Benchmarking Overton Pluralism in LLMs

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce the first framework for measuring Overton pluralism in large language models–the extent to which diverse viewpoints are represented in model outputs. We (i) formalize Overton pluralism as a set-coverage metric (OvertonScore), (ii) conduct a large-scale U.S.-representative human study (N=100; 30 questions; 8 LLMs), and (iii) develop an automated benchmark that reproduces human judgments with high fidelity. Our findings show that while most models achieve comparable pluralism, Gemma 3-27B underperforms and GPT o4-mini achieves the highest OvertonScore. The automated benchmark replicates these human results and generalizes across unseen models, enabling scalable evaluation.

1 Introduction

LLMs shape political discourse, education, and everyday interactions. When they misrepresent or erase viewpoints, they risk distorting deliberation, marginalizing communities, and creating "algorithmic monoculture" [Bommasani et al., 2022, Kleinberg and Raghavan, 2021]. Current alignment strategies exacerbate this by optimizing for the "average" user [Santurkar et al., 2023, Durmus et al., 2024], collapsing genuine disagreements [Kirk et al., 2025, Sorensen et al., 2024a, Bakker et al., 2022] into a single normative stance—an issue known as *value monism* [Gabriel, 2020]. Outputs that appear neutral often encode majority or developer-preferred biases, entrenching representational harms [Chien and Danks, 2024] and heightening safety risks such as susceptibility to propaganda or cultural domination.

Pluralistic alignment offers an alternative: rather than consensus, models should represent a spectrum of reasonable perspectives within the "Overton window" of public discourse. Sorensen et al. [2024b] distinguish *Overton pluralism*, from *steerable pluralism* (users shift toward a perspective), and *distributional pluralism* (outputs reflect population distributions). We focus on Overton pluralism, the most relevant for settings with many legitimate answers. In this work, we focus on Overton pluralism, the most practically relevant for subjective settings where many reasonable answers exist.

Several modeling strategies move in this direction: MaxMin-RLHF ensures minimal group satisfaction [Chakraborty et al., 2024], Modular Pluralism adds community modules for multiple pluralism types [Feng et al., 2024], and Collective Constitutional AI sources rules from diverse publics [Huang et al., 2024]. Yet all depend on robust benchmarks to evaluate pluralistic representation.

Yet there remains no benchmark directly measuring Overton pluralism. Existing datasets provide partial proxies: PRISM [Kirk et al., 2025] and GlobalOpinionQA [Durmus et al., 2024] curate disagreement prompts but focus on preference aggregation; Value Kaleidoscope [Sorensen et al., 2024a] encodes moral principles for distributional pluralism; and Value Profiles [Sorensen et al., 2025] compress value descriptions for steerable personalization. Lake et al. [2025] proxy Overton pluralism with binary yes-no questions, but it is limited and unsuitable for benchmarking.

The closest work is Model Slant [Westwood et al., 2025], which uses pairwise comparisons of perceived political slant. However, their design is limited to bipartisan bias and cannot capture the

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

magnitude or breadth of representation. In contrast, our work estimates Overton pluralism through large-scale human judgements and develops an automated benchmark for scalable evaluation.

Contributions: We develop the first end-to-end framework measuring Overton pluralism in LLMs.

- **Operationalization:** A novel metric, OvertonScore, to quantify pluralism in models (2).
- **Large-scale human study.** Benchmarking Overton pluralism with a U.S.-representative dataset (30 questions, 8 frontier LLMs) measuring perceived representation (3,4).
- **Automated benchmark.** A framework for scalable evaluation of Overton pluralism (5). Our LLM judge OvertonScores achieve high rank correlation with human scores ($\rho \approx 0.90$) and preserves significance conclusions, showing generalization to unseen models (6).

Together, these contributions move pluralistic alignment from a normative goal to a measurable, reproducible benchmark task.

2 Operationalizing Overton Pluralism

Overton pluralism is defined at the level of a *set*: for a given query x and possible answers y, the Overton window $W(x)$ is the set of all *reasonable* answers.[1] A model $\mathcal{M}$'s response to a query x is considered Overton-pluralistic if it contains or synthesizes all answers in the Overton window $W(x)$, i.e. if $\mathcal{M}(x) = W(x)$. Therefore to *quantify* the extent to which a model response is Overton-pluralistic, we can calculate the proportion of Overton window it covers.

Concretely, for a subjective query x, if a majority of humans who believe some answer $y \in W(x)$ feel that a model response $\mathcal{M}(x)$ represents their view, then we consider y to be *covered*, denoted by $y \in \mathcal{M}(x)$. Therefore, we define Overton coverage as

$$\mathrm{OC}(\mathcal{M}, x) = \frac{1}{|W(x)|} \sum_{y \in W(x)} \mathbb{1}\{y \in \mathcal{M}(x)\}$$

The OvertonScore for a model $\mathcal{M}$ over a set of queries $X = \{x_1, \dots x_n\}$ is the average coverage:

$$\mathrm{OvertonScore}(\mathcal{M}, X) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{OC}(\mathcal{M}, x_i)$$

In practice, we conduct a human data study (Section 3) to estimate the Overton window and determine response coverage to form a benchmark (4). However, with the rapid advancement of LLMs, it is not sustainable to collect additional human ratings for each new model. To make progress as a field, we need a way to recreate these representation ratings automatically. Our goal is ultimately to demonstrate that *LLM-as-a-Judge* can accurately and fairly predict how diverse humans perceive representation in model outputs on subjective topics, thus providing a scalable automatic benchmark for Overton pluralism without requiring additional data collection (Section 5).

3 Data Collection

We recruited 100 English-speaking, US-based participants from Prolific, stratified to balance gender (50% female, 50% male) and political spectrum (30% conservative, 30% moderate, 30% liberal, 10% other). Participants were paid a fair wage ($8–$12/hour).

Each participant answered three randomly drawn questions from the 30 prompts in Westwood et al. [2025], which span politically salient domains such as healthcare, climate policy, trans rights, and free speech. These prompts target value-laden tradeoffs that cannot be resolved by factual recall alone.

For each question, participants (i) wrote a short free response (1–3 sentences), (ii) selected their stance via a multiple choice item (liberal, conservative, or neutral;[2]), and (iii) evaluated the outputs of

[1] According to Sorensen et al. [2024b], a reasonable answer is one "for which there is suggestive, but inconclusive, evidence, or one with which significant swaths of the population would agree."

[2] Full endpoints for each topic appear in Table S1 of Westwood et al. [2025].

eight state-of-the-art LLMs in randomized order. For each response they rated: "To what extent is your perspective represented?" (1 = "Not at all" to 5 = "Fully represented").

The eight evaluated LLMs span key axes of development: open vs. closed-source, reasoning vs. non-reasoning, and U.S.- vs. China-based origin. They include GPT-4.1 and o4-mini (OpenAI), Gemma 3-27B (Google), DeepSeek R1 and V3 (DeepSeek), Llama 4 Maverick and Llama 3-70B instruct (Meta), and Claude 3.7 Sonnet (Anthropic). After excluding incomplete responses and timeouts, the final dataset comprised 2,393 user–question–model datapoints.

4 Benchmark Design

In Section 2 we defined the OvertonScore of a model as the average proportion of the Overton window it covers. Calculating this in practice requires (1) identifying *distinct* responses and (2) testing whether a model output covers them in natural language. We approximate distinct answers by clustering, and count a cluster as covered if its average human representation rating is at least 4 (mostly represented) out of 5 (fully represented). In this work, we cluster answers by each user's selected topic stance (Conservative, Neutral, Liberal) as designed in Westwood et al. [2025]. Additional clustering approaches and discussion on their tradeoffs can be found in Appendix B.

4.1 Human Benchmark Results

We analyzed representation ratings with OLS regression, testing whether each model's coverage probability (stance-based clustering) differed from the overall mean. Regressions controlled for question difficulty via fixed effects, with standard errors clustered by question.

Table 1 reports the adjusted OvertonScores (predicted probability of covering a stance cluster, averaged over questions) and each model's deviation from the mean. We find that **o4-mini** achieves significantly higher coverage than average ($+0.07$, $p < 0.01$), while **Gemma 3-27B** is significantly lower (-0.09, $p < 0.05$). The remaining six models do not differ significantly from the average.

5 Automated Benchmarking with LLM Judges

To scale evaluation beyond costly human studies, we test whether LLMs can act as judges of pluralism. The task is to predict a human's perceived representation score (Likert 1–5) for a given model output. We evaluate several prompting variants for the judges: demographics only; free response (FR) plus topic stance; demographics + FR + stance ("Full Profile"); and few-shot prompts using a user's ratings of the other seven responses to the same question.

Judge models are compared against three baselines. The *human baseline* asks another annotator to predict ratings for 300 datapoints given a full profile (no example ratings). The *semantic similarity* baseline selects the closest among the seven other responses to the same question,[3] and assigns its rating. Finally, the *mean-of-others* baseline uses the average of the user's ratings for the other seven responses, rounded to the nearest integer to match the 1–5 scale.[4]

We test GPT-4.1 mini and nano, Gemini Flash, and Gemini Pro as judges under each prompting variant. We predict ratings for all datapoints three times for each configuration and evaluate using the (rounded) average prediction. All experiments were run on CPUs and models accessed via APIs.

6 Benchmark Evaluation

We evaluate judges by accuracy (exact rating), mean absolute error (MAE), mean squared error (MSE), and win-rate (proportion of datapoints with lower error than a baseline). We report 95% CIs via nonparametric bootstrap. Additional ablations appear in Appendix C.

Gemini Pro with Few-Shot is the strongest judge, achieving 59% accuracy. It significantly outperforms the human baseline and profile prompts and matches semantic similarity (56%). Trends hold

[3] Calculated using cosine similarity of response embeddings from OpenAI's `text-embedding-3-large`

[4] Rounding ensures predictions are valid Likert values.

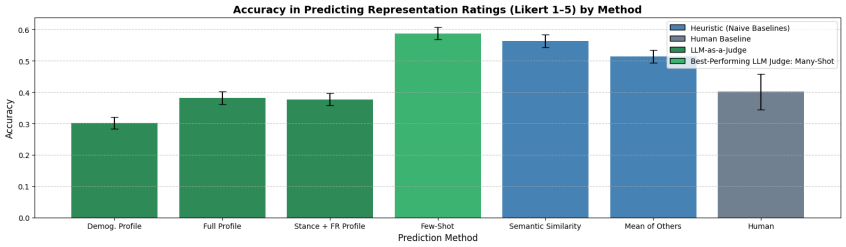

Figure 1: Accuracy of judge LLMs and baselines across prompting conditions (95% CIs).

for MAE and MSE (Figure 2). In terms of win rate, we find again that Gemini Pro with Few-Shot is strongest, winning $> 50\%$ of the time (average 66.12%) against all other methods (Figure 3).

6.1 Subgroup Parity & Generalization

We tested for subgroup disparities using nonparametric permutation ANOVA across demographic and stance categories and found no meaningful fairness issues. Full results are in Appendix D.

To test whether our benchmark generalizes to unseen models, we ran a leave-one-model-out (LOMO) analysis: for each target LLM, we replaced its human ratings with best LLM-judge predictions (Gemini Pro with Few-Shot) and re-ran the stance-based OLS regressions.

Across models, the agreement between the human-only regressions and the judge-substituted regressions was very strong. Rank correlations between human and judge OvertonScores averaged $\rho \approx 0.90$ (Spearman), and the correlation of model coefficients was similarly high ($r \approx 0.89$) with very small mean errors (≈ 0.01). Directional agreement exceeded 90%, indicating the judge consistently preserves which models are above or below the mean. In terms of statistical conclusions, **o4-mini** replicated as significantly above average, while **Gemma 3-27B** did not replicate as significantly below average; the remaining six models all remained non-significant, as in the human-collected data. Taken together, these results show that the judge-based benchmark largely preserves human judgments of pluralistic coverage, with the main discrepancy being that the judge over-rates Gemma compared to human participants.

7 Discussion & Limitations

Our benchmark offers a first framework for quantifying Overton pluralism in LLMs, but several limitations remain. Model-level OvertonScores are defined with respect to the 30 questions in our study, which can be easily broadened to additional topics in future work by simply extending the LLM Judge predictions or collecting additional data. Estimates also depend on how distinct answers are clustered: we report stance-based clustering in this work, while alternative variants are discussed in Appendix B. In addition, our data come from U.S.-based English speakers, and Overton windows are culturally situated; expanding to more diverse global populations is an important direction for future work. Finally, LLM judges approximate but do not perfectly replicate human ratings, as seen with Gemma, and they may inherit biases of the underlying models. We view the present benchmark as a starting point for a cycle of iterative improvement, where pluralism metrics guide pluralistic model development, which in turn enables better benchmarking.

8 Conclusion

We introduce OvertonScore as a principled metric of Overton pluralistic alignment, create a large-scale human dataset across 30 salient questions and 8 LLMs, and validate the first automated benchmark using LLM-as-a-Judge. Human data show OpenAI's o4-mini achieves significantly higher OvertonScores, while Gemma 3-27B scores lower. Automated evaluation with Gemini Pro reproduces these patterns with high correlation to human scores and no major subgroup disparities. By turning pluralistic alignment from a normative aim into a measurable benchmark, our work establishes a foundation for systematic progress. We hope the dataset and public benchmark released alongside this paper foster community engagement and the development of increasingly pluralistic LLMs.

References

Michiel Bakker, Martin Chadwick, Hannah Sheahan, Michael Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matt Botvinick, and Christopher Summerfield. Fine-tuning language models to find agreement among humans with diverse preferences. *Advances in Neural Information Processing Systems*, 35:38176–38189, December 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/f978c8f3b5f399cae464e85f72e28503-Abstract-Conference.html.

Rishi Bommasani, Kathleen A. Creel, Ananya Kumar, Dan Jurafsky, and Percy S. Liang. Picking on the Same Person: Does Algorithmic Monoculture lead to Outcome Homogenization? *Advances in Neural Information Processing Systems*, 35:3663–3678, December 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/17a234c91f746d9625a75cf8a8731ee2-Abstract-Conference.html.

Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Dinesh Manocha, Furong Huang, Amrit Bedi, and Mengdi Wang. MaxMin-RLHF: Alignment with Diverse Human Preferences. In *Proceedings of the 41st International Conference on Machine Learning*, pages 6116–6135. PMLR, July 2024. URL https://proceedings.mlr.press/v235/chakraborty24b.html. ISSN: 2640-3498.

Jennifer Chien and David Danks. Beyond Behaviorist Representational Harms: A Plan for Measurement and Mitigation. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pages 933–946, New York, NY, USA, June 2024. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3658946. URL https://dl.acm.org/doi/10.1145/3630106.3658946.

Esin Durmus, Karina Nguyen, Thomas Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. Towards Measuring the Representation of Subjective Global Opinions in Language Models. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=zl16jLb91v.

Shangbin Feng, Taylor Sorensen, Yuhan Liu, Jillian Fisher, Chan Young Park, Yejin Choi, and Yulia Tsvetkov. Modular Pluralism: Pluralistic Alignment via Multi-LLM Collaboration. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4151–4171, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.240. URL https://aclanthology.org/2024.emnlp-main.240/.

Iason Gabriel. Artificial Intelligence, Values, and Alignment. *Minds and Machines*, 30(3):411–437, September 2020. ISSN 1572-8641. doi: 10.1007/s11023-020-09539-2. URL https://doi.org/10.1007/s11023-020-09539-2.

Saffron Huang, Divya Siddarth, Liane Lovitt, Thomas I. Liao, Esin Durmus, Alex Tamkin, and Deep Ganguli. Collective Constitutional AI: Aligning a Language Model with Public Input. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pages 1395–1417, New York, NY, USA, June 2024. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3658979. URL https://dl.acm.org/doi/10.1145/3630106.3658979.

