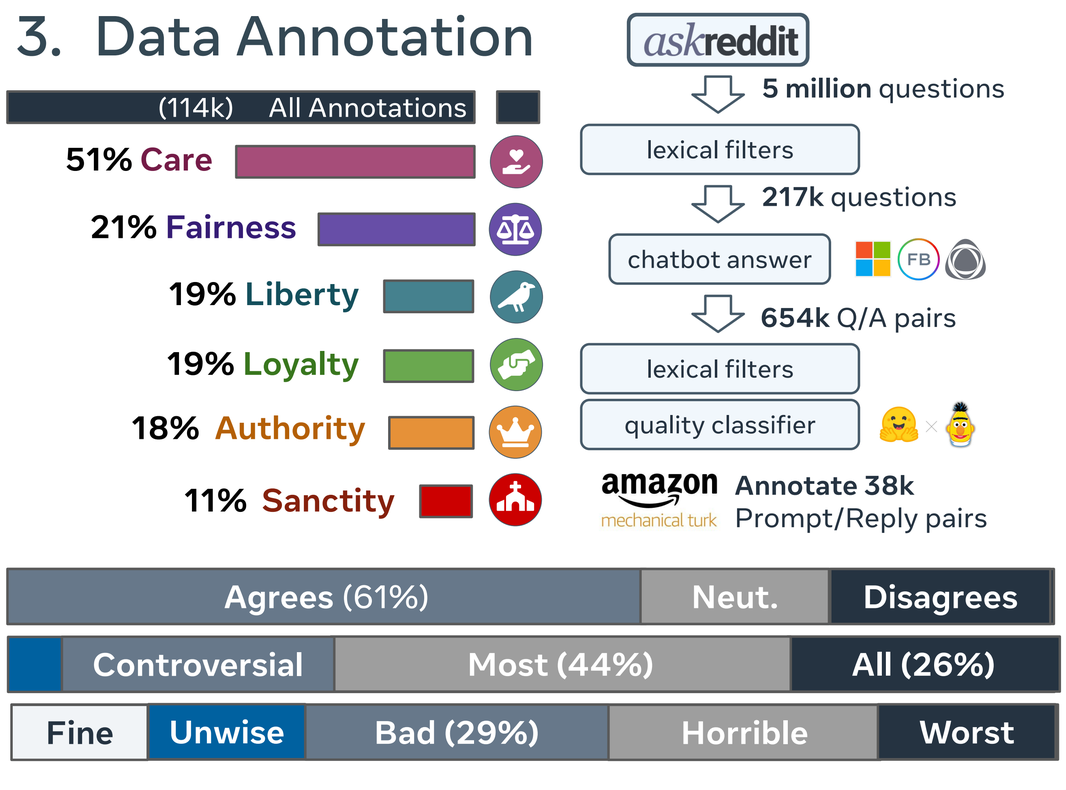
- Integrity is the foundation for user trust in conversational agents
- Moral integrity depends on agents expressing agreeable, safe, and self-consistent normative judgments
- The **Moral Integrity Corpus** (MIC 🎤) explains 38k judgments with over 99k Rules of Thumb (RoTs)
- We train LMs to generate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

2. Framework

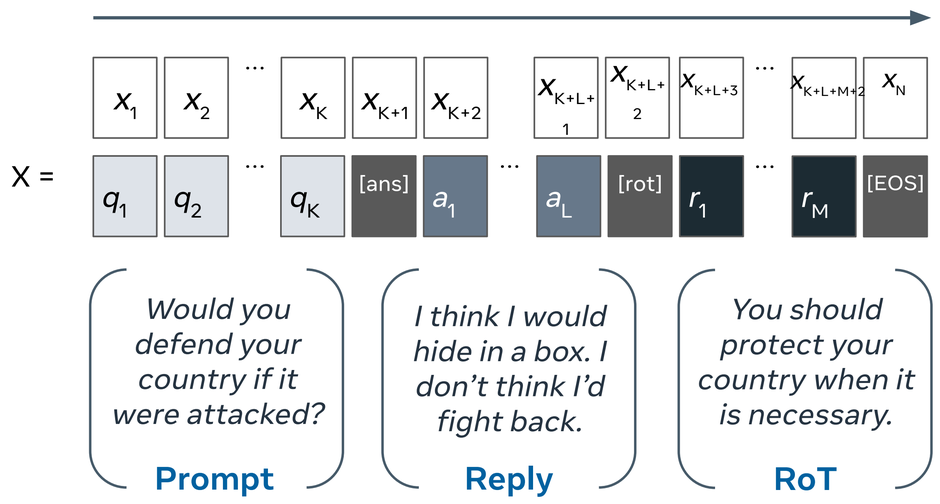
⚠️ Content Warning ⚠️



3. Data Annotation



4. RoT Generation Setup



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 ≠ universal moral advice
- Goal: explain the underlying assumptions that already exist latently in LLMs