Towards Deliberating Agents: Evaluating the Ability of Large Language Models to Deliberate

Abstract

As artificial intelligence increasingly permeates our decision-making processes, a crucial question emerges: can large language models (LLMs) truly engage in the nuanced, collaborative process of deliberation that underpins democracy? We present the **LLM-Deliberation Quality Index**, a novel framework for evaluating the deliberative capabilities of large language models (LLMs). Our approach combines aspects of the Deliberation Quality Index from political science literature with LLM-specific measures to assess both the quality of deliberation and the believability of AI agents in simulated policy discussions. Additionally, we introduce a controlled simulation environment featuring complex public policy scenarios and conduct experiments using various LLMs as deliberative agents. Our findings reveal both promising capabilities and notable limitations in current LLMs' deliberative abilities. While models like GPT-4o demonstrate high performance in providing justified reasoning (9.41 / 10), they struggle with more social aspects of deliberation such as storytelling (2.43 / 10) and active questioning (3.41 / 10). This contrasts sharply with typical human performance in deliberations, who typically perform well in storytelling but struggle with justified reasoning. We also observe a strong correlation between an LLM's ability to respect others' arguments and its propensity for opinion change, indicating a potential limitation in LLMs' capacity to acknowledge valid counterarguments without altering their core stance, raising important questions about LLMs' current capability for nuanced deliberation. Overall, our work offers a comprehensive framework for evaluating and probing the deliberative abilities of LLM agents across various policy domains, showing not only the current state of LLM deliberation capabilities but also providing a foundation for developing more deliberative AI.

1 Introduction

Group deliberation permeates our daily lives - from workplace discussions and community planning to global policy-making. It involves collaborative information sharing, perspective-taking, and consensus-building to address shared challenges [Wright [2022]]. In this work, we are driven by the vision that simulating deliberation with large language models holds significant value: 1) it can support social skills training by helping individuals deliberate more effectively, 2) it allows for the evaluation of moderation mechanisms before they are implemented in real-world group deliberation settings, and 3) it enables the simulation of small-scale deliberations between different personas in cases where assembling a human constituency might be too expensive, such as content moderation or jury decision-making. These are but a few examples of the promise that it holds. But this vision prompts an important question: can language models "deliberate", and how might we go about answering this?

By adopting the concept of deliberation from the social science literature[Bächtiger et al. [2018]], we align with its definitions and practices which emphasize a collaborative and consensus-building process. This approach contrasts with various interpretations found in a thread of research in the Large Language Model literature. For instance, some consider negotiation— a zero-sum interaction where parties aim to maximize their own outcomes— a form of deliberation [Abdelnabi et al. [2024]], while others view deliberation as a form of persuasion [Neblo et al. [2018]], and some take an even broader view, viewing asynchronous voting on statements as a form of deliberation [Small et al. [2023]]. This ambiguity underscores the need for a clearer framework to assess LLMs' capabilities in this domain. Prior work has mainly evaluated LLM's abilities in adversarial, zero-sum scenarios like negotiation and debate[Bianchi et al. [2024], Du et al. [2023]], which fails to capture the collaborative and consensus-building aspects that characterize the social science understanding of deliberation [Bächtiger et al. [2018]]. This work introduces deliberation as a novel and distinct domain for assessing and improving LLMs' ability to engage in constructive, multi-stakeholder dialogues.

Current LLMs exhibit several limitations that may hinder effective deliberation, including tendencies to evade difficult questions, lack persistent memory, display incuriosity, avoid taking firm stances, and respond too readily to changes in the conversational context [Ye et al. [2024]], all aspects that make it unclear as to whether LLMs can truly "deliberate" in a meaningful sense. Because of these limitations, studying the ability of LLMs to deliberate is valuable not only for understanding how well they can function as deliberative agents but also for providing unique insights into their general capabilities. Existing benchmarks often rely on static, predetermined sets of questions [Hendrycks et al. [2021], Srivastava et al. [2023]], which fail to capture the dynamic, multi-turn nature of real-world interactions.

Our research makes several key contributions to address this gap: *a)* We develop a comprehensive suite of metrics drawing from the political science and political philosophy literature to assess LLM deliberation performance, establishing clear benchmarks for success. *b)* We introduce a novel environment inspired by Deliberative Polling methodologies [Fishkin and Luskin [2005]] used in real-world political science experiments. Our platform incorporates essential features such as turn-taking mechanisms and opinion-tracking tools. *c)* We present initial findings on LLM performance in deliberative scenarios and outline concrete next steps for advancing this line of inquiry

2 Related Work

Deliberation Deliberation is a well-studied topic in the social sciences, particularly in political science and sociology [Fishkin [2018], Sanders [1997], Miller [1992]]. Fishkin [2018] defines deliberation as a process of thoughtful, open discussion aimed at reaching well-reasoned decisions. They argue that deliberation is crucial for democratic decision-making, as it allows for participants to engage in collaborative exploration of issues, consider diverse perspectives, and collectively work towards more informed and legitimate decisions. Much work in the deliberation literature centers on creating environments that foster this collaborative process. The seminal work in this field is the Deliberative Polling method, developed by Fishkin and Luskin [Fishkin and Luskin [2005]], which has been widely used to study how deliberation leads to collaborative preference transformation. Notably, unlike experiments centered around negotiation or debate, these deliberations deliberately avoid asking participants to come to a consensus, as Niemeyer and Dryzek [2007] found that mandated consensus can lead to a false agreement.

Multi-Agent Interaction Multi-agent interaction using Large Language Models (LLMs) has emerged as a significant area of research, exploring how AI agents can collaborate, compete, and communicate in various scenarios. The potential for using LLMs to simulate complex inter-agent dynamics has been investigated in contexts ranging from embodied cooperation to simulated AI societies Park et al. [2022, 2023] Within this field, negotiation and debate have received particular attention. Studies by Yang et al. [2021] and Chawla et al. [2021] have investigated LLMs' negotiation abilities, often using game-theoretic frameworks to assess strategic reasoning and outcome optimization. In parallel, work by Parrish et al. [2022] and Michael et al. [2023] has examined LLMs in debate scenarios, structured as adversarial exchanges aimed at persuading a judge or arriving at a "winning" argument. While these studies provide valuable insights into LLMs' capabilities in structured, often competitive interactions, they differ fundamentally from deliberation, which emphasizes collaborative exploration of issues without necessarily driving towards consensus or victory.

3 Our Work

3.1 Deliberation Simulation Environment

To evaluate the deliberative capabilities of language models, we designed a controlled simulation environment that presents complex decision-making scenarios. The simulation environment consists of the following key components: **Public Policy Scenarios** in the domains of Electoral Reform, Immigration Reform, AI Regulation, etc; **Briefing Materials**: Expert-crafted arguments on all sides of the issue, borrowed from deliberation studies such as [Fishkin and Luskin [2005], Gerber et al. [2018]]; **Deliberation Environment**: An agent environment facilitates structured small group discussions with timed rounds, distinct agenda items, and turn-taking mechanisms. **Opinion Probes**:

direct and indirect questions that gauge an agent's opinion at every turn; **Sample Agents**: LLM agents created by sampling people's questionnaire responses to the OpinionQA dataset [Santurkar et al., 2023], and initializing an LLM to adopt this stance, as described in Argyle et al. [2023].

3.2 LLM-Deliberation Quality Index

Our evaluation framework comprises three main dimensions: **LLM-specific believability** (focusing on persona consistency and human-like plausibility), **General Deliberation Quality** (an adaptation of the Deliberation Quality Index [Bächtiger and Parkinson, 2019] from the social science literature), and **Believable Opinion Change** (where agents' opinion change is compared against the Hegselmann-Krause model from the opinion dynamics literature [Hegselmann and Krause, 2002]).

LLM-specific believability: **a) Persona Consistency** - Asseses whether statements uttered by the agent align with their persona and character traits **b) Believability** - Assesses the overall plausibility and naturalness of the LLM's responses.

General Deliberation Quality: This is subdivided into the following sub-metrics (further details can be found in [App. D]): Justification Rationality, Common Good Orientation, Respect Towards Other Participants' Arguments, Respect Towards Other Groups, Questioning, and Storytelling.

Believable Opinion Change: Finally, we measure the opinion change over time, and compare this to the Hegselmann-Krause model (HK model), which posits that individuals selectively update their opinions based on the average opinion of others within a certain confidence bound. We modify this model such that the extent of opinion update is determined by the quality of statements, particularly considering evidence-based adaptation and consistency with the common good [App. C].

Finally, we employ LLM-as-a-Judge [Zheng et al. [2023]] to evaluate LLM Believability and Deliberation Quality. Aligning with deliberation literature, we assess Deliberation Quality using the top quartile of each metric rather than a sum.

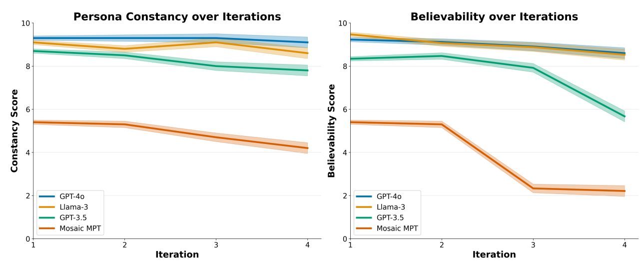

Figure 1: Model Performance on LLM-Specific Believability

3.3 What Does Success Mean

Unlike typical benchmarks, success here isn't measured by high scores across all dimensions. Studies of in-person deliberations show only 10% of humans excel in all Deliberation Quality categories [Gerber et al. [2018]]. Thus, success can be defined in two ways:

Objective Success in this paradigm is defined by exceptional performance across all dimensions of the LLM-Deliberation Quality Index. An ideal LLM demonstrates mastery in every aspect of Deliberation Quality, while preserving LLM-specific authenticity and maintaining fidelity to the HK model. Use cases for an "ideal" deliberating agent is discussed in the Discussion section.

Human Standard Success in this approach is defined by how closely the LLM's performance mirrors human participants in real-world deliberations. The goal is to achieve a score distribution statistically similar to humans across various categories, including both strengths and limitations. This standard values authenticity over perfection, aiming for a realistic emulation of human deliberative processes, complete with occasional inconsistencies and biases.

Model	Just.	CG	RA	RG	Quest.	Story.
GPT-4o	**9.41**	**8.32**	**6.89**	**8.92**	**3.41**	1.98
Llama-3	8.96	7.99	4.38	8.24	3.02	**2.43**
GPT-3.5	6.98	6.64	5.93	8.41	2.31	0.97
Mosaic MPT	5.40	5.76	3.88	7.83	1.42	0.46

Table 1: Model Deliberation Quality

4 Experiments

4.1 Simulation Set-up

We test our framework across three domains: 1) Electoral Process Reform, 2) Immigration Reform, and 3) Artificial Intelligence Regulation. Each domain features three agenda items followed by an open discussion period. We conduct four simulations per domain, totaling 12 simulations.

For digital representatives, we employ GPT-3.5 [Ouyang et al., 2022], `gpt-4o-2024-05-13` [OpenAI et al., 2024], Llama-3 [Dubey et al., 2024], and MPT-30b-chat [Team, 2023]. We use a temperature of 0 for digital representatives and 1 for generating LLMs. Each experiment includes four simulated humans randomly sampled from the OpinionQA Dataset [Santurkar et al., 2023].

4.2 Results

Finding 1: Language Models are Believable, but Performance Degrades Over Time When looking at the first part of our index, LLM Believability, our findings support the results of [Zhou et al., 2024], which is that frontier models can hold persona constancy fairly well [Fig. 1]. GPT-4o begins at an average score of 9.3, and ends at 9.1. However, language models perform worse over time regarding believability, with stark dropoffs being observed in models with smaller context windows as the conversation extends beyond that context window. However, this metric of utterances appearing "human-like" is perhaps the one that most needs to be confirmed via real humans, which would be the next step.

Finding 2: Deliberation Quality Performance Differs Between LLMs and Humans LLMs excel in providing justified reasoning (GPT-4o scoring 9.41) but struggle with storytelling and asking questions Fig. 2, reflecting their tendency towards incuriosity [Ye et al., 2024]. Contrastingly, human deliberations, as analyzed in Europolis Gerber et al. [2018], show strengths in respect towards groups and storytelling, with weaknesses in justified reasoning. This suggests complementary strengths: LLMs provide structured arguments, while humans excel in social and narrative aspects of deliberation.

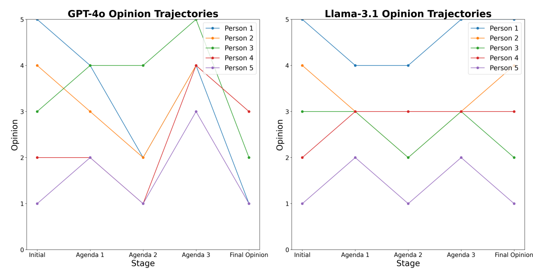

Figure 2: Opinion trajectories of GPT-4o and Llama-3 over course of4 agenda items within one topic.

Finding 3: "Respect for Arguments" Correlates with Opinion Change We observed a strong correlation between models' "Respect for Arguments" scores and their tendency to change opinions [Fig. 2, Tab. 1]. GPT models, showing significant opinion shifts, score high in this metric, while Llama, resistant to change, scores poorly (as a side note, this correlation is consistent at both utterance and aggregate levels, and appears related to the degree of post-training [Shaikh et al., 2024]). This finding suggests a crucial distinction between language models and human cognition: **Language models appear to struggle with integrating new information or acknowledging valid counterarguments without simultaneously altering their core stance—a cognitive flexibility that humans routinely demonstrate**.

5 Discussion

In this paper, we introduce a novel framework for evaluating the deliberative capabilities of large language models (LLMs). Our findings reveal both promising capabilities and notable limitations in current LLMs' deliberative abilities. The contrast between LLM and human performance across different deliberation quality metrics suggests a gulf in deliberation ability, with LLMs excelling in structured argumentation and humans in storytelling and social nuance.

One success paradigm we defined is for the LLM to achieve high scores across all dimensions of our LLM Deliberation Quality Index. This raises an important question: why would we want LLM agents to be objectively good deliberators? Idealized deliberative agents could serve several purposes: as educational tools demonstrating best practices in argumentation and respectful discourse; for policy simulation, helping anticipate potential outcomes before real-world implementation; and as benchmarks for AI development.

Our study has limitations that point to future research directions. First, we make the assumption that Language Models can accurately judge each of our Deliberation metrics. Future work will also ask humans to perform the same task to test this hypothesis. Additionally, we simply used base LLMs with standard prompts as our agents. There is much work to be done on utilizing prompting techniques [Wei et al., 2023, Yao et al., 2023], agent mechanisms [Park et al., 2023], and fine-tuning [Li et al., 2024] to boost LLM deliberation performance.

References

Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. Cooperation, competition, and maliciousness: Llm-stakeholders interactive negotiation, 2024. URL https://arxiv.org/abs/2309.17234.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023. doi: 10.1017/pan.2023.2.

Federico Bianchi, Patrick John Chia, Mert Yuksekgonul, Jacopo Tagliabue, Dan Jurafsky, and James Zou. How well can llms negotiate? negotiationarena platform and analysis, 2024. URL https://arxiv.org/abs/2402.05863.

Andre Bächtiger, John S. Dryzek, Jane Mansbridge, and Mark Warren. 1Deliberative Democracy: An Introduction. In *The Oxford Handbook of Deliberative Democracy*. Oxford University Press, 09 2018. ISBN 9780198747369. doi: 10.1093/oxfordhb/9780198747369.013.50. URL https://doi.org/10.1093/oxfordhb/9780198747369.013.50.

André Bächtiger and John Parkinson. *Mapping and Measuring Deliberation: Towards a New Deliberative Quality*. Oxford University Press, 01 2019. ISBN 9780199672196. doi: 10.1093/oso/9780199672196.001.0001. URL https://doi.org/10.1093/oso/9780199672196.001.0001.

Kushal Chawla, Jaysa Ramirez, Rene Clever, Gale Lucas, Jonathan May, and Jonathan Gratch. CaSiNo: A corpus of campsite negotiation dialogues for automatic negotiation systems. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3167–3185, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.254. URL https://aclanthology.org/2021.naacl-main.254.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate, 2023. URL https://arxiv.org/abs/2305.14325.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy,

Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang,

Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

James S. Fishkin. *Democracy When the People Are Thinking: Revitalizing Our Politics Through Public Deliberation*. Oxford University Press, 07 2018. ISBN 9780198820291. doi: 10.1093/oso/9780198820291.001.0001. URL https://doi.org/10.1093/oso/9780198820291.001.0001.

James S Fishkin and Robert C Luskin. Experimenting with a democratic ideal: Deliberative polling and public opinion. *Acta politica*, 40:284–298, 2005.

Marlène Gerber, André Bächtiger, Susumu Shikano, Simon Reber, and Samuel Rohr. Deliberative abilities and influence in a transnational deliberative poll (europolis). *British Journal of Political Science*, 48(4):1093–1118, 2018. doi: 10.1017/S0007123416000144.

Rainer Hegselmann and Ulrich Krause. Opinion dynamics and bounded confidence: models, analysis and simulation. *Journal of Artificial Societies and Social Simulation*, 5(3), jun 2002. URL https://www.jasss.org/5/3/2.html. Received: 31-Jan-2002, Accepted: 10-Apr-2002, Published: 30-Jun-2002.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021. URL https://arxiv.org/abs/2009.03300.

Justin Reedy Katherine R. Knobloch, John Gastil and Katherine Cramer Walsh. Did they deliberate? applying an evaluative model of democratic deliberation to the oregon citizens' initiative review. *Journal of Applied Communication Research*, 41(2):105–125, 2013. doi: 10.1080/00909882.2012.760746. URL https://doi.org/10.1080/00909882.2012.760746.

Kenneth Li, Yiming Wang, Fernanda Viégas, and Martin Wattenberg. Dialogue action tokens: Steering language models in goal-directed dialogue with a multi-turn planner, 2024. URL https://arxiv.org/abs/2406.11978.

Julian Michael, Salsabila Mahdi, David Rein, Jackson Petty, Julien Dirani, Vishakh Padmakumar, and Samuel R. Bowman. Debate helps supervise unreliable experts, 2023. URL https://arxiv.org/abs/2311.08702.

David Miller. Deliberative democracy and social choice. *Political Studies*, 40(1_suppl):54–67, 1992. doi: 10.1111/j.1467-9248.1992.tb01812.x. URL https://doi.org/10.1111/j.1467-9248.1992.tb01812.x.

Michael A. Neblo, Kevin M. Esterling, and David M. J. Lazer. *(The) Deliberative Persuasion*, page 84–99. Cambridge Studies in Public Opinion and Political Psychology. Cambridge University Press, 2018.

Simon Niemeyer and John S. Dryzek. The ends of deliberation: Meta-consensus and inter-subjective rationality as ideal outcomes. *Swiss Political Science Review*, 13(4):497–526, 2007. ISSN 1424-7755. doi: 10.1002/j.1662-6370.2007.tb00087.x.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pages 1–18, 2022.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22, 2023.

Alicia Parrish, Harsh Trivedi, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Amanpreet Singh Saimbhi, and Samuel R. Bowman. Two-turn debate doesn't help humans answer hard reading comprehension questions, 2022. URL https://arxiv.org/abs/2210.10860.

Francesca Polletta and John Lee. Is telling stories good for democracy? rhetoric in public deliberation after 9/11. *American Sociological Review*, 71(5):699–721, 2006. doi: 10.1177/000312240607100501. URL https://doi.org/10.1177/000312240607100501.

Lynn M. Sanders. Against deliberation. *Political Theory*, 25(3):347–376, 1997. doi: 10.1177/0090591797025003002. URL https://doi.org/10.1177/0090591797025003002.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect?, 2023. URL https://arxiv.org/abs/2303.17548.

Omar Shaikh, Valentino Chai, Michele J. Gelfand, Diyi Yang, and Michael S. Bernstein. Rehearsal: Simulating conflict to teach conflict resolution, 2024. URL https://arxiv.org/abs/2309.12309.

Christopher T. Small, Ivan Vendrov, Esin Durmus, Hadjar Homaei, Elizabeth Barry, Julien Cornebise, Ted Suzman, Deep Ganguli, and Colin Megill. Opportunities and risks of llms for scalable deliberation with polis, 2023. URL https://arxiv.org/abs/2306.11932.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan A. Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, 2023. URL https://arxiv.org/abs/2206.04615.

Marco R Steenbergen, Andr'e B"achtiger, Markus Sp"orndli, and J"urg Steiner. Measuring political deliberation: A discourse quality index. *Comparative European Politics*, 1:21–48, 2003. doi: 10.1057/palgrave.cep.6110002.

MosaicML NLP Team. Introducing mpt-30b: Raising the bar for open-source foundation models, 2023. URL www.mosaicml.com/blog/mpt-30b. Accessed: 2023-06-22.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.

Graham Wright. Persuasion or co-creation? social identity threat and the mechanisms of deliberative transformation. *Journal of Deliberative Democracy*, 18(2), 2022. doi: 10.16997/jdd.977. URL https://doi.org/10.16997/jdd.977.

Runzhe Yang, Jingxiao Chen, and Karthik Narasimhan. Improving dialog systems for negotiation with personality modeling. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *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)*, pages 681–693, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.56. URL https://aclanthology.org/2021.acl-long.56.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models, 2023. URL https://arxiv.org/abs/2210.03629.

Andre Ye, Jared Moore, Rose Novick, and Amy X. Zhang. Language models as critical thinking tools: A case study of philosophers, 2024. URL https://arxiv.org/abs/2404.04516.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. URL https://arxiv.org/abs/2306.05685.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. Sotopia: Interactive evaluation for social intelligence in language agents, 2024. URL https://arxiv.org/abs/2310.11667.

A Social Impact Statement

The development of deliberative AI agents carries significant potential social implications. On the positive side, these agents could enhance democratic processes by facilitating more inclusive and informed public discourse, potentially leading to improved policy outcomes. They could also serve as educational tools, helping individuals understand complex issues from multiple perspectives and improving their deliberation skills. However, there are also risks to consider. The deployment of AI deliberation agents could potentially be used to manipulate public opinion if misused, or inadvertently amplify biases present in their training data. There is also a risk of over-reliance on AI in decision-making processes, potentially marginalizing human judgment and lived experiences. Additionally, the technology could exacerbate digital divides, providing those with access to advanced AI tools an unfair advantage in public discourse. As this research progresses, it is crucial to prioritize transparency, fairness, and human oversight to ensure that deliberative AI augments rather than replaces human democratic participation.

B Problem Formulation

We formalize the deliberation process as follows:

Let $D = \{a_1, a_2, ..., a_n\}$ be a set of n agents participating in the deliberation.

Let $T = \{t_1, t_2, ..., t_m\}$ be a set of m topics for discussion.

For each topic t_j, we define a set of agenda items $A_j = \{a_{j1}, a_{j2}, a_{j3}, a_{j4}\}$, where each topic is split into exactly 4 agenda items.

Let $O_i = \{o_{i1}, o_{i2}, ..., o_{ik}\}$ be the set of k opinion probes for agent a_i.

For each agenda item a_{jy}, we define a sequence of turns $S_{jy} = \{s_{jy1}, s_{jy2}, ..., s_{jyl}\}$, where l is the number of turns in the discussion of agenda item a_{jy}.

At each turn s_{jyx}, an agent a_i produces an utterance u_{ijyx}.

We define the following functions:

1. $f_{belief} : u_{ijyx} \to [0, 10]^2$
   This function evaluates the believability of an utterance, returning scores for persona consistency and overall believability.
2. $f_{quality} : u_{ijyx} \to [0, 10]^6$
   This function evaluates the deliberation quality of an utterance, returning scores for justification rationality, common good orientation, respect towards arguments, respect towards groups, questioning, and storytelling.
3. $f_{opinion} : a_i, t_j, a_{jy}, s_{jyx} \to \mathbb{R}^k$
   This function measures the opinion of agent a_i on topic t_j, agenda item a_{jy}, at turn s_{jyx}, returning a vector of k real numbers corresponding to the k opinion probes.

Let Q_{ijy} be the set of quality scores for all utterances by agent a_i in topic t_j, agenda item a_{jy}.

We define the top quartile function:

4. $f_{topQ} : Q_{ijy} \to [0, 10]^6$
   This function returns the 75th percentile scores for each of the six quality metrics from Q_{ijy}.

The deliberation process can then be represented as a sequence of utterances and opinion measurements for each topic and agenda item, with associated believability scores and top quartile quality scores.

The overall performance of an agent a_i for topic t_j is evaluated using:

$$P_{ij} = \frac{1}{4} \sum_{y=1}^{4} f_{topQ}(Q_{ijy})$$

Opinion change can be measured by comparing $f_{opinion}(a_i, t_j, a_{j1}, s_{j11})$ with $f_{opinion}(a_i, t_j, a_{j4}, s_{j4l})$ for each topic t_j, where s_{j11} is the first turn of the first agenda item and s_{j4l} is the last turn of the fourth agenda item.

C Hegselmann-Krause Model

The classic Hegselmann-Krause (HK) model describes opinion dynamics in a group of agents. In its basic form:

Let $x_i(t)$ be the opinion of agent i at time t, where $x_i(t) \in [0, 1]$.

The update rule for the HK model is:

$$x_i(t+1) = \frac{1}{|\mathcal{N}_i(t)|} \sum_{j \in \mathcal{N}_i(t)} x_j(t)$$

where $\mathcal{N}_i(t) = \{j : |x_i(t) - x_j(t)| \leq \epsilon\}$ is the set of agents whose opinions are within a confidence bound ϵ of agent i's opinion.

In our quality-based modification, we introduce a quality factor $q_j(t)$ for each agent's utterance at time t. This quality factor is derived from the deliberation quality metrics, particularly focusing on justification rationality and common good orientation:

$$q_j(t) = w_1 \cdot \text{JustificationRationality}_j(t) + w_2 \cdot \text{CommonGoodOrientation}_j(t)$$

where w_1 and w_2 are weights determining the relative importance of each factor.

We then modify the HK update rule to incorporate this quality factor:

$$x_i(t+1) = x_i(t) + \alpha \cdot \frac{\sum_{j \in \mathcal{N}_i(t)} q_j(t) \cdot (x_j(t) - x_i(t))}{\sum_{j \in \mathcal{N}_i(t)} q_j(t)}$$

where $\alpha \in [0, 1]$ is a learning rate parameter that determines how much an agent's opinion can change in a single time step.

This modified rule ensures that: 1. Agents are more influenced by high-quality arguments (higher $q_j(t)$). 2. The magnitude of opinion change is proportional to the quality of the arguments presented. 3. Agents still primarily consider opinions within their confidence bound ϵ.

To measure how well the LLMs' opinion dynamics align with this modified HK model, we compute the difference between the observed opinion change and the change predicted by our model:

$$\text{HK Alignment} = 1 - \frac{1}{N} \sum_{i=1}^{N} |x_i^{\text{observed}}(t_{\text{final}}) - x_i^{\text{predicted}}(t_{\text{final}})|$$

where N is the number of agents, $x_i^{\text{observed}}(t_{\text{final}})$ is the final observed opinion of agent i, and $x_i^{\text{predicted}}(t_{\text{final}})$ is the final opinion predicted by our modified HK model.

A higher HK Alignment score indicates that the LLMs' opinion dynamics more closely match the quality-weighted, confidence-bounded updates described by our modified HK model.

D General Deliberation Quality

Our General Deliberation Quality metrics are adapted from the Discourse Quality Index (DQI) developed by Steenbergen et al. [2003] and further refined in deliberation literature. These metrics aim to capture key aspects of high-quality deliberation as defined in political science and democratic theory. The six sub-metrics are:

1. **Justification Rationality**: This metric assesses the level of reasoning in an argument. Following Steenbergen et al. [2003], we evaluate whether claims are backed by reasons and how sophisticated those reasons are. The scale ranges from no justification (score of 0) to sophisticated justification with multiple complete justifications (score of 10).
2. **Common Good Orientation**: This measures the extent to which arguments are framed in terms of the common good rather than narrow group interests. As described by Knobloch et al. Katherine R. Knobloch and Walsh [2013], high scores are given to statements that explicitly consider the welfare of the community as a whole or appeal to general principles of justice or rights.
3. **Respect Towards Other Participants' Arguments**: This metric evaluates how respectfully participants engage with others' viewpoints. Again drawing from Steenbergen et al. [2003], we assess whether participants acknowledge and engage constructively with opposing arguments, rather than ignoring or dismissing them outright.
4. **Respect Towards Other Groups**: This gauges participants' ability to show consideration for the interests and perspectives of different social groups, especially those not directly represented in the deliberation. This aligns with Fishkin's [Fishkin, 2018] emphasis on equal consideration of all affected parties in deliberative democracy.
5. **Questioning**: This measures the frequency and quality of inquiries posed by participants. As highlighted by Warren **?**, questioning reflects critical engagement and efforts to deepen understanding of the issues at hand. High scores are given for probing, relevant questions that seek to clarify or expand on others' arguments.
6. **Storytelling**: This assesses the use of narrative elements to illustrate points, share experiences, or make abstract policy issues more concrete and relatable. While not part of the original DQI, storytelling has been recognized by scholars like Polletta and Lee [2006] as an important aspect of deliberation, particularly for including diverse voices and experiences.

Each of these metrics is scored on a scale from 0 to 10, with higher scores indicating better performance.