Hannah Rose Kirk, Alexander Whitefield, Paul Rottger, Andrew M. Bean, Katerina Margatina, Rafael Mosquera-Gomez, Juan Ciro, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott Hale. The PRISM Alignment Dataset: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models. *Advances in Neural Information Processing Systems*, 37:105236–105344, January 2025. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/be2e1b68b44f2419e19f6c35a1b8cf35-Abstract-Datasets_and_Benchmarks_Track.html.

Jon Kleinberg and Manish Raghavan. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22):e2018340118, June 2021. doi: 10.1073/pnas.2018340118. URL https://www.pnas.org/doi/full/10.1073/pnas.2018340118. Publisher: Proceedings of the National Academy of Sciences.

Thom Lake, Eunsol Choi, and Greg Durrett. From Distributional to Overton Pluralism: Investigating Large Language Model Alignment. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *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)*, pages 6794–6814, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 9798891761896. doi: 10.18653/v1/2025.naacl-long.346. URL https://aclanthology.org/2025.naacl-long.346/.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose Opinions Do Language Models Reflect? In *Proceedings of the 40th International Conference on Machine Learning*, pages 29971–30004. PMLR, July 2023. URL https://proceedings.mlr.press/v202/santurkar23a.html. ISSN: 2640-3498.

Christopher Small, Michael Bjorkegren, Timo Erkkilä, Lynette Shaw, and Colin Megill. Polis: Scaling Deliberation by Mapping High Dimensional Opinion Spaces. *RECERCA. Revista de Pensament i Anàlisi*, 26(2), July 2021. ISSN 2254-4135, 1130-6149. doi: 10.6035/recerca.5516. URL https://www.e-revistes.uji.es/index.php/recerca/article/view/5516. Publisher: Universitat Jaume I.

Taylor Sorensen, Liwei Jiang, Jena D. Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, Maarten Sap, John Tasioulas, and Yejin Choi. Value Kaleidoscope: Engaging AI with Pluralistic Human Values, Rights, and Duties. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(18):19937–19947, March 2024a. doi: 10.1609/aaai.v38i18.29970. URL https://ojs.aaai.org/index.php/AAAI/article/view/29970. Section: AAAI Technical Track on Philosophy and Ethics of AI.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. Position: a roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML'24*, pages 46280–46302, Vienna, Austria, July 2024b. JMLR.org.

Taylor Sorensen, Pushkar Mishra, Roma Patel, Michael Henry Tessler, Michiel Bakker, Georgina Evans, Iason Gabriel, Noah Goodman, and Verena Rieser. Value Profiles for Encoding Human Variation, March 2025. URL http://arxiv.org/abs/2503.15484. arXiv:2503.15484 [cs].

Sean J Westwood, Justin Grimmer, and Andrew B Hall. Measuring Perceived Slant in Large Language Models Through User Evaluations, May 2025. URL https://modelslant.com/paper.pdf.

A Full Human Benchmark Results

Table 1 reports the detailed OLS estimates of OvertonScores under stance clustering for the human data.

B Clustering

There are several approaches to clustering we considered, but ultimately chose the stance clusters to present in the main paper. Each has their own benefits and limitations, which we discuss below. Regardless, both our metrics and automated benchmark setup are flexible with respect to the clustering method chosen, allowing future work to build upon and improve on our estimation of Overton pluralism!

Model	Adjusted Coverage	Coef vs. Mean	p-value
o4-mini	**0.526**	**+0.072**	**0.003**
Llama 3-70B instruct	0.513	+0.059	0.149
Claude 3.7 Sonnet	0.474	+0.021	0.630
DeepSeek V3	0.462	+0.008	0.818
Llama 4 Maverick	0.449	–0.005	0.888
GPT-4.1	0.436	–0.018	0.562
DeepSeek R1	0.410	–0.043	0.223
Gemma 3-27B	**0.359**	**–0.095**	**0.022**

Table 1: OLS estimates of OvertonScores under stance clustering. Adjusted coverage is the marginal predicted probability of a stance cluster being covered. Coefficients are differences relative to the grand mean (question fixed effects included; SEs clustered by question). Significant deviations are shown in **bold**.

B.1 Individual Clusters

Each user is treated as a unique cluster.

$$W(x) = \{\{u\} : u \in \text{users for question } x\}$$

The benefits of treating all users individually is that it is maximally strict: every user's perception is its own "view," guaranteeing that minority views are flattened into majority views. However, this results in an overly fragmented estimation of the Overton window. By nature of surveying a population, majority perspectives will occur multiple times in the dataset, resulting in an over-weighting of the majority views when calculating the Overton coverage. Therefore, LLM responses that cover popular views can look *inflated* (since many users share that view). Conversely, models that cover rare distinct positions but not the mainstream can look *deflated*. In sum, treating each user as their own individual distinct response results in a weighted score. Since this deviates from the definition of Overton pluralism, we chose not to pursue this in the scope of this work.

B.2 Stance Clusters

Users are grouped only by their selected stance: Conservative, Neutral, Liberal. This is the method we use in our main paper.

$$W(x) = \{C_{\text{Cons}}, C_{\text{Neut}}, C_{\text{Lib}}\}$$

The limitations of this method are the potential risk of flattening minority views within the same general stance group. This may lead to an overestimation of coverage if there exists some minority subgroup. While less likely, it can also result in underestimation in the case that the minority subgroup(s) pull down the average representation rating below 4, causing the entire group to be considered not covered.

However, the benfits are that it is incredibly simple and straightforward to calculate. Moreover, in the context of the questions we use in this work, they are designed to more or less have 3 distinct answers.

B.3 Stance + Minority Split

Similar to the stance clusters, but we try to split off minority subgroups, which are detected by consistently diverging representation scores. Concretely, for each target model m, clusters are defined using ratings from all other models $j \neq m$. If a group of users is consistently $\geq \delta$ Likert points above/below stance medians for at least 4 of 7 other models, they form a minority cluster.

The benefits to this approach is that it can correct some of the overly coarse merging and increase the resolution of the Overton window estimation. However, the thresholding hyperparameter δ is difficult to tune without additional validation, and the coverage metric is potentially very sensitive to this because we are generally working with a small number of clusters. If too many splits are created,

this will deflate the score, whereas if it is too relaxed, it will not create any splits and result in the same as the original stance cluster approach.

B.4 NLP-based clustering

Another approach would be to use NLP methods such as semantic embeddings or entailment (NLI) to cluster the humans' free responses into distinct groups. However, this is tricky to tune and validate without additional human validation. While it is out of the scope of this work, we believe it is probably the best route forward in the future!

B.5 Crowdsourced clustering

It would be ideal if the humans themselves could determine which distinct response category fits them best. Platforms such as Pol.is [Small et al., 2021] allow for users to vote on other's responses, and clustering is done automatically based on similar vote patterns. In other words, participants who hold the same view will likely agree/disagree with similar statements, and therefore should be clustered.

However, this is potentially expensive and / or logistically complex to carry out in practice due to the additional time taken to pay the crowd workers as well as coordinate a second round of collection with the same workers to do the voting.

C Detailed LLM Judge Results

Experiment Design. Our prompting experiments are structured to systematically reducing or perturbing the full user profile $\mathbf{u}$, *demog_select_freereponse* (composed of demographics, topic/selection stance, and free-text response) by default, to identify which components contribute most to accurate and fair prediction of user's perceived representation $\hat{y}$ relative to ground truth y. We evaluate this in both zero-shot and few-shot setups. For zero-shot, i.e., general-user-profile-only setups, we designed ablations and perturbations as follows:

- *null*: user profile information fully ablated, which means the only input would be prompt instruction $\mathbf{p}$ and LLM's response $\mathbf{r}$ to be evaluated.

$$\hat{y}_{\text{null}} = P(y \mid \mathbf{r}, \mathbf{p})$$

- *demog*: include only demographic information $\mathbf{u}_{\text{demo}}$, i.e., sex, ethnicity, and political identity.

$$\hat{y}_{\text{demo}} = P(y \mid \mathbf{u}_{\text{demo}}, \mathbf{r}, \mathbf{p})$$

- *freeresponse*: include only the free-text response $\mathbf{u}_{\text{freeresponse}}$.

$$\hat{y}_{\text{resp}} = P(y \mid \mathbf{u}_{\text{freeresponse}}, \mathbf{r}, \mathbf{p})$$

- *random_full* and *cluster_full*: perturbation setups that substitute the user profile with a random profile $\mathbf{u}'$ or cluster-centroid/majority profile $\mathbf{u}_{\text{centroid},k}$.

$$\hat{y}_{\text{random}} = P(y \mid \mathbf{u}', \mathbf{r}, \mathbf{p}), \quad \mathbf{u}' \sim \mathcal{U}(\mathcal{U}_{\text{train}})$$

  or

$$\hat{y}_{\text{cluster}} = P(y \mid \mathbf{u}_{\text{centroid},k}, \mathbf{r}, \mathbf{p})$$

Under few-shot setups, the LLM judges are additionally provided with the user's answers to other question (QAs), as well as their ratings of other model responses to the same question (Ratings). We again ablated the original profile fields along with these example fields through the prompt variants below:

- *demog_select_freeresponse_ms_**: few-shot variants with full user profile.
- *freeresponse_ms_**: few-shot variants with the user's free-text response only.
- *ms_**: general user profile has been ablated entirely, leaving only example fields as the "pseudo-profile."

