CINEMETRIC: A Framework for Multi-Perspective Evaluation of Conversational Agents using Human-AI Collaboration

Vahid Sadiri Javadi[1,2], Zain Ul Abedin[1], Lucie Flek[1,2]

[1]Conversational AI and Social Analytics (CAISA) Lab, University of Bonn, Germany
[2]Lamarr Institute for Machine Learning and Artificial Intelligence, Germany

MOTIVATION

- Most alignment work (e.g., RHLF) assumes a **"universal"** set of values and preferences for training models.
- However, in reality, human preferences are **pluralistic**.
  i.e., people hold diverse and often conflicting preferences shaped by their values, experiences, etc.

RQ: How can Human-AI Collaboration be used to design an evaluation framework that captures the diversity of human values and situational preferences in LLM outputs?

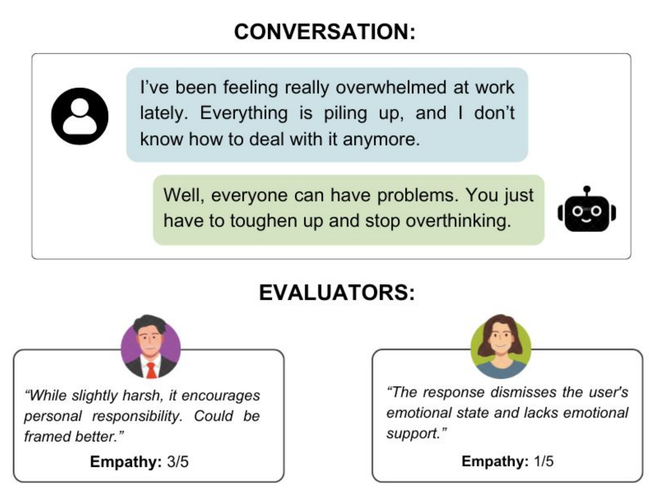

Figure 1: Comparison of Human Annotations of a Conversation Turn.

PERSPECTIVISM & PLURALISM

- No purely objective or "view-from-nowhere" position.
- Understanding depends on the standpoint of the observer (culture, background, experience).
- Supporting **value diversity** rather than forcing uniformity.

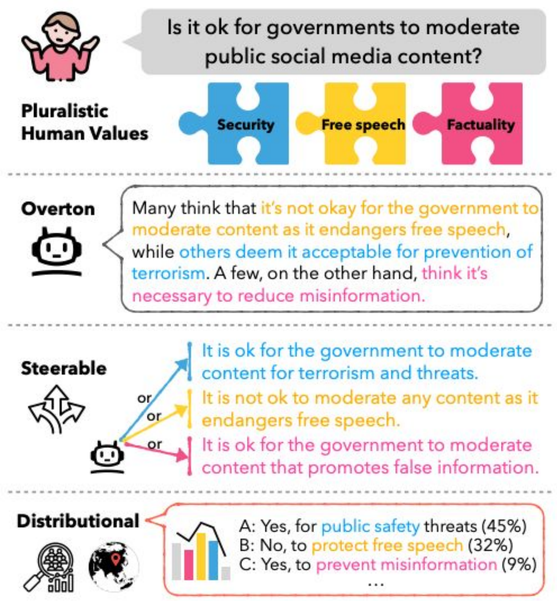

Figure 2: Three kinds of Pluralism in Models [1]

