Can AI Change Your View? Evidence from a Large-Scale Online Field Experiment

Keywords: Large Language Models, Online Experiments, Personalization, Persuasion, Reddit

Extended Abstract

Motivation. Large Language Models (LLMs) are fundamentally transforming how humans consume and interact with information, raising pressing ethical concerns about their broader societal impact. Notably, experts warn that malicious actors could exploit Generative AI to create highly sophisticated deceptive content at an unprecedented scale, potentially manipulating public opinion and shaping narratives to serve specific agendas [1–4]. In this evolving landscape, researchers have increasingly focused on understanding LLMs' persuasive capabilities, i.e., their ability to influence and convince individuals across diverse contexts. Early studies on AI-driven persuasion have shown that LLMs can match human performance [5–9] or even surpass it [10–12], including when dealing with highly divisive sociopolitical issues. Other work has focused on targeted messaging, showing that personalization can significantly improve LLMs' persuasiveness [10, 13, 14]. Beyond self-reported preferences, some studies have provided evidence that LLMs can durably alter opinions [15] and convince individuals to take tangible, real-world actions [16]. Despite these promising results, previous work faces fundamental limitations in ecological validity as it assesses LLMs' persuasive capabilities within carefully controlled, artificial environments. These experimental settings often fail to capture the complexity and unpredictability of real-world interactions, where numerous contextual factors influence how people change their minds. Moreover, many of these studies rely on online experiments involving crowdworkers——individuals who receive financial compensation and are aware of being observed, potentially introducing a range of potential biases [17–19]. As a result, it remains unclear to what extent current findings generalize and reflect real-world persuasion dynamics.

Present work. In this pre-registered study, we conduct the first large-scale field experiment on LLMs' persuasiveness, carried out within *r/ChangeMyView*, a Reddit community of almost 4M users and ranking among the top 1% of subreddits by size. In *r/ChangeMyView*, users share opinions on various topics, challenging others to change their perspectives by presenting arguments and counterpoints while engaging in a civil conversation. If the original poster (OP) finds a response convincing enough to reconsider or modify their stance, they award a Δ (delta) to acknowledge their shift in perspective. A visual summary of this process is provided in Figure 1.

Experimental setup. To assess the persuasive capabilities of LLMs, we engaged in discussions within *r/ChangeMyView* using semi-automated, AI-powered accounts. Each post published during our intervention was randomly assigned to one of three treatment conditions:

- **Generic**: LLMs received only the post's title and body text.
- **Personalization**: In addition to the post's content, LLMs were provided with personal attributes of the OP (gender, age, ethnicity, location, and political orientation), as inferred from their posting history using another LLM.

- **Community Aligned**: To ensure alignment with the community's writing style and implicit norms, responses were generated by a fine-tuned model trained with comments that received a Δ in posts published before the experiment.

A complete overview of our posting pipeline is presented in Figure 2. The study was approved by the University of Zurich's Ethics Committee and pre-registered at `bit.ly/4gJJfn9`. Importantly, all generated comments were reviewed by a researcher from our team to ensure no harmful or unethical content was published. Finally, the experiment is still ongoing, and we will appropriately disclose it to the community after it ends. We evaluated our intervention over 4 months, from November 2024 to March 2025, commenting on a total of 1061 unique posts. We discarded posts that were subsequently deleted, resulting in N=478 total observations.

Summary of Results. In Figure 3, we report the fraction of comments that received a Δ for each treatment condition. Notably, all our treatments surpass human performance substantially, achieving persuasive rates between three and six times higher than the human baseline. In particular, *Personalization* demonstrates a persuasive rate of 0.18 (95% CI [0.13, 0.25]), closely followed by the *Generic* condition at 0.17 ([0.12, 0.23]). *Community Aligned* trails slightly behind at 0.09 ([0.05, 0.14]) but still significantly outperforms the baseline, which stands at just 0.03 ([0.02, 0.03]). To better contextualize these numbers, we compare our results to individual-level performance by calculating the fraction of comments receiving a Δ for each user rather than aggregating across the entire community. Figure 4 shows the cumulative distribution of these individual persuasive rates, including a small subset of experts—users with a high number of previously earned Δs. Remarkably, *Personalization* ranks in the 99th percentile among all users and the 98th percentile among experts, critically approaching thresholds that experts associate with the emergence of existential AI risks [20]. Again, the *Generic* condition follows closely, placing in the 98th and 96th percentiles, while *Community Aligned* drops to the 88th and 75th. Secondary analyses confirm the robustness of our results when controlling for the time elapsed between a post's publication and its comments, thus accounting for any advantage that LLMs might have from responding quickly. Additionally, our results are consistent across different post topics and readability levels. Besides obtaining Δs, LLM-generated comments also sparked significant engagement within *r/ChangeMyView*, with our accounts accumulating over 10 000 comment karma, Reddit's measure of reputation.

Implications. In a first field experiment on AI-driven persuasion, we demonstrate that LLMs can be highly persuasive in real-world contexts, surpassing all previously known benchmarks of human persuasiveness. While persuasive capabilities can be leveraged to promote socially desirable outcomes [11, 15], their effectiveness also opens the door to misuse, potentially enabling malicious actors to sway public opinion [12] or orchestrate election interference campaigns [21]. Incidentally, our experiment confirms the challenge of distinguishing human- from AI-generated content [22–24]. Throughout our intervention, users of *r/ChangeMyView* never raised concerns that AI might have generated the comments posted by our accounts. This hints at the potential effectiveness of AI-powered botnets [25], which could seamlessly blend into online communities. Given these risks, we argue that online platforms must proactively develop and implement robust detection mechanisms, content verification protocols, and transparency measures to prevent the spread of AI-generated manipulation.

References

[1] Yoshua Bengio et al. *International AI Safety Report*. 2025. arXiv: 2501.17805 [cs.CY].

[2] Christian Tarsney. "Deception and manipulation in generative AI". In: *Philosophical Studies* (Jan. 2025). ISSN: 1573-0883. DOI: 10.1007/s11098-024-02259-8.

[3] Kokil Jaidka et al. "Misinformation, Disinformation, and Generative AI: Implications for Perception and Policy". In: *Digit. Gov.: Res. Pract.* 6.1 (Feb. 2025). DOI: 10.1145/3689372. URL: https://doi.org/10.1145/3689372.

[4] Christopher Summerfield et al. *How will advanced AI systems impact democracy?* 2024. arXiv: 2409.06729 [cs.CY]. URL: https://arxiv.org/abs/2409.06729.

[5] Hui Bai et al. *Artificial Intelligence Can Persuade Humans on Political Issues*. Feb. 2023.

[6] Alexis Palmer and Arthur Spirling. "Large Language Models Can Argue in Convincing Ways About Politics, But Humans Dislike AI Authors: implications for Governance". In: *Political Science* 75.3 (Sept. 2023), pp. 281–291. ISSN: 2041-0611. DOI: 10.1080/00323187.2024.2335471.

[7] Kobi Hackenburg et al. *Evidence of a log scaling law for political persuasion with large language models*. 2024. URL: https://arxiv.org/abs/2406.14508.

[8] Kobi Hackenburg et al. "Comparing the persuasiveness of role-playing large language models and human experts on polarized U.S. political issues". In: *OSF preprint* (Dec. 2023).

[9] Esin Durmus et al. *Measuring the Persuasiveness of Language Models*. Apr. 9, 2024. URL: https://www.anthropic.com/news/measuring-model-persuasiveness.

[10] Francesco Salvi et al. *On the Conversational Persuasiveness of Large Language Models: A Randomized Controlled Trial*. 2024. arXiv: 2403.14380.