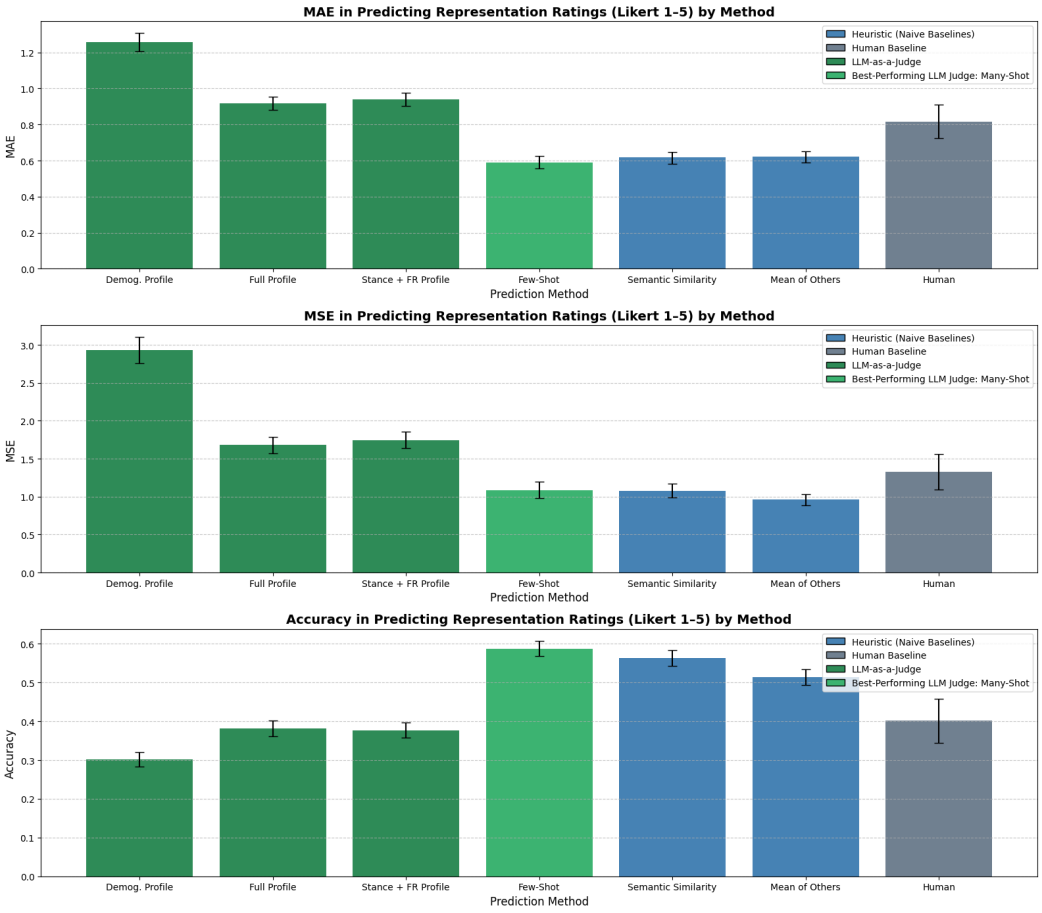

Figure 2: Average accuracy, MAE, and MSE among baselines and Gemini Pro LLM judge across prompting methods. The Few-Shot method generally outperforms all other methods across metrics except the Semantic Similarity. Higher accuracy and lower MAE/MSE is considered better. The error bars are 95% confidence intervals estimated via bootstrapping.

This design allows us the probe into the marginal contribution of demographics, per-topic stance, and free-text response of the user to LLM Judges' performances; it also indicates whether few-shot contextualization can substitute for or amplify these user-specific information fields.

Table 2: Detailed LLM-as-a-Judge Results

Prompt Variant	Metric	gpt-4.1-mini	gpt-4.1-nano	**gemini-2.5-pro**	gemini-2.5-flash
null	Accuracy	0.316	0.276	**0.247**	0.276
	MAE	1.180	0.967	**1.322**	1.087
	MSE	2.612	1.594	**2.989**	2.145
demog	Accuracy	0.256	0.280	**0.219**	0.281
	MAE	1.100	0.936	**1.381**	0.966
	MSE	2.012	1.474	**3.121**	1.584
freeresponse	Accuracy	0.344	0.268	**0.348**	0.336
	MAE	0.944	1.029	**1.053**	0.937
	MSE	1.624	1.747	**2.105**	1.611
demog_select _freeresponse	Accuracy	0.348	0.268	**0.344**	0.384

Continued on next page

Table 2 – continued from previous page

Prompt Variant	Metric	gpt-4.1-mini	gpt-4.1-nano	**gemini-2.5-pro**	gemini-2.5-flash
	MAE	0.948	1.032	**0.972**	0.872
	MSE	1.668	1.748	**1.772**	1.449
random_full	Accuracy	0.348	0.244	**0.359**	0.369
	MAE	0.948	1.035	**0.955**	0.876
	MSE	1.676	1.743	**1.747**	1.454
cluster_full	Accuracy	0.352	0.252	**0.321**	0.392
	MAE	0.940	1.032	**1.008**	0.852
	MSE	1.652	1.738	**1.837**	1.401
demog_select _freeresponse_ms _qas_ratings	Accuracy	0.408	0.357	**0.579**	0.506
	MAE	0.856	0.912	**0.579**	0.663
	MSE	1.552	1.627	**0.979**	1.056
demog_select _freeresponse_ms _qas_ratings2	Accuracy	0.384	0.284	**0.498**	0.456
	MAE	0.884	1.084	**0.733**	0.810
	MSE	1.588	2.124	**1.348**	1.464
demog_select _freeresponse_ms _ratings	Accuracy	0.396	0.324	**0.574**	0.544
	MAE	0.824	0.972	**0.591**	0.636
	MSE	1.400	1.764	**1.017**	1.060
demog_select _freeresponse_ms _ratings2	Accuracy	0.372	0.280	**0.483**	0.468
	MAE	0.908	1.084	**0.780**	0.808
	MSE	1.628	2.108	**1.493**	1.496
freeresponse_ms_ ratings	Accuracy	0.420	0.352	**0.539**	0.536
	MAE	0.804	0.892	**0.643**	0.644
	MSE	1.332	1.580	**1.108**	1.092
ms_ratings	Accuracy	0.588	0.396	**0.588**	0.576
	MAE	0.544	0.784	**0.592**	0.564
	MSE	0.864	1.280	**1.080**	0.916

Full Pilot Results across Prompts and Models. We first ran the prompt grid on **250 rows** to control the time and cost while stress-testing design choices. The results already show systematic differences across both models and prompt types: the dominance of *ms_** over all zero-shot prompts and over profile-augmented few-shot prompts. We selected Gemini-2.5-Pro for scaling to the full data since it demonstrates the strongest predictive fidelity, with a consistently high accuracy and substantially smaller MAE and MSE relative to alternatives in few-shot setups in particular.

Zero-Shot Results and Analysis. Overall, the results align closely with the logic of our prompt setups. Removing the user profile entirely (*null*) leads to a collapse in accuracy and inflated errors across all models, suggesting that some form of user information input is necessary. Similarly, all models performs poorly under demographic-only input (*demog*), with the best accuracy by Gemini-2.5-Flash reaching only 0.281, which indicates that demographics cannot by themselves capture fine-grained perceptions of representation in LLM-as-a-Judge framework. By contrast, the free-text-response-only prompt (*freeresponse*) produces better results of accuracy around 0.34 across models, suggesting that written opinion might be a more informative single channel of the user profile.

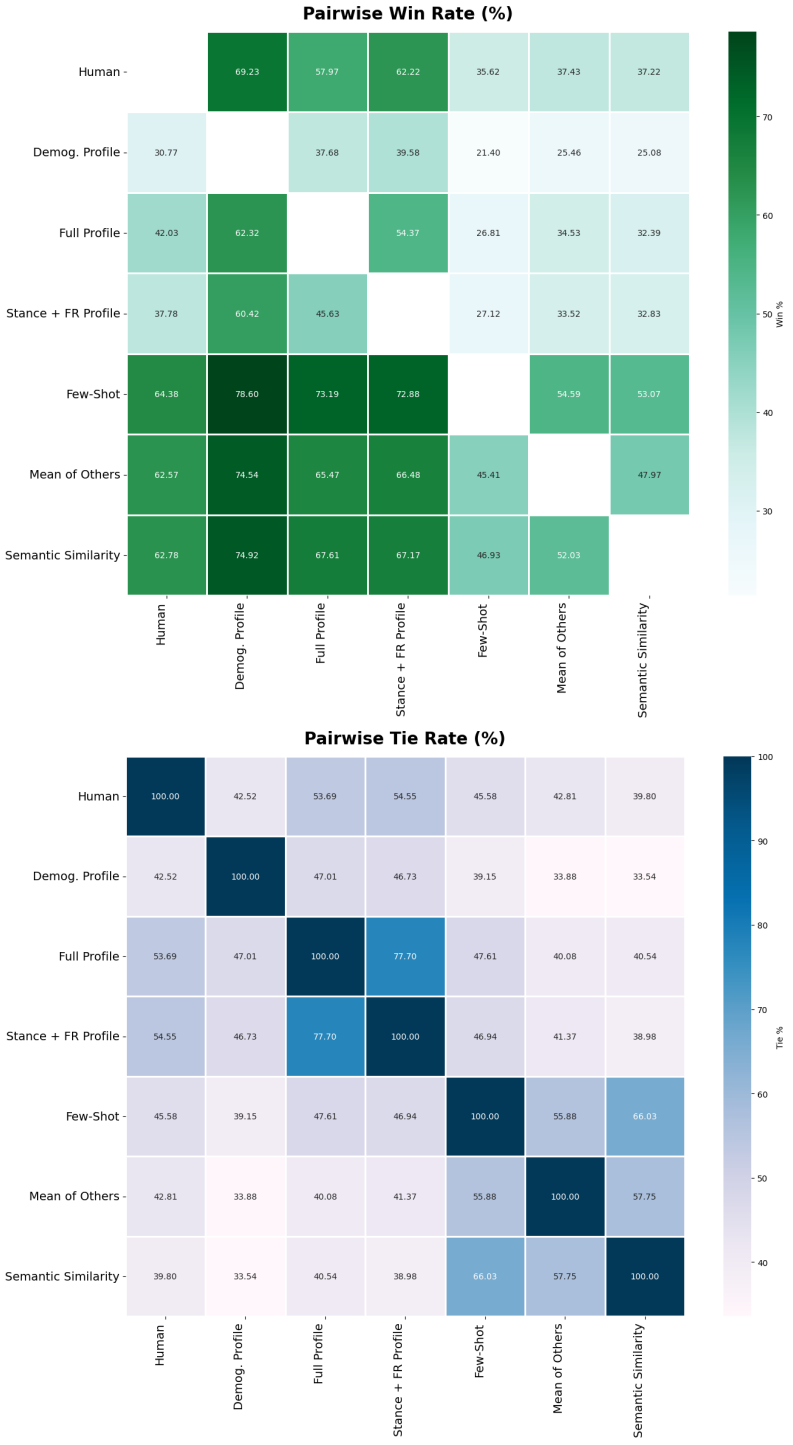

Figure 3: Win and tie rates for each method. To interpret the results, the win rate is the proportion of the time the method in the row "beats" the method in the column by having a strictly smaller prediction error, excluding ties. For example, Few-Shot has a closer prediction than the Human baseline 64.38% of the time, and ties (equal error) 45.58% of the time.

Interestingly, the perturbation prompts (*random_full* and *cluter_full*) do not degrade performance as sharply as one might expect if user profiles were highly informative. Substituting a random profile (*random_full*) yields results around the same as those of full accurate user profile (*demog_select_freeresponse*), and substituting a cluster-centroid/majority profile (*cluster_full*) has even slightly better performance than using the authentic profile fields (e.g., Gemini-2.5-Flash reached 0.392 accuracy with *cluster_full*, outperforming free-response and full-profile variants when yielding the lowest zero-shot MAE of around 0.940 in the meantime).

Few-Shot Results and Analysis. The few-shot setting amplifies these trends. Notably, profile-ablated *ms_** prompts deliver the best performance–for Gemini-2.5-Pro, *ms_qas_ratings* attains the highest accuracy (0.588), and low MAE/MSE (0.592/1.080). The same performance pattern appears in other models as well. The gaps over single-shot reductions are large and mirrored across MAE/MSE, indicating a practically significant effect rather than statistical noise.

Interpreted through our experiment design, this says the users' rating examples are an exceptionally information-dense surrogate for the user and likely function as a calibration signature of how the user maps content to perceived representation on certain topics. Crucially, adding profile contexts to the few-shot setups (*demog_select_freeresponse_ms_** and *freeresponse_ms_**) does **not** help and can slightly hurt the performance relative to *ms_** alone. For Gemini-2.5-Pro, *demog_select_freeresponse_ms_ratings* trails *ms_ratings* on accuracy, suggesting overconditioning or source-weighting mismatch.

We acknowledge the trade-offs in the 250-row pilot experiment: subgroup coverage is uneven, confidence intervals are wider, and ranking among close prompt variants can wobble. To reduce overfitting to the pilot subset, we **re-ran only the strongest set of prompt variants on the full dataset**, anchoring our most significant conclusions in a larger, more representative sample.

D Subgroup Parity Checks

To assess subgroup parity, we conducted nonparametric permutation ANOVA tests (5,000 permutations) for each category (sex, ethnicity, political spectrum, selection position, and model) and each metric (Accuracy, MAE, MSE). This approach tests whether group means differ overall, without relying on normality assumptions. Results are summarized in Table 3.

We find no evidence of disparities across the eight target models on accuracy or MAE (all $p > 0.40$), and only a borderline effect for MSE ($p = 0.055$). By contrast, several participant characteristics show significant differences. Accuracy varies significantly across ethnic groups ($p = 0.014$), political identities ($p = 0.015$), and between male and female participants ($p = 0.025$). For sex, both error magnitude metrics (MAE, MSE) are also significant. The largest and most consistent effects occur for question stance: accuracy, MAE, and MSE all show significant variation across liberal, neutral, and conservative participants ($p < 0.01$).

Importantly, effect sizes are uniformly small ($\eta^2 < 0.01$ in all cases). Thus, while statistically detectable subgroup differences exist—especially by stance—the magnitude of disparities in judge performance is marginal. These results suggest that the LLM-as-a-judge benchmark does not exhibit large systematic fairness issues, but that participant stance and certain demographics can introduce subtle variation.

Category	Metric	F	p_{perm}	η^2	# Groups
Ethnicity	Accuracy	**3.10**	**0.014**	**0.0053**	5
	MAE	1.11	0.338	0.0019	5
	MSE	0.74	0.553	0.0013	5
Political spectrum	Accuracy	**3.49**	**0.015**	**0.0044**	4
	MAE	0.33	0.799	0.0004	4
	MSE	0.79	0.508	0.0010	4
Sex	Accuracy	**5.42**	**0.025**	**0.0023**	2
	MAE	**9.12**	**0.003**	**0.0039**	2
	MSE	**7.77**	**0.007**	**0.0033**	2
Model	Accuracy	0.98	0.443	0.0029	8
	MAE	1.00	0.432	0.0030	8
	MSE	1.95	0.055	0.0058	8
Stance (selection)	Accuracy	**7.10**	**0.001**	**0.0060**	3
	MAE	**8.10**	**0.001**	**0.0069**	3
	MSE	**4.85**	**0.007**	**0.0041**	3

Table 3: Permutation ANOVA results for subgroup fairness checks. Significant results ($p_{\text{perm}} < .05$) are bolded. Effect sizes (η^2) are small in all cases ($< .01$).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes, the claims for the operationalization of the metric are in Section 2, which are empirically carried out in Section 3 and then analyzed with appropriate statistics in Section 4. We then show our automated benchmark performance evaluation in Section 6 and all additional ablations in the appendix.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes, see Section 7.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings,

model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.

- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: We don't have theoretical *results*.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We include all experimental details in the paper and plan to release our code publicly via GitHub. We unfortunately can't submit it to the workshop directly but are happy to share it with reviewers upon request!

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.

- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Yes, we plan to release it all publicly after the double blind reviewing is over!

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We include all details in the paper and plan to release our code publicly via GitHub. We unfortunately can't submit it to the workshop directly but are happy to share it with reviewers upon request!

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We perform several different types of statistical analyses and report 95% bootstrapped confidence intervals on all figures, and include all detailed effect sizes and information in the appendix for the generalization results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We only used CPUs and minimal storage. This is mentioned in Section 5

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes, we conform.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Yes, our work is extremely motivated by societal impacts and we discuss the positive intended impact of our benchmark. We do not anticipate any misuse or negative unintended impacts.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA] .

    Justification: We do not believe our paper poses such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We faithfully cite and credit all data used in this work, which is publicly available under the Creative Commons Attribution-NonCommercial license.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We cannot upload the zip file for the workshop submission but are happy to do so if reviewers require!

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [No]

Justification: We paid our workers a fair wage, ranging from $8-$12/hr. The full instructions are omitted for brevity but all the main important phrasings are included in the main paper. We are happy to include it if reviewers request! Moreover, the full data collection pipeline (including all instructions text) will be released publicly with the codebase after the double blind review period is over.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [Yes]

Justification: There are no potential risks, and so we are exempt from our institution's IRB.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We describe the LLM usage in the methodology for the automated benchmark in Section 5.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.