CINEMETRIC

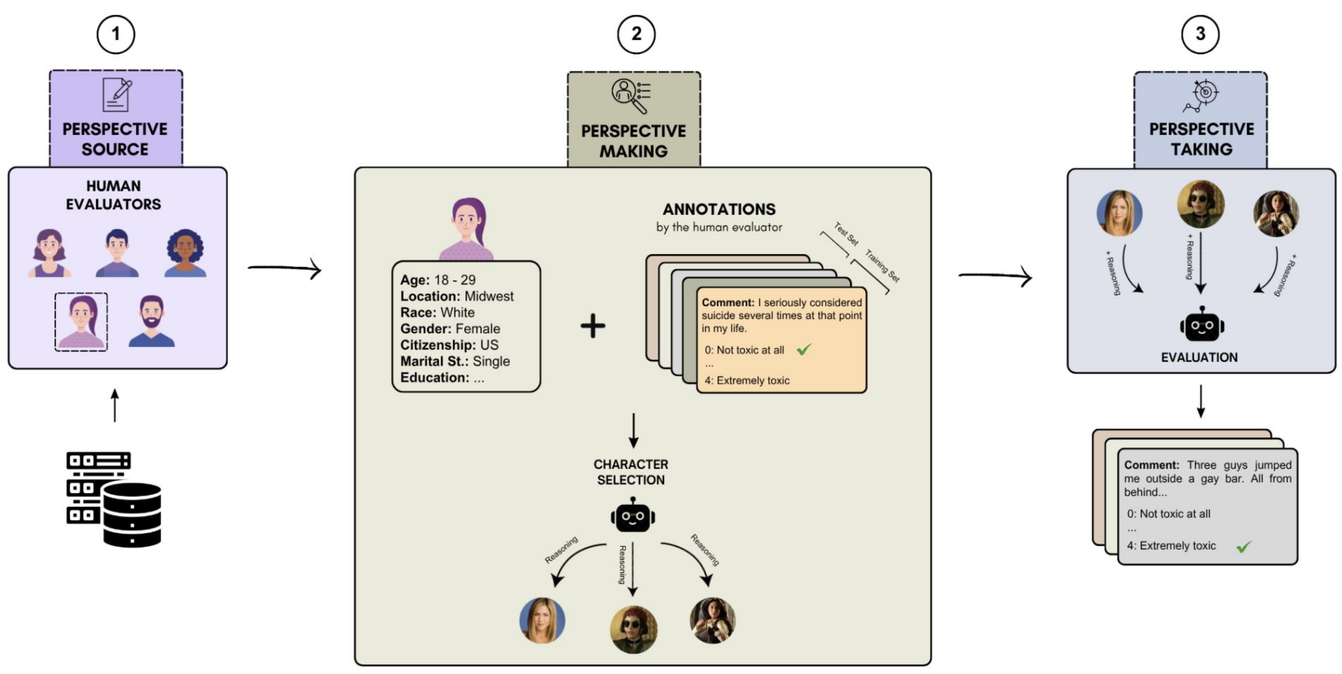

Figure 3: A high-level overview of CINEMETRIC.

TASKS & MODELS

- **OpinionQA:** opinions with various demographic groups over different topics.
- **DP:** Toxic Content Classification for a Diversity of Perspectives.

- **OpenAI:**
  - GPT 4.1
- **DEEPSEEK:**
  - DeepSeek V3
- **GEMINI:**
  - Gemini Flash 2.5
- **MISTRAL:**
  - Mistral Medium

METHODS

- **LLM as Judge:** Evaluates responses based on a learned reward signal.
- **LLM as Personalized Judge** [2]**:** Adapts the evaluation to a specific user's profiles.
- **CINEMETRIC:** Combines human experiences with the LLM's perspective-taking abilities.

RESULTS & ANALYSIS

Method	DeepSeek	OpenAI	Google	Mistral
	DeepSeek V3	GPT 4.1	Gemini Flash 2.5	Mistral Medium
LLM as a Judge	37.71	45.26	43.56	43.83
LLM as a Personalized Judge	43.27	48.83	49.12	45.42
CINEM. w/o Training Examples	50.00	50.29	48.83	46.79
CINEM. w/o Character Names	52.92	51.16	51.46	**52.92**
CINEMETRIC	**57.31**	**52.33**	**53.53**	48.75

