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

Vahid Sadiri Javadi **Zain Ul Abedin** **Lucie Flek**
Conversational AI and Social Analytics (CAISA) Lab, University of Bonn
{vahid.sadirij, zainabedin, lflek}@uni-bonn.de

Abstract

Despite advances in conversational systems, the evaluation of such systems remains a challenging problem. Current evaluation paradigms often rely on costly homogeneous human annotators or oversimplified automated metrics, leading to a critical gap in socially aligned conversational agents, where pluralistic values (i.e., acknowledging diverse human experiences) are essential to reflect the inherently subjective and contextual nature of dialogue quality. In this paper, we propose **CINEMETRIC**, a novel framework that operationalizes pluralistic alignment by leveraging the perspectivist capacities of large language models. Our approach introduces a mechanism where LLMs simulate a diverse set of evaluators, each with distinct personas constructed by matching real human annotators to movie characters based on both demographic profiles and annotation behaviors. These role-played characters independently assess subjective tasks, offering a scalable and human-aligned alternative to traditional evaluation. Empirical results show that our approach consistently outperforms baseline methods, including LLM as a Judge and as a Personalized Judge, across multiple LLMs, showing high and consistent agreement with human ground truth. CINEMETRIC improves accuracy by up to 20% and reduces mean absolute error in toxicity prediction, demonstrating its effectiveness in capturing human-like perspectives.

1 Introduction

What makes a conversation good? If we ask ten people, we might get ten different answers. As shown in Figure 1 (Human Evaluators), a response that one person finds relatively empathetic might strike another as less empathetic or even offensive. These differences highlight that the quality of the dialogue is inherently subjective and multifaceted (Foster et al., 2009). Yet the way we evaluate conversational systems today often assumes that there is an objective fact (by using automatic evaluation metrics) or a single definitive measure of quality (by aggregating ratings from crowdworkers or domain experts) to rate qualities such as coherence or overall satisfaction (Siro, 2023). However, this "one-size-fits-all" human annotation approach has its own blind spots. It implicitly assumes a homogeneous pool of evaluators, averaging out individual differences. In reality, the background, values, cognitive styles, or personality of an annotator may significantly influence how they perceive the quality of a conversation (Prabhakaran et al., 2021; Gautam and Srinath, 2024). Thus, both purely automatic metrics and aggregated human ratings risk missing the plurality of perspectives inherent in dialogue quality, struggling to capture the nuanced dimensions in a scalable and robust way.

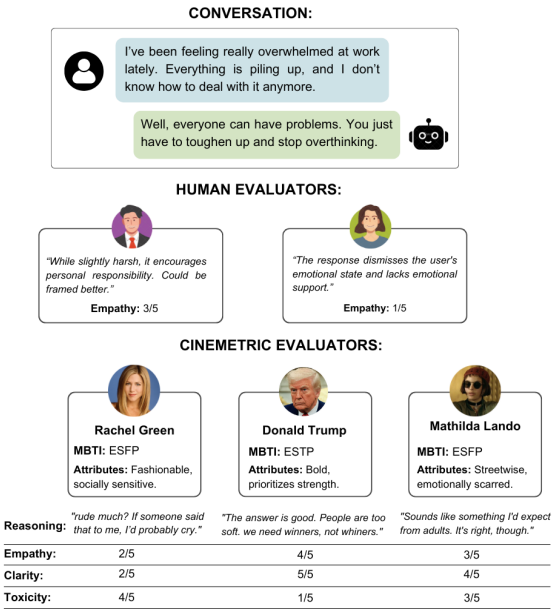

Figure 1: Comparison of Human vs. CINEMETRIC Role-Played Evaluators. A user expresses emotional distress, and the agent responds. Human evaluators and three distinct characters evaluate the agent's response. In the CINEMETRIC evaluation, each character reflects a unique personality profile, resulting in diverse ratings and subjective commentary across various dimensions.

Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP (NLPerspectivists 2025), pages 15–26

To address this gap, we introduce CINEMETRIC, a novel framework grounded in perspectivist principles for pluralistic alignment (Feng et al., 2024b; Castricato et al., 2024). This idea breaks away with from the singularity of current methods by embracing pluralism (Feng et al., 2024a). As illustrated in Figure 1, the core idea behind this framework is to simulate a panel of diverse evaluators (i.e., movie and public characters), each embodied as a distinct perspective-driven persona, capable of assessing a conversation through different lenses. These personas are defined by interpretable attributes such as gender, personality traits (e.g., MBTI types), thinking style (e.g., analytical vs. intuitive), and more. We then task the LLM to role-play these personas and evaluate conversational turns along multiple dimensions, including but not limited to toxicity, persuasiveness, clarity, and empathy. This work is guided by the following research questions:

RQ1. How can the perspectivist role-playing of diverse personas by large language models enhance the pluralistic alignment of conversational agents?

RQ2. How can Human-AI Collaboration be used to design an evaluation framework that captures the diversity of human values and preferences in LLM outputs?

RQ3. To what extent can perspective-driven evaluations, as instantiated by CINEMETRIC, approximate human judgments and enhance alignment with diverse human evaluative preferences?

In this paper, we propose several steps that help answer our RQs: *(i)* We investigate the conceptual foundations of perspectivism and the need for pluralistic alignment in conversational systems. (Section 2), *(ii)* We detail the design of the CINEMETRIC framework, outline our methodology for persona construction and evaluation design (Section 3). *(iii)* We describe our experimental setup, including how we simulate annotator perspectives through character personas, construct evaluation tasks, and define comparison baselines (Section 4). *(iv)* We present empirical results demonstrating the effectiveness of our method in capturing individual annotator perspectives compared to the baselines (Section 5).

2 Background

In this section, we discuss how perspectivism can be operationalized by the role-playing technique. We investigate the current approaches for evaluating conversational systems, and finally, we explore the pluralistic alignment in conversational agents.

2.1 Perspectivism and Role-Playing

Perspectivism, rooted in Nietzsche's philosophical tradition, refers to the idea that there is no singular objective viewpoint for many problems (Anderson, 1998; Cox, 1997), instead, understanding is shaped by diverse perspectives. Recent work in NLP has embraced this notion by treating annotator disagreements not as noise but as a valuable signal (Uma et al., 2021). For example, Basile (2020) advocates disaggregating annotation labels to preserve individual annotators' viewpoints instead of enforcing a single "ground truth," thereby capturing genuine differences in opinion. This perspectivist approach aims to avoid marginalizing minority opinions and to train models that recognize a spectrum of valid interpretations (Muscato et al., 2025).

One effective way to apply perspectivism in conversational systems is through role-playing or persona-based prompting of LLMs. Role-playing represents a core human ability to simulate different viewpoints and engage in perspective-taking(Jones, 1973). Prior work has shown that role prompts can improve the clarity and relevance of responses by aligning them with the implied perspective of the role (e.g., a "doctor" role yielding medically grounded explanations) (Tseng et al., 2024; Wang et al., 2024; Sun et al., 2024). At the same time, researchers caution that persona prompts can reinforce stereotypes if the model's training data contains biased representations of that role. (Park et al., 2025; Tseng et al., 2024; Tan and Lee, 2025). These studies highlight how role-playing with LLMs provides a versatile framework to inject perspectivism into conversational agents.

2.2 Evaluating Conversational Systems

Evaluating dialogue systems remains a difficult problem in NLP. Traditional evaluation methods for conversational systems have typically fallen into two categories: automated metrics and human evaluation. Automated metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and more recent neural embedding-based approaches like BERTScore (Zhang et al., 2019), which com-

putes similarity scores using contextual embeddings, and contextually sensitive models such as ADEM (Lowe et al., 2018) and DynaEval (Lowe et al., 2018), which enhance correlation with human judgments by considering dialogue context and structure. However, these metrics often fail to capture the nuanced aspects of dialogue quality that matter to humans (Liu et al., 2016). Human evaluation, while more aligned with user experiences, faces challenges of cost, scalability, and annotator variance (Smith et al., 2022; Liu et al., 2022).

A recent trend is to leverage large language models themselves as automatic judges of dialogue (i.e., LLM-as-a-Judge) (Gu et al., 2024; Chan et al., 2023). Instead of a fixed metric formula, one can prompt an advanced LLM (e.g., GPT-4) with a conversation and ask it to provide a rating or feedback, possibly with an explanation. For example, Chiang and Lee (2023) showed that ratings given by ChatGPT-based evaluators correlated more strongly with human judgments than traditional metrics like BLEU or BERTScore. Furthermore, Dong et al. (2024) demonstrated that the standard LLM-as-a-Judge setting is not sufficiently reliable for personalization tasks, showing low agreement with human ground truth. They identified persona sparsity as a major cause of this unreliability. Thus, efforts to infuse evaluation with multiple perspectives from different backgrounds are a direct motivation for the CINEMETRIC framework.

2.3 Pluralistic Alignment in Conversational Agents

As conversational agents become more powerful and widespread, the goal of alignment, i.e., ensuring that an agent's behavior is consistent with human values and intentions, has taken center stage. Traditional alignment approaches, such as reinforcement learning from human feedback (RLHF) (Bai et al., 2022), typically optimize models to perform on average according to the preferences of a broad user base or a set of guidelines (e.g., being helpful, truthful, and harmless). Kirk et al. (2023) show that standard RLHF tends to collapse a model's behavior towards a central norm, reducing the richness of responses it can generate.

The concept of pluralistic alignment begins by questioning *"Whose values?"* current systems are aligned to (Bergman et al., 2024), and it has emerged as a response to the limitations of monolithic evaluation approaches (Conitzer et al., 2024). Gabriel (2020) argues that conversational agents should be designed to acknowledge and respect the diversity of human values rather than optimizing for a single objective function. This perspective aligns with Rawls (1971) concept of *"reasonable pluralism"*, which recognizes that a just society must accommodate diverse and sometimes conflicting conceptions of the good. Moreover, Feng et al. (2024b) have argued that alignment must be reconceived as a socially situated process, acknowledging the pluralism of society rather than pretending there is a single correct value system for a conversational agent. Therefore, researchers proposed diversity-aware alignment frameworks. For instance, pluralistic alignment as defined by Sorensen et al. (2024) is the capacity of conversational agents to handle a plurality of values or preferences, instead of being narrowly tuned to one.

Given the fact that evaluation is inherently multi-perspective, and that we can now harness LLMs to simulate those perspectives in a principled and reproducible way, CINEMETRIC offers a novel solution, namely, a framework that explicitly encodes pluralism into the evaluation pipeline.

3 Methodology: CINEMETRIC

The CINEMETRIC framework proceeds in three steps as shown in Figure 2: *(i)* **Perspective Source** (i.e., selecting representative human evaluators from multi-perspective datasets (Section 3.1)), *(ii)* **Perspective Making** (i.e., creating a set of "persona" movie characters and their corresponding perspectives for each sampled human evaluator (Section 3.2)), and *(iii)* **Perspective Taking** (i.e., leveraging the movie characters and their perspectives by LLMs to make predictions on held-out human evaluator annotations (Section 3.3)).

3.1 Perspective Source

The first step in our framework involves sourcing diverse human perspectives that serve as the grounding for the rest of the framework. To construct this perspective source, we draw from existing datasets that include both: (i) annotations made by individual human evaluators on subjective tasks such as toxicity classification or multiple-choice opinion questions, and (ii) demographic metadata about each evaluator (e.g., age, gender, location, race, political orientation, marital status, education level, etc.). From each dataset, we randomly sample a fixed number of evaluators.

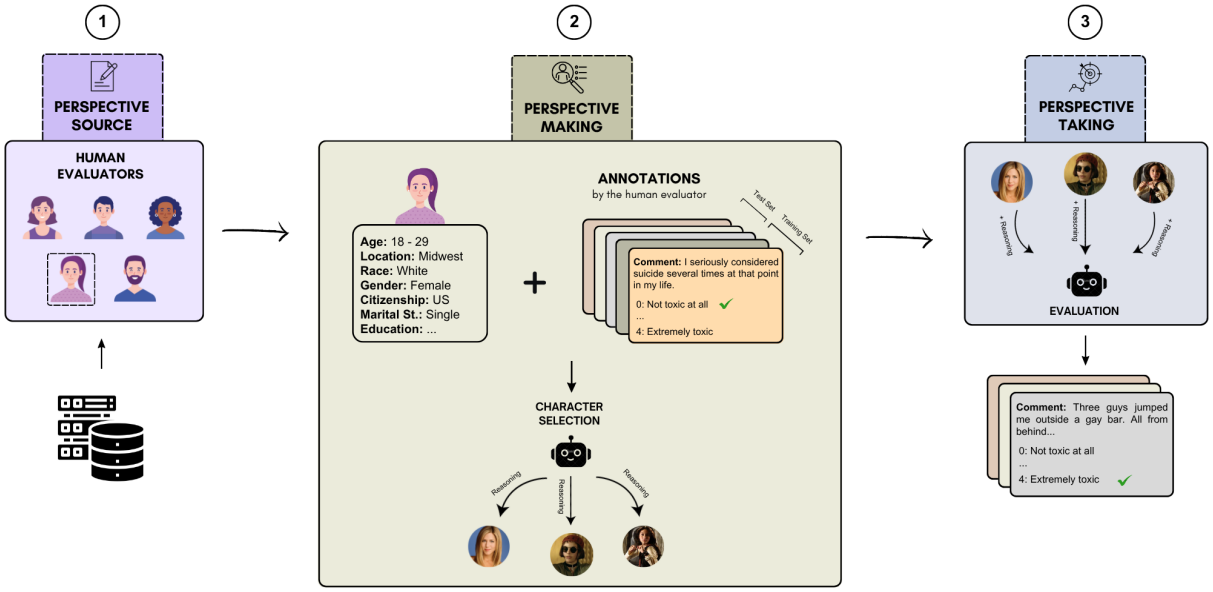

Figure 2: A high-level overview of **CINEMETRIC**, consisting of three steps: ① Perspective Source (See Section 3.1), ② Perspective Making (See Section 3.2), ③ Perspective Taking (See Section 3.3). See Appendix B for prompts for each step. To demonstrate how CINEMETRIC operates in practice, we provide a detailed example in Appendix C centered on a human annotator and the corresponding personas assigned to represent their perspective. We also include the reasoning behind the selection of each movie character, highlighting how their traits align with the evaluator's values and annotation.

3.2 Perspective Making

In this step, we transform each human evaluator into a small set of movie characters who share that evaluator's demographic profile and annotation tendencies.

Concretely, for each sampled human evaluator, we perform the following steps using a large language model: First, we compile the annotator's demographic metadata (e.g. age, gender, region, education, etc.) and some examples of their annotations or responses (training set). Then, we task the LLM to list five movie characters who are demographically and behaviorally similar to this person, given the demographic data as well as the examples of annotations. The LLM also provides a detailed reasoning for each character selection, explaining how the movie character's traits, background, and personality connect to the annotator's profile and annotation patterns, meaning that the characters together capture multiple facets of the human evaluator's perspective. Finally, as the LLM may hallucinate non-existent characters, we verify each suggested character by asking the LLM to check the existence of the movie character in the suggested movie. Any suggested character that the LLM cannot confirm is discarded. From the original five, we keep the first three characters that pass validation. In rare cases where fewer than three valid characters are found, we repeat the generation to produce additional candidates.

After this process, each real annotator is represented by three personas (movie characters), each described by a name and a short rationale for the match. These personas are intended to embody different, plausible user perspectives aligned with the original annotator's demographics and behavior.

3.3 Perspective Taking

Finally, we use the LLM to role-play each of the three movie characters to predict how the persona would respond to each held-out test query of an annotator (test set). Therefore, each of the three personas produces a predicted label or answer for the query. We compute the final prediction by majority vote among the three. If all three differ, we break ties by a fixed rule, choosing the prediction of the first-listed persona.

4 Experimental Setup

To comprehensively evaluate the effectiveness of our proposed approach, we conduct experiments across a diverse set of tasks and models, employing various techniques for comparison.

4.1 Evaluation Tasks

We focus our evaluation on two distinct subjective tasks that require reasoning over human perspectives. In particular, we utilize:

OpinionQA (Santurkar et al., 2023): a multiple-choice question-answering dataset based on U.S. public opinion surveys. OpinionQA contains responses from thousands of respondents, each annotated with 12 demographic features (e.g. age, gender, region, education, political ideology, race, etc.). Each respondent answered 50 questions on various topics, with an average of 3–4 answer choices per question. In our setting, the LLM selects the most likely answer based on a simulated annotator's perspective.

DP (Diversity of Perspectives) (Kumar et al., 2021): a large toxicity annotation dataset. In DP, 17,280 participants each assigned a score from 0 (not toxic) to 4 (very toxic) to 20 social-media comments (drawn from Twitter/Reddit/4chan) and each annotator provided demographic information and personal background. In total, the dataset contains 107,620 comment judgments linked to annotator metadata. The task captures how annotators with diverse backgrounds perceive offensive content differently. In our setting, the LLM assigns a score to the online comments.

In our experiments, for each dataset, we randomly select 100 annotators. Each annotator has a set of annotations (for DP) or answers (for OpinionQA) and demographic information. To simulate their perspectives using CINEMETRIC, we randomly sample 5 examples from each annotator's data for training, used to select the movie characters and construct the perspective-driven reasoning, and 10 examples for testing the evaluation performance. This results in a total of 1,000 test instances (100 annotators × 10 examples), each representing a comment or question to be evaluated by our framework.

4.2 Benchmarking Models

To evaluate the performance of our approach across a wide range of Large Language Models, we experiment with the following LLM families:

1. **DeepSeek**: DeepSeek-V3
2. **OpenAI**: GPT-4.1
3. **Google**: Gemini 2.5 Flash
4. **Mistral**: Mistral Medium 3

These models were selected to cover a wide spectrum of capabilities, sizes and families, enabling us to test CINEMETRIC's robustness across different LLMs. We use the same persona-generation and inference prompts across models. More details on the implementation can be found in Appendix A.

4.3 Methods Studied

We compared our proposed approach against several techniques. These approaches are chosen to cover a range of strategies.

LLM-as-a-Judge (Zheng et al., 2023): In this approach, the LLM is directly prompted to answer questions or assess toxicity without any personalization. This represents the default, non-perspectivist evaluation strategy.

LLM-as-a-Personalized-Judge (Dong et al., 2024): In this approach, the LLM receives demographic metadata of the target annotator to judge user preferences based on personas. This technique constructs personas but does not simulate perspective-taking. This represents a personalization baseline.

Ours: CINEMETRIC: This approach represents the core of our proposed method (described in Section 3), which uses training examples and demographics to match each annotator to movie characters. These characters are then role-played to predict labels on the test set, with majority voting.

Ours: CINEM. w/o Training Examples: To measure the effectiveness of our proposed approach, we remove behavioral data (i.e., the annotator's annotations) from the perspective making step, using only demographic metadata to select movie characters.

Ours: CINEM. w/o Character Names: An ablation, in which the LLM receives the annotator's metadata and behavioral examples, but does not use character names for perspective-taking (i.e., only for perspective-making). Instead, the LLM directly simulates the annotator.

These variants allow us to examine the impact of behavior-based persona construction (i.e., incorporating examples of evaluators' annotations) and the usefulness of using well-known movie characters for grounding perspectives.

4.4 Evaluation Metrics

For OpinionQA, we report accuracy in predicting the annotator's answer. For DP, we report both accuracy (exact match with the annotator's score) and mean absolute error (MAE) to account for near-miss predictions in ordinal toxicity judgments.

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

Table 1: The performance (accuracy) of different methods with various LLMs on the OpinionQA dataset.

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 2: On DP, CINEMETRIC consistently outperforms other techniques. We present the performance (accuracy & MAE) of different methods with various LLMs on the DP dataset.

5 Results & Analysis

We evaluate the performance of CINEMETRIC and competing methods on two datasets across a diverse set of LLMs. As shown in Tables 1 and 2, CINEMETRIC consistently outperforms all baseline approaches, demonstrating its robustness and adaptability across different model families and evaluation formats.

5.1 Performance on OpinionQA

Table 1 presents detailed experimental results on OpinionQA. CINEMETRIC achieves the highest accuracy across all LLMs, surpassing both the LLM as a Judge and LLM as a Personalized Judge baselines. For instance, on DeepSeek V3, CINEMETRIC achieves 57.31% accuracy, which is a substantial improvement over the strongest baseline (Personalized Judge) by significant margins of about 15%. Similar gains are observed on GPT-4.1 (52.33% vs. 48.83%), and on a smaller model like Mistral Medium (48.75% vs. 45.42%). These results indicate that the combination of character-based simulation and perspective-driven alignment significantly enhances model performance in capturing annotator perspectives.

5.2 Performance on DP

Results on DP are presented in Table 2. Performance on the DP dataset reinforces our findings on OpinionQA. CINEMETRIC achieves the highest accuracy on every model, with notable improvements in both categorical prediction and mean absolute error (MAE). For example, on DeepSeek V3, CINEMETRIC reaches 46.94% accuracy with a MAE of 0.747, compared to 37.22% and 0.981 for the Personalized Judge baseline. On GPT-4.1, CINEMETRIC maintains its lead with 49.61% accuracy and a MAE of 0.808. Gemini shows particularly strong results, where CINEMETRIC achieves 52.78% accuracy, again with the lowest MAE (0.683), reflecting better ordinal sensitivity. Even on Mistral, the least capable model in our suite, CINEMETRIC improves performance to 38.27% accuracy with a MAE of 0.891, surpassing all alternative approaches.

5.3 Analysis of CINEMETRIC Aspects

Our baselines (i.e., CINEM. w/o Training Examples & CINEM. w/o Movie Characters) highlight the individual contributions of CINEMETRIC's components. Removing the behavioral training examples (CINEM. w/o Training Examples) consistently reduces accuracy and increases MAE across models, underscoring the value of using

human-authored examples to align LLM behavior. When movie characters are excluded (CINEM. w/o Movie Characters), performance generally drops as well, though the magnitude of the decline varies by model. Notably, for Mistral on OpinionQA, the version without movie characters slightly outperforms the full model. This suggests that in resource-constrained models, reducing simulation complexity may be beneficial, possibly due to prompt length limitations or reduced capacity for role-play reasoning. Nevertheless, across all other settings, the full CINEMETRIC framework provides the best overall performance, reaffirming the utility of combining character-based simulation with perspective-driven alignment.

6 Conclusion

In this work, we introduced CINEMETRIC, a novel evaluation framework that operationalizes perspectivist alignment by simulating diverse evaluative standpoints through LLM role-play. By drawing on a rich set of character-based personas, our approach provides a scalable, pluralistic alternative to monolithic evaluation practices. Through comprehensive experiments on two diverse benchmarks and across four leading LLM families, we demonstrated that CINEMETRIC consistently outperforms existing evaluation strategies in both accuracy and MAE. Our results highlight the value of perspective-driven simulation in enhancing the human-likeness and value-diversity sensitivity of automated evaluations. In particular, CINEMETRIC achieves stronger agreement with human judgments than standard LLM-based or personalized-LLM evaluation baselines.

Limitations

Dataset limitations: In this study, we evaluated CINEMETRIC using only two benchmark tasks (OpinionQA and DP), which are diverse in format and domain, but do not exhaust the full range of scenarios in which perspectivist evaluation may be useful. Further evaluation on broader datasets, including open-domain conversations and underrepresented demographic viewpoints, will be explored in future work to strengthen the generalizability of our framework.

Analysis limitations: While our results show that CINEMETRIC exhibits higher agreement with human annotators compared to existing approaches, our current analysis focuses primarily on aggregate accuracy and mean absolute error. We do not yet conduct fine-grained error analyses on persona-specific disagreements or examine how specific attributes (e.g., gender, neurotype) contribute to evaluation variance. Additionally, our agreement metrics are indirect (e.g., accuracy on human-labeled responses), rather than derived from inter-rater correlation with actual human raters on a per-instance basis. A deeper investigation into persona-level contributions and alignment dynamics will help better characterize the interpretability and fairness of CINEMETRIC.

References

R Lanier Anderson. 1998. Truth and objectivity in perspectivism. *Synthese*, 115:1–32.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Valerio Basile. 2020. It's the end of the gold standard as we know it: Leveraging non-aggregated data for better evaluation and explanation of subjective tasks. In *International Conference of the Italian Association for Artificial Intelligence*, pages 441–453. Springer.

Stevie Bergman, Nahema Marchal, John Mellor, Shakir Mohamed, Iason Gabriel, and William Isaac. 2024. Stela: a community-centred approach to norm elicitation for ai alignment. *Scientific Reports*, 14(1):6616.

Louis Castricato, Nathan Lile, Rafael Rafailov, Jan-Philipp Fränken, and Chelsea Finn. 2024. Persona: A reproducible testbed for pluralistic alignment. *arXiv preprint arXiv:2407.17387*.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*.

Vincent Conitzer, Rachel Freedman, Jobst Heitzig, Wesley H Holliday, Bob M Jacobs, Nathan Lambert, Milan Mossé, Eric Pacuit, Stuart Russell, Hailey Schoelkopf, et al. 2024. Social choice should guide ai alignment in dealing with diverse human feedback. *arXiv preprint arXiv:2404.10271*.

Christoph Cox. 1997. The" subject" of nietzsche's perspectivism. *Journal of the History of Philosophy*, 35(2):269–291.

Yijiang River Dong, Tiancheng Hu, and Nigel Collier. 2024. Can LLM be a personalized judge? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10126–10141, Miami, Florida, USA. Association for Computational Linguistics.

KJ Feng, Inyoung Cheong, Quan Ze Chen, and Amy X Zhang. 2024a. Policy prototyping for llms: Pluralistic alignment via interactive and collaborative policymaking. *arXiv preprint arXiv:2409.08622*.

Shangbin Feng, Taylor Sorensen, Yuhan Liu, Jillian Fisher, Chan Young Park, Yejin Choi, and Yulia Tsvetkov. 2024b. Modular pluralism: Pluralistic alignment via multi-llm collaboration. *arXiv preprint arXiv:2406.15951*.

Mary Ellen Foster, Manuel Giuliani, and Alois Knoll. 2009. Comparing objective and subjective measures of usability in a human-robot dialogue system. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 879–887.

Iason Gabriel. 2020. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3):411–437.

Sanjana Gautam and Mukund Srinath. 2024. Blind spots and biases: exploring the role of annotator cognitive biases in nlp. *arXiv preprint arXiv:2404.19071*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Delmos J Jones. 1973. Culture fatigue: The results of role-playing in anthropological research. *Anthropological Quarterly*, pages 30–37.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2023. Understanding the effects of rlhf on llm generalisation and diversity. *arXiv preprint arXiv:2310.06452*.

Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. 2021. Designing toxic content classification for a diversity of perspectives. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, pages 299–318.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Xiangyang Liu, Tianxiang Sun, Junliang He, Jiawen Wu, Lingling Wu, Xinyu Zhang, Hao Jiang, Zhao Cao, Xuanjing Huang, and Xipeng Qiu. 2022. Towards efficient NLP: A standard evaluation and a strong baseline. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3288–3303, Seattle, United States. Association for Computational Linguistics.

Ryan Lowe, Michael Noseworthy, Iulian V. Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2018. Towards an automatic turing test: Learning to evaluate dialogue responses. *Preprint*, arXiv:1708.07149.

Benedetta Muscato, Praveen Bushipaka, Gizem Gezici, Lucia Passaro, Fosca Giannotti, and Tommaso Cucinotta. 2025. Embracing diversity: A multi-perspective approach with soft labels. *arXiv preprint arXiv:2503.00489*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Sanghyun Park, Boris Maciejovsky, and Phanish Puranam. 2025. Thinking with many minds: Using large language models for multi-perspective problem-solving. *Preprint*, arXiv:2501.02348.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. *arXiv preprint arXiv:2110.05699*.

John Rawls. 1971. An egalitarian theory of justice. *Philosophical Ethics: An Introduction to Moral Philosophy*, pages 365–370.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR.

Clemencia Siro. 2023. Evaluating task-oriented dialogue systems with users. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3495–3495.

Eric Michael Smith, Orion Hsu, Rebecca Qian, Stephen Roller, Y-Lan Boureau, and Jason Weston. 2022. Human evaluation of conversations is an open problem: comparing the sensitivity of various methods for evaluating dialogue agents. *arXiv preprint arXiv:2201.04723*.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. 2024. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*.

Guangzhi Sun, Xiao Zhan, and Jose Such. 2024. Building better ai agents: A provocation on the utilisation of persona in llm-based conversational agents. *Preprint*, arXiv:2407.11977.

Bryan Chen Zhengyu Tan and Roy Ka-Wei Lee. 2025. Unmasking implicit bias: Evaluating persona-prompted llm responses in power-disparate social scenarios. *arXiv preprint arXiv:2503.01532*.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in llms: A survey of role-playing and personalization. *arXiv preprint arXiv:2406.01171*.

Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2024. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. *Preprint*, arXiv:2307.05300.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

A Model Implementation Details

All experiments were conducted using the OpenRouter API [1]. Across all models, the results are averages over 5 runs with a temperature of 1.0 and a maximum number of tokens of 2048. The other parameters are set to their default values.

B Prompts Used in CINEMETRIC

We describe the prompts used for each step in the CINEMETRIC framework.

B.1 Perspective Making

B.1.1 Character Selection w/ Training Examples

```
### GOAL: ###
You are a movie character matching expert. Your
     goal is to suggest {num_characters} well-
    known movie characters that match a user
    profile.

### USER'S PROFILE: ###
{user_info}

### TRAINING EXAMPLES: ###
Here are some examples of how this user
    annotated comments/ answered questions:
{examples with ground truth}

### TASK: ###
Based on both the user's profile AND the
    examples shown above, suggest {
    num_characters} movie characters who would
    likely have similar perspectives on what
    content is toxic or harmful.

Consider:
- How this person's background might influence
    their views.
- Characters whose values and experiences would
     align with these judgment patterns.

### IMPORTANT RULES: ###
1. Only suggest real characters from real,
    widely known movies.
2. Choose characters from popular, mainstream
    movies that most people would recognize.
3. Make sure all {num_characters} characters
    are DIFFERENT.

### OUTPUT FORMAT ###
Format your response EXACTLY as follows:
Character Name:[character name 1]
Movie Title:[movie title 1]
Reasoning:[explain why you chose this character,
     considering the person's profile and the
    examples shown]
"""
```

B.1.2 Character Selection w/o Training Examples

```
### GOAL: ###
You are a movie character matching expert. Your
     goal is to suggest {num_characters} well-
    known movie characters that match a user
    profile.

### TASK: ###
Based on the following user profile, suggest {
    num_characters} well-known movie characters
     that match these characteristics:
{user_info}

### IMPORTANT RULES: ###
1. Only suggest real characters from real,
    widely known movies.
2. Choose characters from popular, mainstream
    movies that most people would recognize.
3. Make sure all {num_characters} characters
    are DIFFERENT.

### OUTPUT FORMAT ###
Format your response EXACTLY as follows:
Character Name:[character name 1]
Movie Title:[movie title 1]
```

B.1.3 Character Validation

```
### TASK: ###
You are a movie expert. You should verify if
    the character "{character_name}" is a
    character from the movie "{movie_title}".

### IMPORTANT RULES: ###
1. Only respond with "VALID" if you are 100%
    certain this character exists in this movie
    .
2. Respond with "INVALID" if:
   - The character doesn't exist in this movie
   - The movie doesn't exist
3. Do not provide any explanation, just respond
     with VALID or INVALID.
```

B.2 Perspective Taking

B.2.1 Evaluation w/ Reasoning

```
### GOAL: ###
You should role-play the following movie
    character and solve the task from the movie
     character's perspective, informed by how
    similar real people have approached these
    evaluations.

### MOVIE CHARACTER: ###
You should role-play {character_name} from the
    movie {movie_name}.
You must fully embody the role of {
    character_name} throughout the entire
    evaluation process. This means your
    judgment and reasoning must reflect how {
    character_name} behaves and thinks in {
    movie_name}.

### CHARACTER SELECTION RATIONALE: ###
The reason why this character was chosen is as
    follows:
{Reasoning}
```

[1] https://openrouter.ai/models

```
### TASK: ###
As {character_name}, evaluate ...

### INSTRUCTION: ###
Consider the character selection rationale
    above and the character's values and
    personality when making judgments. This
    means your evaluation must reflect what the
     movie character would genuinely prefer.
    You should not say your personal opinion or
     an objective analysis.

### OUTPUT FORMAT: ###
Your output should have the following format:
Reasoning: [The reason why you choose the
    response.]
Character's aspect: [The aspect of the movie
    character's perspective you use to make
    this judgment.]
Answer: [Your choice]
```

B.2.2 Evaluation w/o Reasoning

```
### GOAL: ###
You should role-play the following movie
    character and evaluate the toxicity of the
    following comment from the movie character'
    s perspective, informed by how similar real
     people have approached these evaluations.

### MOVIE CHARACTER: ###
You should role-play {character_name} from the
    movie {movie_name}.
You must fully embody the role of {
    character_name} throughout the entire
    evaluation process. This means your
    judgment and reasoning must reflect how {
    character_name} behaves and thinks in {
    movie_name}.

### TASK: ###
As {character_name}, evaluate ...

### OUTPUT FORMAT: ###
Your output should have the following format:
Reasoning: [The reason why you choose the
    response.]
Character's aspect: [The aspect of the movie
    character's perspective you use to make
    this judgment.]
Answer: [Your choice]
```

B.2.3 Evaluation w/o Movie Character

```
### TASK: ###
Your task is to evaluate ... from the
    perspective of a user who has the following
     characteristics: {user_info}

### USER RATIONALE: ###
{Reasoning}

### OUTPUT FORMAT: ###
Reasoning: [The reason why you choose the
    response.]
User's aspect: [The aspect of the user's
    perspective you use to make this judgment.]
Answer: [Your choice]
```

B.2.4 LLM-as-a-Personalized-Judge

```
### GOAL: ###
Based on the user profile provided below,
    evaluate ...

### User Profile: ###
{user_info}

### OUTPUT FORMAT: ###
Reasoning: [Explain your reasoning based on
    your profile]
Answer: [Your choice]
```

B.2.5 LLM-as-a-Judge

```
### TASK: ###
Evaluate ...

### OUTPUT FORMAT: ###
Answer: [Your choice]
```

C Case Study Example

To illustrate the workings of CINEMETRIC in a concrete scenario, we present an example involving a single real-world annotator and their three selected personas. The goal is to showcase how the model simulates different character perspectives during toxicity evaluation.

C.1 Persona Matching and Evaluation Walkthrough

Human Annotator Profile: The person is female. Racially, the person is Asian. She is 55 - 64 years old and has some college but no degree. The person views technology's impact as somewhat positive and uses social media, news media, video media, online forums. The person has personally seen toxic content online and considers toxic comments to be occasionally a problem. Politically, the person is independent and identifies as heterosexual. The person is a parent. Religion is very important to the person.

Selected Movie Characters and Rationales:

- **Mulan (Mulan, 1998)**: A disciplined, honor-driven Asian woman with strong moral values, aligned with the user's cultural identity and parenting role. Likely sensitive to personal attacks and sexually inappropriate content.
- **Helen Parr / Elastigirl (The Incredibles, 2004)**: A pragmatic parent and protector, inclined to flag comments that threaten familial or social wellbeing, while tolerating factual or benign competition-related statements.

- **Evelyn Wang (Everything Everywhere All At Once, 2022)**: A resilient Asian immigrant navigating chaos and family, with deep moral grounding. Strong aversion to disrespect, but tolerant of less consequential critiques or political slogans.

Task:
Rating the toxicity of the following comment: `Comment:` *"Denying genocide is a crime!"*

Role-Played Evaluations:

- **Mulan**:
  - **Reasoning**: Mulan, as a soldier and symbol of justice, sees genocide denial as morally egregious. Upholding truth and defending the vulnerable are central to her character.
  - **Character Aspect**: Military background and moral absolutism.
  - **Toxicity Score**: **2**
- **Helen Parr**:
  - **Reasoning**: As a mother and protector, Helen supports the comment as a factual and justice-aligned assertion. She interprets it as a call to accountability rather than toxicity.
  - **Character Aspect**: Balanced, justice-oriented parental perspective.
  - **Toxicity Score**: **0**
- **Evelyn Wang**:
  - **Reasoning**: Evelyn, grounded in spiritual resilience and firsthand experience with injustice, perceives the comment as morally justified. Her empathy for victims drives her interpretation.
  - **Character Aspect**: Lived trauma, spiritual values, and parental protectiveness.
  - **Toxicity Score**: **0**