

← Go to **ACL ARR 2026 January** homepage (</group?id=aclweb.org/ACL/ARR/2026/January>)

2 Versions ▾

Evaluating Language Model Pluralism through In-the-wild Crowd Discussions



Gagan Mundada (/profile?id=~Gagan_Mundada1),
Rohan Surana (/profile?id=~Rohan_Surana1),
Nandhini Swaminathan (/profile?id=~Nandhini_Swaminathan1),
Bodhisattwa Prasad Majumder (/profile?id=~Bodhisattwa_Prasad_Majumder1),
Junda Wu (/profile?id=~Junda_Wu1),
Julian McAuley (/profile?id=~Julian_McAuley1),
Zhouhang Xie (/profile?id=~Zhouhang_Xie1)

06 Jan 2026 (modified: 17 Mar 2026) ACL ARR 2026 January Submission
 January, Senior Area Chairs, Area Chairs, Reviewers, Authors, Commitment Readers
 Revisions (</revisions?id=t9TipOYRUI>) CC BY 4.0 (<https://creativecommons.org/licenses/by/4.0/>)

Keywords: LLM, Pluralistic Alignment

Abstract:

When answering subjective questions, an ideal LLM should surface diverse plausible perspectives rather than favoring a single viewpoint, a characteristic known as pluralism. Recent studies show that modern LLMs optimized through preference alignment systematically favor certain positions on subjective queries, making pluralism evaluation increasingly important. However, existing evaluation methods focus dominantly on multiple-choice and question-answering tasks, leaving open-ended generation largely unaddressed.

We propose PLURALEVAL, an evaluation framework that assesses LLM pluralism in open-ended generation by comparing outputs against free-form crowd responses. Our approach decomposes ground-truth responses into atomic, non-overlapping claims, then evaluates whether LLMs adequately cover this diverse claim space. We then introduce WildSCOPE, a multi-domain dataset of natural crowd responses, and demonstrate that PLURALEVAL captures novel insights, such as the collapse of pluralism through sycophancy, where LLM systematically degrades in overtone pluralism when a user's belief is revealed. Finally, we discuss the value and actionable insights for preserving and encouraging pluralism from LLM deployers' side.

Paper Type: Long

Research Area: Ethics, Bias, and Fairness

Research Area Keywords: model bias/fairness evaluation

Contribution Types: Model analysis & interpretability, Data resources

Languages Studied: English

Reassignment Request Area Chair: This is not a resubmission

Reassignment Request Reviewers: This is not a resubmission

A1 Limitations Section: This paper has a limitations section.

A2 Potential Risks: Yes

A2 Elaboration: Section 9 Limitation

B Use Or Create Scientific Artifacts: Yes

B4 Data Contains Personally Identifying Info Or Offensive Content: N/A

B6 Statistics For Data: Yes

B6 Elaboration: Section 5 Collecting WildScope

C Computational Experiments: Yes

C2 Experimental Setup And Hyperparameters: Yes

C2 Elaboration: Section 6

C3 Descriptive Statistics: Yes

C3 Elaboration: Section 6

D Human Subjects Including Annotators: Yes

D1 Instructions Given To Participants: Yes

D1 Elaboration: Section A.4

D2 Recruitment And Payment: N/A

D3 Data Consent: Yes

D4 Ethics Review Board Approval: N/A

E Ai Assistants In Research Or Writing: Yes

E1 Information About Use Of Ai Assistants: Yes

E1 Elaboration: Section 8

Author Submission Checklist: yes

TLDR: We introduce PLURALEVAL to evaluate whether LLMs adequately cover the diverse viewpoints people naturally express in crowd discussions.

Preprint: no

Preprint Status: We are considering releasing a non-anonymous preprint in the next two months (i.e., during the reviewing process).

Preferred Venue: ACL

Consent To Share Data: yes

Consent To Share Submission Details: On behalf of all authors, we agree to the terms above to share our submission details.

Association For Computational Linguistics - Blind Submission License Agreement: On behalf of all authors, I agree

Submission Number: 10831

Discussion (?id=t9TipOYRUI#discussion)

Filter by reply type...

Filter by author...

Search keywords...

Sort: Newest First

Everyone Submission10831... Submission10831... Submission10831...

10 / 10 replies shown

Submission10831... Program Chairs Submission10831... Submission10831...

Submission10831... Submission10831... Submission10831...

Meta Review of Submission10831 by Area Chair ormh

Meta Review by Area Chair ormh 02 Mar 2026, 21:00 (modified: 17 Mar 2026, 08:40)

Senior Area Chairs, Area Chairs, Authors, Reviewers Submitted, Program Chairs, Commitment Readers

Revisions (/revisions?id=qvH6BXMJRJ5)

Metareview:

The paper introduces PluralEval, a framework for evaluating "pluralism" (viewpoint diversity) in LLM generations, alongside WildSCOPE, a dataset of 1.2K Reddit threads across subjective domains. The methodology decomposes human comments into atomic units to measure how well LLM responses cover the spectrum of human opinions. The reviewers are highly positive. They laud the paper for addressing a timely and socially significant topic—the tendency of LLMs to stifle diverse perspectives in favor of sycophancy. The pipeline for data collection and the reframing of open-ended generation as a set-coverage problem are viewed as significant technical contributions.

Summary Of Reasons To Publish:

- The work provides a rigorous framework to evaluate whether LLMs capture or suppress the diversity of human viewpoints, which is critical as these models increasingly shape public discourse.
- The "atomic opinion" decomposition and clustering approach effectively bridges the gap between granular human comments and high-level model responses without requiring expensive manual rubrics.
- The experiments convincingly demonstrate that current preference alignment (e.g., RLHF) tends to induce asymmetric sycophancy and distorts the model's ability to judge the popularity of different viewpoints.
- WildSCOPE serves as a valuable resource for future research into subjective and argumentative AI behavior.

Summary Of Suggested Revisions:

- The authors should provide evidence (perhaps through a small-scale human study) that higher PluralEval scores actually correlate with human perceptions of pluralism. Additionally, the reliability of the "entailment-based matching" (Section 5) needs validation to account for potential false positives/negatives.
- Address Reviewer 83dG's confusion regarding why LLM-generated "guesses" of popularity were used instead of actual Reddit upvote data. Provide a clear definition of how the "least popular cluster" is determined and whether deltas in upvote counts (e.g., 300 vs 301) are handled.


Overall Assessment: 4 = Conference: I think this paper could be accepted to an *ACL conference.


Reported Issues:  No


Publication Ethics Policy Compliance: I did not use any generative AI tools for this review

General Response

Official Comment

by Authors ( Gagan Mundada (/profile?id=~Gagan_Mundada1), Bodhisattwa Prasad Majumder (/profile?id=~Bodhisattwa_Prasad_Majumder1), Zhouhang Xie (/profile?id=~Zhouhang_Xie1), Julian McAuley (/profile?id=~Julian_McAuley1), +3 more (/group/edit?id=aclweb.org/ACL/ARR/2026/January/Submission10831/Authors))

 21 Feb 2026, 17:31 (modified: 17 Mar 2026, 08:40)

 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers, Reviewers Submitted, Authors, Commitment Readers

 Revisions (/revisions?id=EtRSOzxIts)

Comment:


Dear reviewers, as author response period comes almost to an end, we'd like to thank you again for your careful reading and constructive feedback. We are encouraged by the enthusiastic and supportive evaluations. Reviewers recognize the timeliness and importance of the problem (83dG, wLb6), the novelty of PluralEval (JxCp), and its clear set coverage formulation (wLb6), the thoughtful framework and data collection pipeline (83dG, wLb6), and that experiments effectively demonstrate sycophancy and support our claims (83dG, JxCp).

We note that a few points raised as weaknesses are clarification-focused and can be resolved with a few additional sentences in the paper, such as the Overton pluralism convention we follow (83dG) and the scoping of Section 6.1's filtering criteria (83dG, JxCp) and the framing of our core contributions (83dG). Reviewers also suggested additional validation experiments (wLb6, 83dG), such as human evaluation of entailment reliability (wLb6) and robustness checks on popularity deltas (83dG), which we have conducted and confirm that our findings remain robust. We also appreciate the presentation suggestions and will incorporate them in the revision. Please see the individual responses for additional details and other points raised by reviewers, including cluster quality validation, suggested citations, and code release.

Please don't hesitate to reach out if you have any additional questions - we look forward to discussing with you.

Official Review of Submission10831 by Reviewer 83dG

Official Review by Reviewer 83dG  12 Feb 2026, 15:07 (modified: 17 Mar 2026, 08:40)

 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer 83dG, Commitment Readers

 Revisions (/revisions?id=HYnyTT4gjs)

Paper Summary:

The paper attempts to provide insight into whether responses provided by an LLM capture or stifle pluralism (defined here as diversity in views expressed in three primary domains). A dataset is collected from Reddit posts and broken down into clusters of “atomic” opinion units which express a single perspective. These clusters are summarized, and multiple experiments demonstrate that LLMs have a tendency towards: (1) sycophancy, in which the diversity of views is collapsed in favor of expressions that agree with the prompt; and (2) inability to accurately rank opinions by popularity (capturing the Overton window view of pluralism).

Summary Of Strengths:

This is an important topic area, as LLMs are progressively more widely used for a variety of tasks that may lead to having an impact on users’ opinions and perspectives. Pluralism, and more generally exposure to a range of viewpoints, has broadly positive impacts, while narrowing exposure can both lead people to misunderstand the range of views held by others and to strengthen their own opinions.

Significant work went into this paper, and the pipeline of data collection, data subdivision, clustering, and summarization is well constructed.

The experimental results provided effectively demonstrate a tendency towards sycophancy and a reduction of viewpoint diversity exposure—which is not surprising, but is important for working toward improvement, and the primary contribution is the mechanism for evaluating pluralism reduction rather than this immediate result.

For the most part, enough information is provided to allow good analysis and reproduction of the systems described. (However, while prompts are provided, examples of some of the variations present in the “personal belief” statements would be informative.)

Summary Of Weaknesses:

As far as I can determine, the majority of the work describes captures the presence of divergent viewpoints, but not the prevalence of those viewpoints. A small number of Reddit comments espousing certain beliefs (above a cutoff) are treated for the most part the same as a much more popular opinion (below a cutoff). These cutoffs are a somewhat simple measure when actual percentages could be used. There are definitions of pluralism that disregard prevalence, but the Overton window is not in that philosophical family.

There are ways in which the presentation could be clearer. For example, whether and how opinion prevalence was taken into account (see above) was not obvious until roughly section 6.2, which is the area in which the paper does attempt to consider prevalence, albeit still with a broad-brush “top ten” approach; the retention of discussions with a significant number of contradictory viewpoints (434–437) is a fairly major filter that could be exposed earlier in the discussion of the architecture. (Also, see detailed notes below.)

I am also not sure I understand the sentence “models rank cluster summaries from most to least popular based on expected Reddit upvotes.” The actual upvotes are available; why are LLM-generated guesses used (I think that’s what this means)?

6.3 depends on pairs of opinions in which one has more votes than the other, but it’s not obvious whether there is a minimum delta. If one opinion has 300 votes and another has 301, that should lead to different conclusions than larger deltas, and it’s not obvious that is taken into account.

Potential mitigation strategies are claimed as a contribution, but (1) are only present in an appendix and (2) are relatively high-level suggestions. This should be reframed to avoid making unsupported claims.

Comments Suggestions And Typos:

- The results presented in the text in lines 473–487 would be clearer as a table.
- Figure 3, by contrast, is rather hard to understand. Is there a different mechanism by which these deltas could be presented more explicitly?

Confidence: 3 = Pretty sure, but there's a chance I missed something. Although I have a good feel for this area in general, I did not carefully check the paper's details, e.g., the math or experimental design.

Soundness: 3.5

Excitement: 3.5

Overall Assessment: 4 = Conference: I think this paper could be accepted to an *ACL conference.

Limitations And Societal Impact:

This paper's primary contribution is an approach to measuring a popular family of models' tendency to reduce viewpoint diversity, which is a strong societal negative, and as such has the potential to have positive societal impact. However, reddit is significantly skewed demographically and ideologically (for example, the population has a heavy male skew). As such, any dataset collected therefrom should clearly state how this skew affects the interpretation of results (even if only to say that it does not, with justification).

Ethical Concerns:

There are no concerns with this submission

Needs Ethics Review: No

Reproducibility: 4 = They could mostly reproduce the results, but there may be some variation because of sample variance or minor variations in their interpretation of the protocol or method.

Datasets: 3 = Potentially useful: Someone might find the new datasets useful for their work.

Software: 1 = No usable software released.

Knowledge Of Or Educated Guess At Author Identity: No

Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Knowledge Of Paper Source: N/A, I do not know anything about the paper from outside sources

Impact Of Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Reviewer Certification: I certify that the review I entered accurately reflects my assessment of the work. If you used any type of automated tool to help you craft your review, I hereby certify that its use was restricted to improving grammar and style, and the substance of the review is either my own work or the work of an acknowledged secondary reviewer.

Publication Ethics Policy Compliance: I did not use any generative AI tools for this review



Response to Reviewer 83dG (2/2)

Official Comment

by Authors (Gagan Mundada (/profile?id=~Gagan_Mundada1), Bodhisattwa Prasad Majumder (/profile?id=~Bodhisattwa_Prasad_Majumder1), Zhouhang Xie (/profile?id=~Zhouhang_Xie1), Julian McAuley (/profile?id=~Julian_McAuley1), +3 more (/group/edit?id=aclweb.org/ACL/ARR/2026/January/Submission10831/Authors))

20 Feb 2026, 21:08 (modified: 17 Mar 2026, 08:40)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer 83dG, Commitment Readers

Revisions (/revisions?id=vXrc7QVMKr)

Comment:

Comments/Suggestions:

[C1] Clarity

- We have represented the results in Lines 473-487 as Table 4, as mentioned in Line 473

[C2] Figure 3 Correction

- We appreciate this feedback. We also spent some time exploring the best way to present these results during the drafting process. We considered alternatives such as heatmaps and grouped bar charts showing raw recall values side by side, but found that directly plotting the deltas better isolates the sycophancy effect from baseline variation across models and datasets. That said, we will add a clarifying sentence to the figure caption explaining how to read the deltas and what each quantity represents.

Limitations:

[L1] Reddit Limitation

- While we share the reviewer's concern that Reddit is not universally representative, which is an ongoing discussion in the community [9], many prior works in adjacent areas such as long-form question answering [10], fine-grained emotion detection [11], and dialogue generation [12] rely on

Reddit for the same reasons: it enables large-scale reproducible collection with natural discourse and voting signals. We will explicitly acknowledge the demographic skew and interpret our results as platform-specific in the revised paper.

- Additionally, PluralEval is a general framework not tied to Reddit. It can be readily extended to new data sources as they become available, and we hope the community can work together to build more representative opinion corpora for pluralism evaluation.

References:

- [1] Johnson, Elliott Aidan, Irene Hardill, Matthew T. Johnson, and Daniel Nettle. "Breaking the Overton Window: on the need for adversarial co-production." *Evidence & Policy* 20, no. 3 (2024): 393-405.
- [2] Morgan, Daniel J. "The Overton window and a less dogmatic approach to antibiotics." *Clinical Infectious Diseases* 70, no. 11 (2020): 2439-2441.
- [3] Youvan, Douglas C. "Shifting boundaries of acceptability: Examining the Overton Window and its modern manipulators in US discourse." unpublished paper (2024).
- [4] Singh Chauhan, Shivank. "Fragmented Overton Windows: Rethinking Political Viability in Polarised Public Spheres." Available at SSRN 5097975 (2024).
- [5] Sorensen, Taylor, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Miresghallah, Christopher Michael Rytting, Andre Ye et al. "A roadmap to pluralistic alignment." arXiv preprint arXiv:2402.05070 (2024).
- [6] Feng, Shangbin, Taylor Sorensen, Yuhang Liu, Jillian Fisher, Chan Young Park, Yejin Choi, and Yulia Tsvetkov. "Modular pluralism: Pluralistic alignment via multi-llm collaboration." In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4151-4171. 2024.
- [7] Lake, Thom, Eunsol Choi, and Greg Durrett. "From distributional to overton pluralism: Investigating large language model alignment." In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6794-6814. 2025.
- [8] Poole-Dayana, Elinor, Jiayi Wu, Taylor Sorensen, Jiaxin Pei, and Michiel A. Bakker. "Benchmarking Overton Pluralism in LLMs." arXiv preprint arXiv:2512.01351 (2025).
- [9] Pew Research Center. *Seven-in-Ten Reddit Users Get News on the Site*. 2016.
- [10] Fan, Angela, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. "ELI5: Long form question answering." In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 3558-3567. 2019.
- [11] Demszky, Dorottya, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. "GoEmotions: A dataset of fine-grained emotions." In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4040-4054. 2020.
- [12] Huryn, Daniil, William M. Hutsell, and Jinho D. Choi. "Automatic generation of large-scale multi-turn dialogues from Reddit." In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 3360-3373. 2022.



Response to Reviewer 83dG (1/2)

Official Comment

by Authors (Gagan Mundada (/profile?id=~Gagan_Mundada1), Bodhisattwa Prasad Majumder (/profile?id=~Bodhisattwa_Prasad_Majumder1), Zhouhang Xie (/profile?id=~Zhouhang_Xie1), Julian McAuley (/profile?id=~Julian_McAuley1), +3 more (/group/edit?id=aclweb.org/ACL/ARR/2026/January/Submission10831/Authors))

20 Feb 2026, 21:06 (modified: 17 Mar 2026, 08:40)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer 83dG, Commitment Readers

Revisions (/revisions?id=otz8qor1pS)

Comment:

We sincerely thank the reviewer for the careful reading and thoughtful comments. We are especially encouraged that the evaluation framework and overall pipeline were found to be well-constructed and meaningful for studying pluralism in open-ended generation. For your convenience, we numbered each bullet in your original comments and addressed them as follows:

Weakness:**[W1] On prevalence vs. presence of viewpoints**

- We agree that prevalence-sensitive pluralism is an important consideration. However, we wanted to note that while classically the Overton Window refers to the range of policy ideas considered politically acceptable, with boundaries shaped by public support and legitimacy [1–4], recent use of the term Overton pluralism in NLP has generalized this to an unweighted formulation, defining $W(x)$ as the set of reasonable answers and evaluating precision and recall over $W(x)$ without prevalence weighting [5,6,7], with [8] as a notable exception that also explores prevalence-weighted evaluation. To this end, our usage follows these recent works, but we will clarify our use of the term with appropriate citations in the final version.
- That said, while our primary framing focuses on viewpoint coverage, our experiments do go beyond presence and account for prevalence as well. Specifically, Sections 6.2 and 6.3 evaluate prevalence calibration by testing whether models accurately rank opinions by vote-grounded popularity and identify the more popular opinion in pairwise comparisons. We leave more fine-grained distributional analysis, such as recovering the exact proportion of opinions, to future work.

[W2] Presentation Issues

- We appreciate this suggestion. However, we wanted to clarify that Lines 434–437 describe a data-selection filter applied only to Section 6.1, where measuring opinion suppression requires threads with sufficient contradictory viewpoints for the analysis to be meaningful. This filter does not apply to the ranking (Section 6.2) or pairwise comparison (Section 6.3) experiments. We will make this distinction explicit in the final version.

[W3] Sentence Clarity

- Thank you for pointing this out. By "expected Reddit upvotes," we meant that the task is a ranking task by nature, and the idea is to evaluate whether the LLM knows which opinion would be more popular (i.e., LLM-expected opinion popularity), given a variety of opinions. This simulates a situation where the user understands the potential opinions relevant to a topic but wants to know which is more commonly accepted.
- In practice, each comment in a thread carries a vote score, which transfers to the atomic opinions extracted from it. Each opinion cluster in PluralEval then has a vote score aggregated from its constituent opinions, enabling us to set up this ranking task where the model should rank clusters with higher overall vote scores above those with lower ones. We will revise the wording in the paper to make this clearer.

[W4] Popularity Deltas Clarification

- Thanks for raising this point. We wanted to clarify that in our pairwise setup, we pair the most popular cluster with a randomly sampled cluster that is strictly less popular (not necessarily the least popular). The average vote gap across pairs is 64.7 for AskEconomics, 101.2 for AITA, and 144.3 for AskPolitics. We will clarify this selection criterion and report these statistics in the revised paper.
- To further verify robustness, we ran an additional experiment on AITA with GPT-4.1 using fully random cluster pairs with a minimum vote gap of 10, rather than fixing one side to the most popular cluster. The sycophancy effect persists, with neutral accuracy dropping from 77.4% to 71.2% under bias injection, confirming that our findings hold regardless of the pairing strategy. We agree this could be stated more clearly and will add a clarifying sentence along with these statistics in the final version.

[W5] Contribution Mismatch

- Thank you for pointing this out. We wanted to clarify that we do not claim mitigation strategies in the Appendix as our primary contribution. As outlined in Lines 119–141, our core contributions are the PluralEval framework, the WildScope dataset, and the empirical findings on sycophancy and prevalence calibration. The appendix discussion on mitigation is intended as an exploratory analysis.
- We will make this distinction more explicit around Line 118 in the introduction of the revised paper.

Official Review of Submission10831 by Reviewer wLb6

Official Review by Reviewer wLb6 📅 11 Feb 2026, 20:46 (modified: 17 Mar 2026, 08:40)

👁 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer wLb6, Commitment Readers

📄 Revisions (/revisions?id=ycV2DFyAKN)

Paper Summary:

The paper introduces **PluralEval**, an evaluation framework for measuring overton pluralism in open-ended LLM generations. The framework constructs a reference set of human viewpoints from Reddit posts by decomposing comments into atomic opinions, clustering the opinions, and summarizing the clusters into distinct thread-level viewpoints. Then, LLM responses are evaluated using coverage retrieval against the reference set using precision, recall, F-1, etc. The authors also introduce WILDSCOPE, a dataset of 1.2K Reddit threads across three domains spanning over subjective topics. Experiments reveal that belief injection induces asymmetric sycophancy by suppressing opposing views and distorting popularity judgments.

Summary Of Strengths:

- The paper is well-written and easy to understand.
- A well-motivated, timely topic that examines pluralism in open-ended generation tasks. Also, the paper reframes the open-ended generation task into a set coverage problem using standard retrieval metrics, which is very clear and avoids manual rubric design.
- A thoughtful design of **PluralEval**. The authors mined natural discussions from Reddit across three sub-communities, which reflects real discourse diversity in highly subjective and argumentative topics. Their decision to decompose comments into atomic opinion units before clustering can reduce granular mismatch between model responses and crowd responses.

Summary Of Weaknesses:

- The paper does not validate that higher PluralEval scores correspond to what humans also perceive as greater overton pluralism. Without showing correlation with human evaluations, it remains unclear whether the metric truly captures perceived pluralism.
- The metrics of coverage rely on entailment-based matching between model outputs and cluster summaries. However, the paper does not present an analysis of the reliability of this entailment matching. For example, there is no human validation of match correctness, nor false positives/false negatives in the matching results, etc.
- The PluralEval framework assumes that the proposed online clustering approach produces mutually distinct viewpoints, but this approach can raise ambiguity regarding viewpoint granularity. For example, one cluster can be a paraphrased restatement of another (e.g., morally right vs socially acceptable according to norms). It is unclear whether the authors checked and validated the mutual exclusivity and semantic distinctness of the generated clusters.

Comments Suggestions And Typos:

- Please address the above concerns.
- Please include in the Related Work section the following article related to the LLM pluralism topic - How Far Can We Extract Diverse Perspectives from Large Language Models? (<https://aclanthology.org/2024.emnlp-main.306/>) (Hayati et al., EMNLP 2024). This paper also evaluates pluralism in open-ended LLM generation tasks based on the recall-based coverage against human-generated opinions.
- Missing citations for GPT model families (lines 299, 426-427).

Confidence: 4 = Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.

Soundness: 4 = Strong: This study provides sufficient support for all of its claims. Some extra experiments could be nice, but not essential.

Excitement: 4 = Exciting: I would mention this paper to others and/or make an effort to attend its presentation in a conference.

Overall Assessment: 4 = Conference: I think this paper could be accepted to an *ACL conference.

Ethical Concerns:

There are no concerns with this submission

Reproducibility: 3 = They could reproduce the results with some difficulty. The settings of parameters are underspecified or subjectively determined, and/or the training/evaluation data are not widely available.

Datasets: 5 = Enabling: The newly released datasets should affect other people's choice of research or development projects to undertake.

Software: 4 = Useful: I would recommend the new software to other researchers or developers for their ongoing work.

Knowledge Of Or Educated Guess At Author Identity: No

Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Knowledge Of Paper Source: N/A, I do not know anything about the paper from outside sources

Impact Of Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Reviewer Certification: I certify that the review I entered accurately reflects my assessment of the work. If you used any type of automated tool to help you craft your review, I hereby certify that its use was restricted to improving grammar and style, and the substance of the review is either my own work or the work of an acknowledged secondary reviewer.

Publication Ethics Policy Compliance: I used a privacy-preserving tool exclusively for the use case(s) approved by PEC policy, such as language edits



Response to Reviewer wLb6 (2/2)

Official Comment

by Authors (Gagan Mundada (/profile?id=~Gagan_Mundada1), Bodhisattwa Prasad Majumder (/profile?id=~Bodhisattwa_Prasad_Majumder1), Zhouhang Xie (/profile?id=~Zhouhang_Xie1), Julian McAuley (/profile?id=~Julian_McAuley1), +3 more (/group/edit?id=aclweb.org/ACL/ARR/2026/January/Submission10831/Authors))

20 Feb 2026, 20:53 (modified: 17 Mar 2026, 08:40)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer wLb6, Commitment Readers

Revisions (/revisions?id=jxHDBJPsUf)

Comment:

References:

[1] Sanyal, Soumya, Tianyi Xiao, Jiacheng Liu, Wenya Wang, and Xiang Ren. "Are machines better at complex reasoning? unveiling human-machine inference gaps in entailment verification." In Findings of the Association for Computational Linguistics: ACL 2024, pp. 10361-10386. 2024.

[2] Greco, Candida Maria, Lucio La Cava, and Andrea Tagarelli. "Talking the talk does not entail walking the walk: On the limits of large language models in lexical entailment recognition." In Findings of the Association for Computational Linguistics: EMNLP 2024, pp. 14991-15011. 2024.

[3] Min, Sewon, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. "Factscore: Fine-grained atomic evaluation of factual precision in long form text generation." In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 12076-12100. 2023.

[4] Honovich, Or, Roei Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. "TRUE: Re-evaluating factual consistency evaluation." In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 3905-3920. 2022.

[5] Laban, Philippe, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. "SummaC: Re-visiting NLI-based models for inconsistency detection in summarization." Transactions of the Association for Computational Linguistics 10 (2022): 163-177.

[6] Wang, Zihan, Jingbo Shang, and Ruiqi Zhong. "Goal-driven explainable clustering via language descriptions." In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 10626-10649. 2023.

[7] Tamkin, Alex, Miles McCain, Kunal Handa, Esin Durmus, Liane Lovitt, Ankur Rathi, Saffron Huang et al. "Clio: Privacy-preserving insights into real-world ai use." arXiv preprint arXiv:2412.13678 (2024).

[8] Zheng, Lianmin, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin et al. "Judging llm-as-a-judge with mt-bench and chatbot arena." Advances in neural information processing systems 36 (2023): 46595-46623.



Response to Reviewer wLb6 (1/2)

Official Comment

by Authors (Gagan Mundada (/profile?id=~Gagan_Mundada1), Bodhisattwa Prasad Majumder (/profile?id=~Bodhisattwa_Prasad_Majumder1), Zhouhang Xie (/profile?id=~Zhouhang_Xie1), Julian McAuley (/profile?id=~Julian_McAuley1), +3 more (/group/edit?id=aclweb.org/ACL/ARR/2026/January/Submission10831/Authors))

20 Feb 2026, 20:52 (modified: 17 Mar 2026, 08:40)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer wLb6, Commitment Readers

Revisions (/revisions?id=AYqmqQ6Iu9)

Comment:

Thank you for the careful reading and constructive feedback. We are encouraged that you find the problem timely, the framework and dataset design thoughtful, and the set coverage formulation clear and well-motivated. We address your specific concerns below. For your convenience, we numbered each bullet in your original comments and addressed them as follows:

Weakness:

[W1] Human Perception

- Thank you for pointing this out. We note that PluralEval uses standard, well-established set-based metrics (precision, recall, F1) over human-constructed reference sets, rather than introducing a novel scoring function [3]. The core contribution lies in automating reference set construction, and we validate this component through human evaluation (Section A.4, Table 10), where annotators confirmed that our clustering produces semantically distinct viewpoints.
- On the other hand, the challenge with directly calibrating human perception of overall pluralism is that it requires annotators to be aware of all possible crowd viewpoints to judge coverage adequately, a setup that is difficult to realize in practice. We acknowledge this as an important direction for future work and will discuss it explicitly in the paper.

[W2] Entailment Matching

- Thank you for raising this point. We wanted to note that using LLMs for text-to-text entailment and matching is now standard practice across a range of NLP tasks [1,2], including factuality evaluation [3,4,5], LLM-guided clustering [6,7], and LLM-as-a-judge evaluation [8]. Our setup follows the same paradigm, using an LLM to judge whether a model-generated response entails an opinion from the ground-truth reference set.
- To validate reliability, we conducted a manual evaluation on a class-balanced set of 100 LLM-judged pairs with two annotators. The LLM judge agreed with human raters 80% of the time. We then manually inspected the disagreements and found them to be predominantly cases of leniency, where

the LLM matcher judged topically related but subtly different opinions as matching. Since this tendency applies across all evaluated models, the relative comparisons underlying our findings remain unaffected. We will report these results along with a detailed reliability analysis in the Appendix.

[W3] Viewpoint Granularity

- We agree that near-duplicate clusters and granularity ambiguity are important concerns. To this end, we conducted a detailed comparison of our clustering method against several baselines in Appendix A.4, including a human evaluation (Table 10) where annotators assessed whether clusters capture semantically distinct viewpoints rather than paraphrased restatements. Our method was preferred 65.5% of the time over a state-of-the-art LLM-based clustering method, suggesting it better identifies genuinely distinct opinions. We will make this validation more prominent in the main text.

Comments/Suggestions:

[C1],[C2],[C3] - Minor Writing Issues:

- We thank the reviewer for these suggestions. We will add Hayati et al. (EMNLP 2024) to the Related Work, and add the missing citation for the GPT model.

Official Review of Submission10831 by Reviewer JxCp

Official Review by Reviewer JxCp 📅 02 Feb 2026, 15:17 (modified: 17 Mar 2026, 08:40)

👁️ Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer JxCp, Commitment Readers

📄 Revisions (/revisions?id=ddUZpaL80W)

Paper Summary:

The authors argue that for subjective prompts, LLMs should provide a response that consists of multiple viewpoints (i.e., pluralism). However, preference alignment inherently introduces bias such that LLMs may be more aligned to a certain viewpoint while abandoning others, which marginalizes valid albeit perhaps unpopular viewpoints. However, previous approaches to evaluating LLM pluralism do not reflect the in-the-wild scenario. Thus, the authors introduce a novel evaluation framework **PluralEval**, and a multi-domain dataset of natural crowd response **WildSCOPE**.

The paper provides evidence that sycophancy inhibits pluralism in LLMs, affecting their responses (more aligned with the user's belief) and their reasoning/judgment capabilities.

Outside the already mentioned contributions, the paper also provides a new goal-oriented clustering algorithm that rivals GoalEx, although the current evaluation of this new clustering algorithm remains minimal (understandable, due to it not being the focus of the paper).

The paper concludes by stating that current preference alignment methods exacerbate sycophancy behaviour in LLMs by avoiding outputs that contradict user beliefs.

Summary Of Strengths:

1. Introduces a novel framework to evaluate LLM pluralism in open-ended generation, which may be the first one that doesn't require manual human evaluation.
2. Experiments are well-motivated and support the claims the authors made.

Summary Of Weaknesses:

1. Software/code artifacts are unavailable, which hurts applicability
2. Section 6.1 evaluates pluralism by evaluating how many view points did the model covered across k generated samples. However, in-the-wild use cases, LLMs only generate "one" response. I believe pluralism should be evaluated by **how much view points do the model cover on average, instead of coverage across all generations.**

Comments Suggestions And Typos:

Experimental Suggestions:

1. The authors concluded that LLMs avoid "confrontational" responses that diverge from the user's beliefs. It would be an additional plus if the authors tested this on open-sourced LLMs with pre-preference alignment and post-preference alignment.

Clarity Suggestions:

1. In Section 6.2.1, how is the least popular cluster determined? Is it the sum of (upvote - downvote) per summary in the cluster? Is this normalized? Not a major methodological concern, but enhancing clarity on this will be appreciated.
2. A flow chart for Experiment 6.1 will be greatly appreciated, even if it is in the appendix. There are many used terms, and the clarity becomes muddled. Personally, lines 434-437 are especially hard to understand.

Confidence: 3 = Pretty sure, but there's a chance I missed something. Although I have a good feel for this area in general, I did not carefully check the paper's details, e.g., the math or experimental design.

Soundness: 4 = Strong: This study provides sufficient support for all of its claims. Some extra experiments could be nice, but not essential.

Excitement: 4 = Exciting: I would mention this paper to others and/or make an effort to attend its presentation in a conference.

Overall Assessment: 4 = Conference: I think this paper could be accepted to an *ACL conference.

Ethical Concerns:

There are no concerns with this submission

Needs Ethics Review: No

Reproducibility: 3 = They could reproduce the results with some difficulty. The settings of parameters are underspecified or subjectively determined, and/or the training/evaluation data are not widely available.

Datasets: 3 = Potentially useful: Someone might find the new datasets useful for their work.

Software: 1 = No usable software released.

Knowledge Of Or Educated Guess At Author Identity: No

Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Knowledge Of Paper Source: N/A, I do not know anything about the paper from outside sources

Impact Of Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Reviewer Certification: I certify that the review I entered accurately reflects my assessment of the work. If you used any type of automated tool to help you craft your review, I hereby certify that its use was restricted to improving grammar and style, and the substance of the review is either my own work or the work of an acknowledged secondary reviewer.

Publication Ethics Policy Compliance: I did not use any generative AI tools for this review



Response to Reviewer JxCp

Official Comment

by Authors (Gagan Mundada (/profile?id=~Gagan_Mundada1), Bodhisattwa Prasad Majumder (/profile?id=~Bodhisattwa_Prasad_Majumder1), Zhouhang Xie (/profile?id=~Zhouhang_Xie1), Julian McAuley (/profile?id=~Julian_McAuley1), +3 more (/group/edit?id=aclweb.org/ACL/ARR/2026/January/Submission10831/Authors))

20 Feb 2026, 20:43 (modified: 17 Mar 2026, 08:40)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer JxCp, Commitment Readers

Revisions (/revisions?id=Sm7Xu5sizA)

Comment:

Thank you for the careful reading and thoughtful feedback. We are encouraged that you find PluralEval and WildSCOPE novel and that the experiments are well motivated and support our claims. We address the specific comments below.

Weakness:

[W1] Artifacts:

- We plan to release the code and relevant artifacts upon acceptance.

[W2] Single Response Setting:

- We wanted to start by clarifying that, in our setting, we prompt the LLM to generate k different plausible responses to a given query in a single overall response (as, e.g., a bullet list). For example, asking the LLM “please give me 10 potential opinions for situation X”, to simulate the scenario where the user explicitly and purposefully asks the LLM for potential crowd opinions in decision making scenarios (e.g., to understand whether he/she’s behavior is appropriate).
- To this end, LLM will only be queried once in our setting (though it was prompted for k responses), simulating in-the-wild usage, rather than being sampled k times.
- That said, we do think exploring other alternatives, such as evaluating single-response generation from LLMs, is valuable, and we leave this to future work that our proposed PluralEval framework could support. However, we believe this seems to fit future extensions, rather than a weakness of the current work.

Comments/Suggestions:

[E1] Comparative Setting:

- We agree this would be a valuable extension. Our current scope focuses on measuring the effect of belief injection and preference-aligned models on pluralism within a consistent evaluation setup. Evaluating pre-trained open-source models against post-trained open-source models is important future work that guides the significance of post-training in pluralism, and we will note this explicitly.

[S1] Comparison Clarity:

- We determine cluster popularity by aggregating Reddit vote scores. Each comment in a thread carries a vote score which transfers to the atomic opinions extracted from it. Each cluster’s popularity is the sum of scores from its constituent opinions. Since we only compare clusters within the same thread, normalization is not necessary. We will make this more explicit in the final version.

[S2] Flow Chart:

- We thank the reviewer for this suggestion. We will add a flowchart for Experiment 6.1 to the appendix and clarify the text around Lines 434 to 437.

News (/group?id=OpenReview.net/News&referrer=

Privacy Policy (/legal/privacy)

[Homepage](/))

[OpenReview \(/about\)](#) is a long-term project to advance science through improved peer review with legal nonprofit status. We gratefully acknowledge the support of the [OpenReview Sponsors \(/sponsors\)](#). © 2026 OpenReview