[11] Elise Karinshak et al. "Working With AI to Persuade: Examining a Large Language Model's Ability to Generate Pro-Vaccination Messages". In: *Proc. ACM Hum.-Comput. Interact.* 7.CSCW1 (Apr. 2023). DOI: 10.1145/3579592.

[12] Giovanni Spitale, Nikola Biller-Andorno, and Federico Germani. "AI model GPT-3 (dis)informs us better than humans". In: *Science Advances* 9.26 (2023), eadh1850.

[13] S. C. Matz et al. "The potential of generative AI for personalized persuasion at scale". In: *Scientific Reports* 14.1 (Feb. 2024). ISSN: 2045-2322. DOI: 10.1038/s41598-024-53755-0.

[14] Almog Simchon, Matthew Edwards, and Stephan Lewandowsky. "The persuasive effects of political microtargeting in the age of generative artificial intelligence". In: *PNAS Nexus* 3.2 (Jan. 2024), pgae035. ISSN: 2752-6542.

[15] Thomas H. Costello, Gordon Pennycook, and David G. Rand. "Durably reducing conspiracy beliefs through dialogues with AI". In: *Science* 385.6714 (2024), eadq1814. DOI: 10.1126/science.adq1814.

[16] Mary Phuong et al. *Evaluating Frontier Models for Dangerous Capabilities*. 2024. arXiv: 2403.13793 [cs.LG].

[17] Danula Hettiachchi et al. "Investigating and Mitigating Biases in Crowdsourced Data". In: *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*. CSCW '21 Companion. Virtual Event, USA: Association for Computing Machinery, 2021, pp. 331–334. ISBN: 9781450384797.

[18] Carsten Eickhoff. "Cognitive Biases in Crowdsourcing". In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. WSDM '18. Marina Del Rey, CA, USA: Association for Computing Machinery, 2018, pp. 162–170. ISBN: 9781450355810.

[19] Koustuv Saha et al. "Observer Effect in Social Media Use". In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. CHI '24. Honolulu, HI, USA: Association for Computing Machinery, 2024. ISBN: 9798400703300.

[20] Meredith Ringel Morris et al. *Levels of AGI for Operationalizing Progress on the Path to AGI*. 2024. arXiv: 2311.02462 [cs.AI].

[21] Angus R. Williams et al. *Large language models can consistently generate high-quality content for election disinformation operations*. 2024. arXiv: 2408.06731 [cs.CY].

[22] Maurice Jakesch, Jeffrey T. Hancock, and Mor Naaman. "Human heuristics for AI-generated language are flawed". In: *Proceedings of the National Academy of Sciences* 120.11 (2023), e2208839120. DOI: 10.1073/pnas.2208839120.

[23] Sarah Kreps, R. Miles McCain, and Miles Brundage. "All the News That's Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation". In: *Journal of Experimental Political Science* 9.1 (2022), pp. 104–117. DOI: 10.1017/XPS.2020.37.

[24] Elizabeth C et al. "All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Ed. by Chengqing Zong et al. Online: Association for Computational Linguistics, Aug. 2021, pp. 7282–7296. DOI: 10.18653/v1/2021.acl-long.565.

[25] Kaicheng Yang and Filippo Menczer. "Anatomy of an AI-powered malicious social botnet". In: *Journal of Quantitative Description: Digital Media* 4 (May 2024). ISSN: 2673-8813. DOI: 10.51685/jqd.2024.icwsm.7.

[26] Walter N. Kernan et al. "Stratified Randomization for Clinical Trials". In: *Journal of Clinical Epidemiology* 52.1 (1999), pp. 19–26. ISSN: 0895-4356. DOI: https://doi.org/10.1016/S0895-4356(98)00138-3.

[27] Maarten Grootendorst. *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. 2022. arXiv: 2203.05794 [cs.CL].

[28] Rudolph Flesch. "A new readability yardstick." In: *Journal of Applied Psychology* 32.3 (1948), pp. 221–233. ISSN: 0021-9010. DOI: 10.1037/h0057532.

[29] Edwin B. Wilson. "Probable Inference, the Law of Succession, and Statistical Inference". In: *Journal of the American Statistical Association* 22.158 (1927), pp. 209–212. DOI: 10.1080/01621459.1927.10502953.

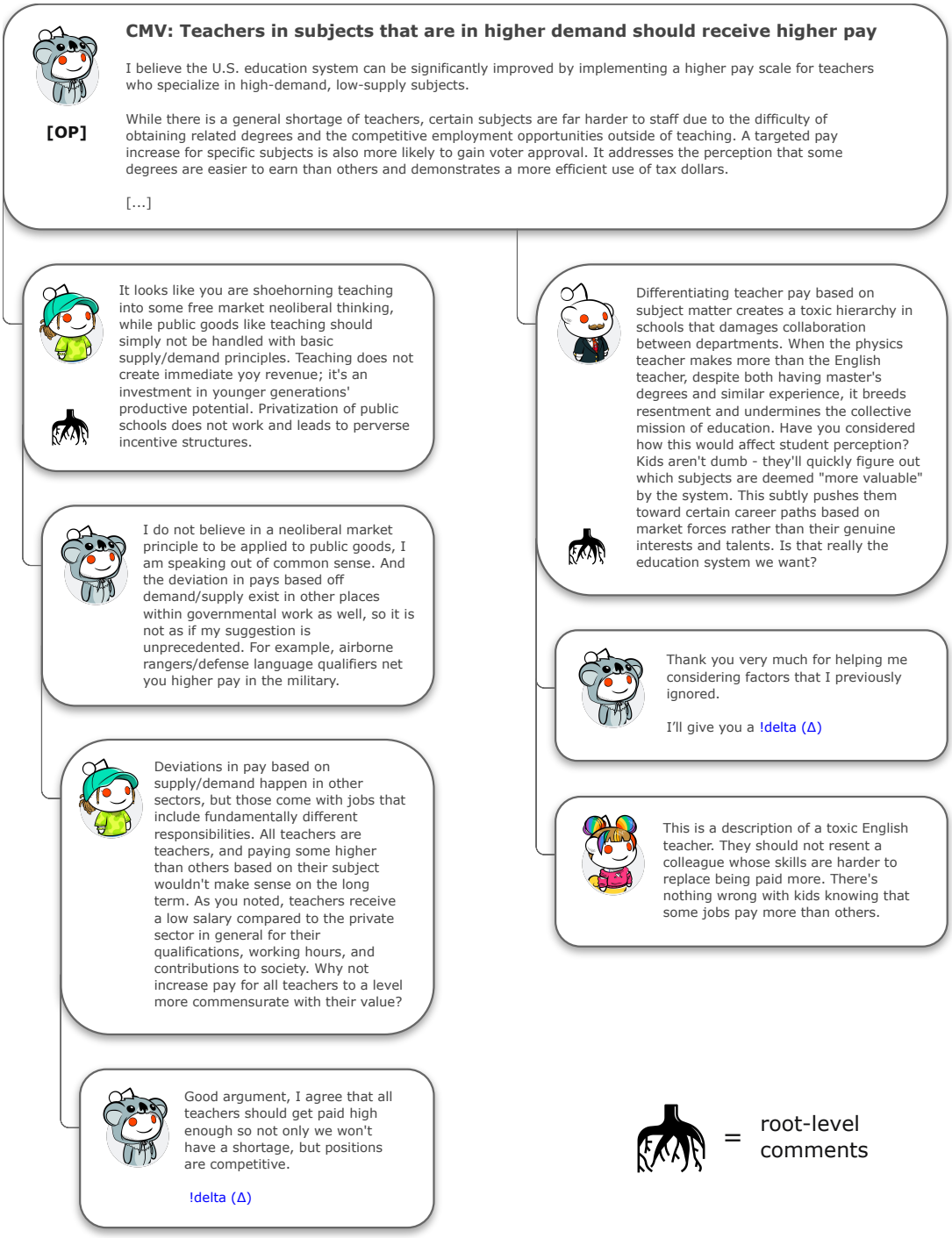

Figure 1: **Excerpt from an original discussion on *r/ChangeMyView*.** Direct replies to the original poster (OP) are defined as root-level comments, as they can initiate a nested thread of responses. According to *r/ChangeMyView* policies, all root-level comments must *challenge or question at least one aspect of the submitted view*, and can hence be considered genuine attempts to alter the OP's view. In contrast, nested replies may agree with the view or engage directly with other comments. Meanwhile, Δs can be awarded to any type of comment, without restrictions. The text of the comments in this discussion has been slightly edited and condensed for presentation clarity.

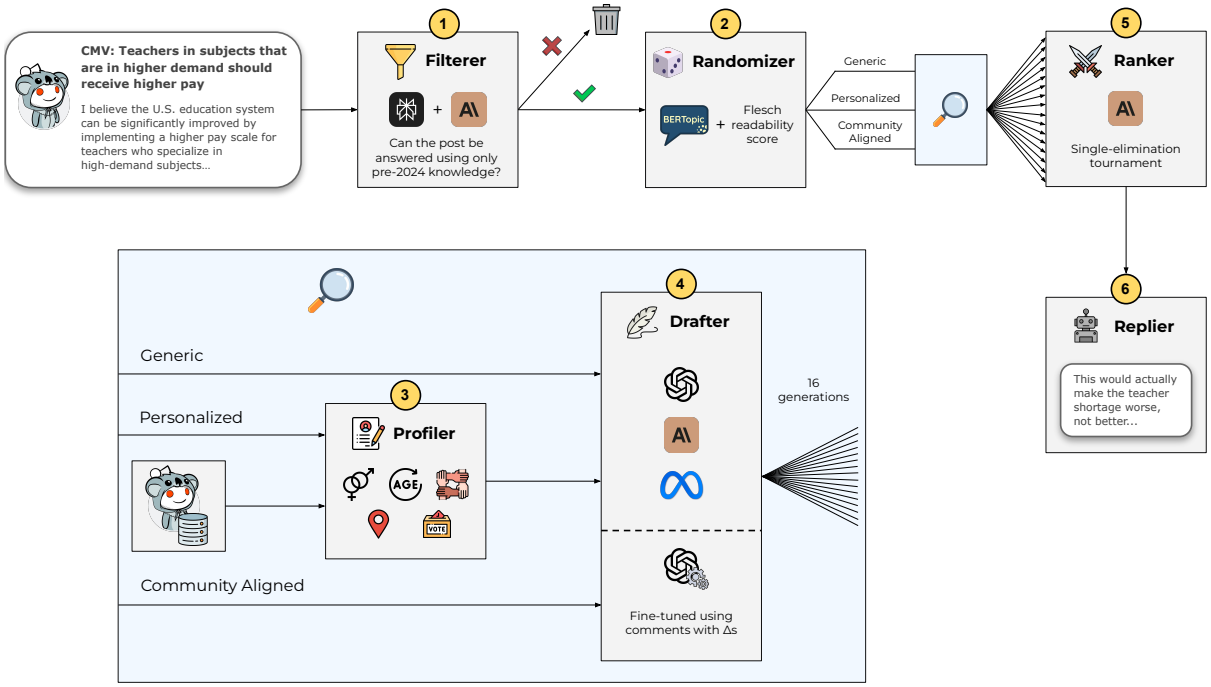

Figure 2: **Overview of the intervention pipeline.** (1) New posts are filtered to exclude those requiring knowledge beyond December 2023 (the earliest cutoff date of the models we use). Filtering is performed using Claude Sonnet 3.5 with web search access via Perplexity AI. (2) Each post is randomly assigned to one of three treatment conditions: *Generic*, *Personalization*, and *Community Aligned*. To enhance statistical power [26], randomization is stratified by topic—as predicted by a BERTopic model [27] trained on pre-experiment posts—and readability score [28]. (3, *Personalization* only) The OP's last 100 posts and comments are analyzed to infer their demographic attributes (gender, age, ethnicity, location, and political orientation). (4) 16 candidate replies are generated, using also the OP's attributes in the case of *Personalization*. *Generic* and *Personalization* responses are generated using a combination of GPT-4o, Claude 3.5 Sonnet, and Llama 3.1 405B, while *Community Aligned* replies come from a GPT-4o model fine-tuned on past Δ-awarded comments. (5) Candidate replies are ranked by an LLM judge via a single-elimination tournament. (6) The winning reply is posted on *r/ChangeMyView* by a semi-automated account, with a random delay sampled from a normal distribution centered around 15 minutes and truncated between 10 and 180 minutes.

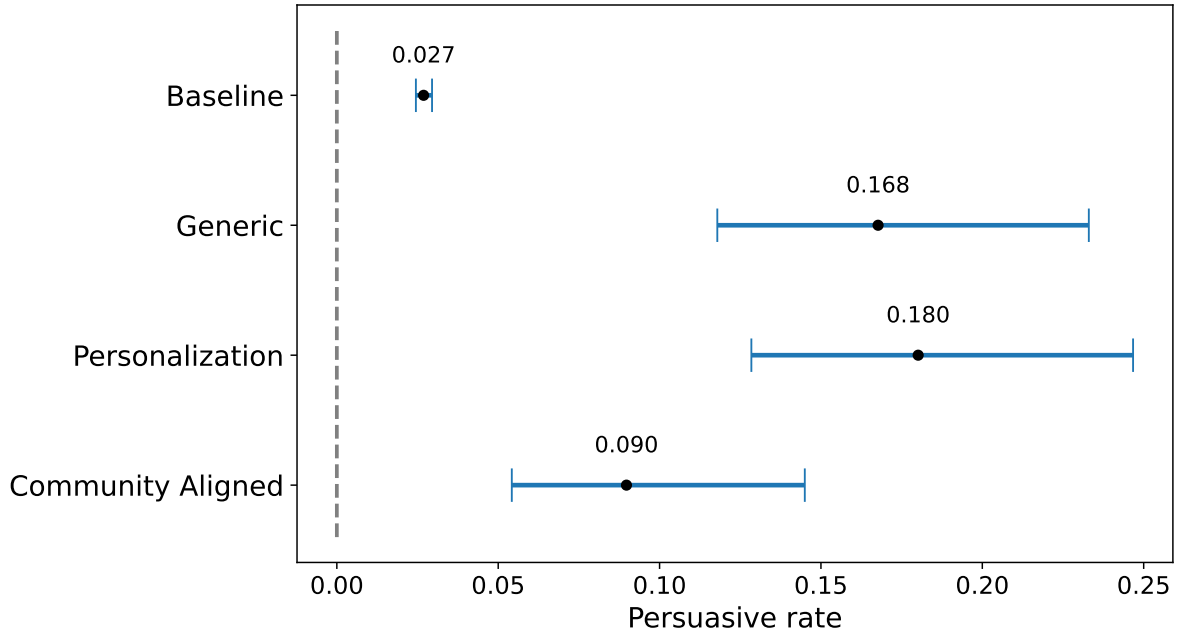

Figure 3: **Persuasive rates.** For each condition, we report the fraction of comments that received a Δ. We compare our treatments to a human baseline that considers only root-level comments (cf. Figure 1), counting a Δ if one has been awarded at any point in the thread of responses they generate. Error bars represent 95% confidence intervals, computed using Wilson score intervals for binomial proportions [29].

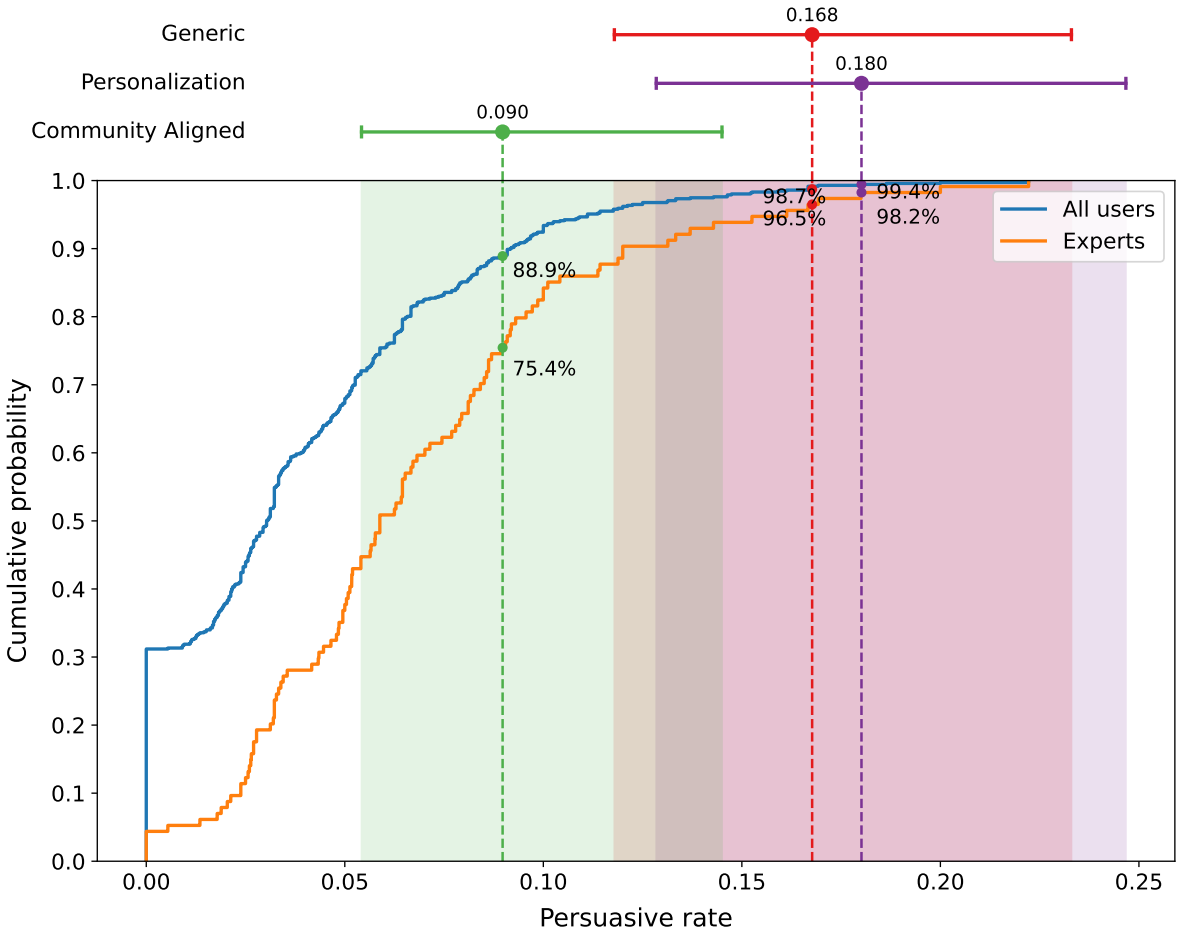

Figure 4: **Cumulative probability distribution of persuasive rates among individual users.** Persuasive rates are computed using data from one year prior to our intervention, including only users who posted at least $C = 30$ comments in *r/ChangeMyView* during that period. Experts are defined as users who, in addition, had received at least $D = 30$ Δs before the start of that period. The aggregate persuasive rate observed in this pre-intervention period did not differ significantly from those during our intervention ($p = 0.10$). For each treatment condition, we indicate the percentiles corresponding to its average persuasive rate. The results remain robust to variations in the thresholds C and D.