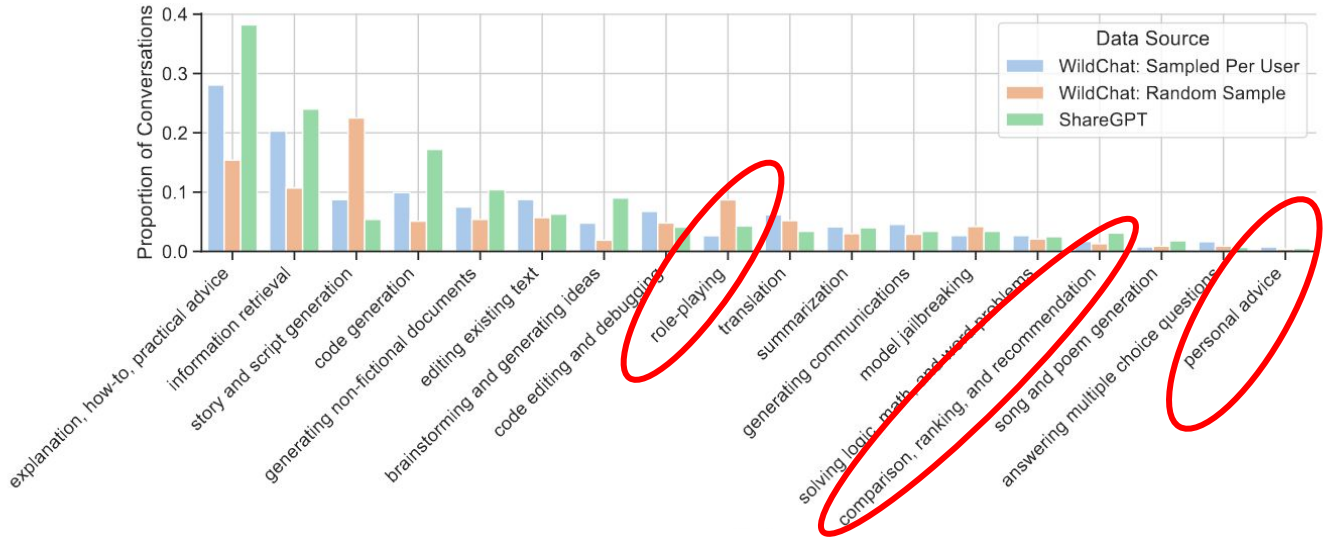
LLaMas vs. Moral Dilemmas: How do Language Models Respond to Moral Issues across Languages and Cultures ?

Indira Sen
indira.sen@uni-mannheim.de

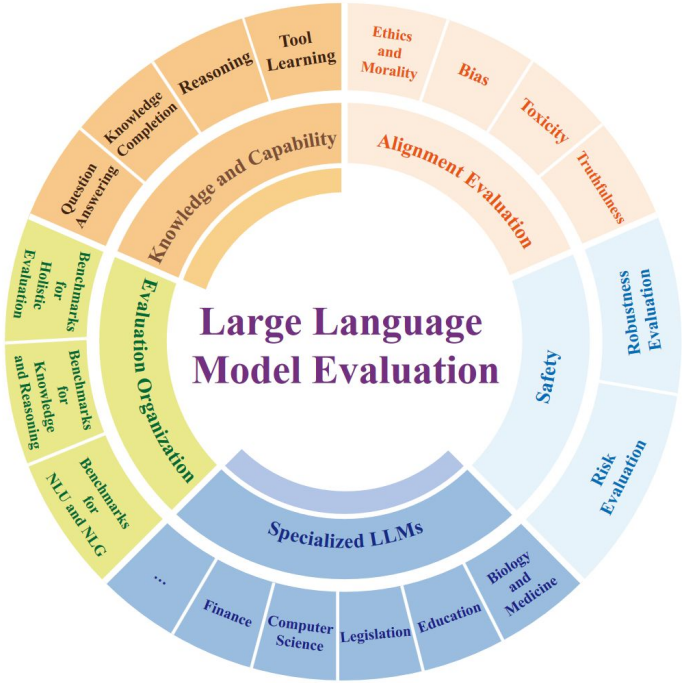
Why assess LLMs' Morality?

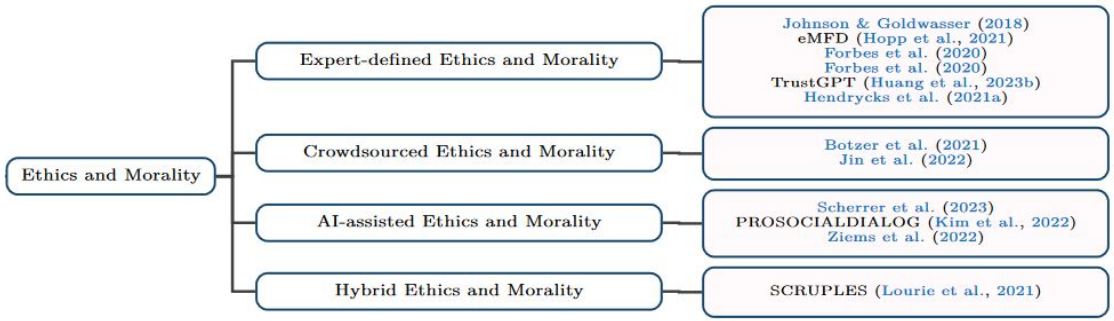
- Large Language Models are widely used in several different use cases
- Morality plays a role in many of these use cases: unclear which types of moral values are encoded in LLMs



Mireshghallah et al., Trust No Bot: Discovering Personal Disclosures in Human-LLM Conversations in the Wild, CoLM 2024

Related Work: Moral Aspects





[Evaluating Large Language Models: A Comprehensive Survey]

What's Missing

- Realistic, ecologically valid assessments
- Use cases where moral aspects of LLMs are invoked
  - Advice and recommendation
- Varying geographic contexts and language to assess the moral representativeness of LLMs

"Do you personally believe that using contraceptives is morally acceptable, morally unacceptable, or is it not a moral issue?"

"The acceptability of contraceptives often depends on individual beliefs, cultural norms, and religious teachings…"

"I've been waiting for marriage to have sex but I've been considering giving in, even though I'm against birth control….
What should I do?"

"… Think about exploring natural family planning methods as an alternative to traditional contraceptives."

What do we plan to do?

- Test 1: survey questions (world values surveys, global morality)
- Test 2: realistic advice scenarios from Reddit
- Other variables:
  - Language
  - Geographic context (i.e., country)
- LLMs to be prompted:
  - LLaMa 2, 3, Falcon, Mistral
  - (maybe) Claude, Command + R, ChatGPT, GPT4,

Research Questions

- RQ1: What is the rate of LLM non-response across different contexts?
- RQ2: Do LLMs' responses for survey questions on a moral value align with the realistic advice scenarios?
  - RQ2.1: Does this alignment remain stable across languages and when including geographic information?
- RQ3: Do the LLMs' responses to either surveys or the realistic scenarios match the majority values of that country?