Method	DeepSeek	OpenAI	Google	Mistral
	DeepSeek V3	GPT 4.1	Gemini Flash 2.5	Mistral Medium
LLM as a Judge	31.11 (1.183)	45.83 (0.9)	43.06 (0.967)	31.37 (1.07)
LLM as a Personalized Judge	37.22 (0.981)	45.00 (0.9)	41.34 (0.934)	27.33 (1.064)
CINEM. w/o Training Examples	37.50 (0.972)	46.11 (0.872)	43.89 (0.844)	31.11 (1.05)
CINEM. w/o Character Names	43.33 (0.847)	47.50 (0.867)	52.78 (0.683)	35.46 (0.904)
CINEMETRIC	**46.94** (0.747)	**49.61** (0.808)	**54.72** (0.653)	**38.27** (0.891)

Table 1: Performance (Accuracy & MAE) of LLMs across different methods on OpinionQA & DP.

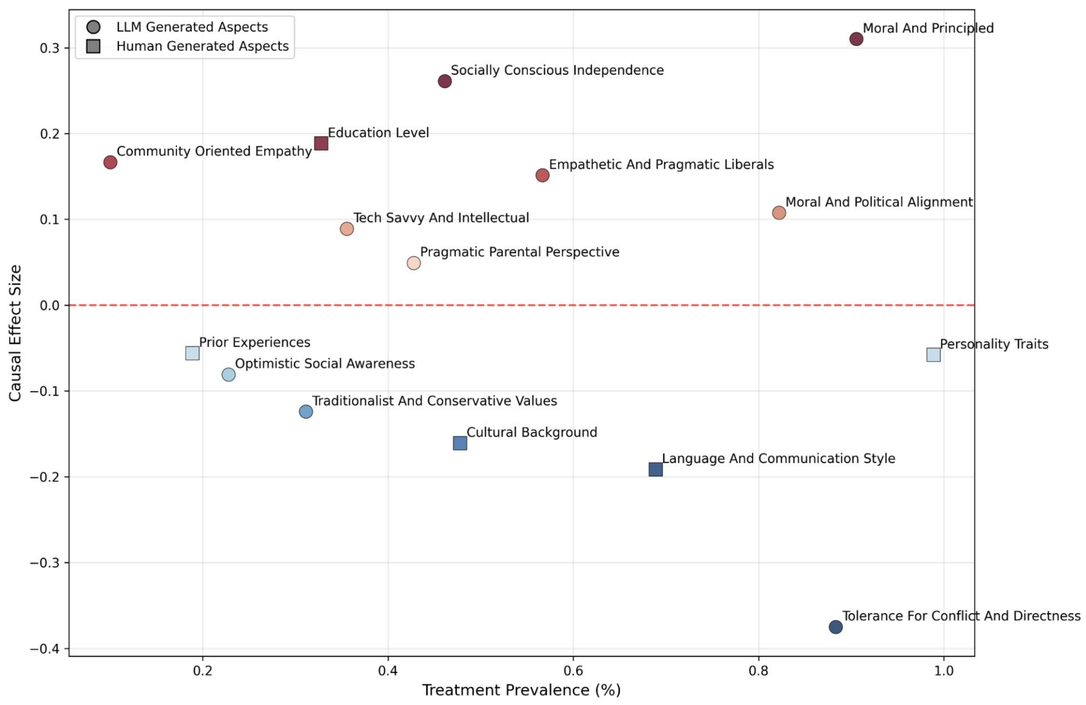

Figure 4: Treatment Prevalence vs. Causal Effect Size.

References

1 Taylor Sorensen et al. (2024), Position: a roadmap to pluralistic alignment. In Proceedings of the 41st International Conference on Machine Learning (ICML'24).
2 Yijiang River Dong et al. (2024), Can LLM be a Personalized Judge?. Findings of the Association for Computational Linguistics: EMNLP 2024.

Partner institutions:

technische universität dortmund
Fraunhofer IAIS
UNIVERSITÄT BONN
Fraunhofer IML

Institutionally funded by:

Bundesministerium für Bildung und Forschung
Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen