CIVICPARSE: A Benchmark and Pipeline for Structured Online Deliberation

Anonymous Author(s)
Affiliation
Address
email

Abstract

Online deliberation platforms promise scalable collective intelligence, yet their free-form threads are difficult to navigate, summarize, and moderate. We argue that progress requires treating *structured deliberation* as a formal natural language processing (NLP) problem with civic significance: reliably mapping raw discussions into a deliberation-native schema so that key barriers, solutions, metrics, and stances are visible at scale. We introduce **CIVICPARSE**, a two-stage pipeline that operationalizes this problem as extraction and classification over a domain-grounded schema. Stage 1 extracts distinct points from threads; Stage 2 assigns *Barrier*, *Solution*, or *Metric* types together with *Pro*/*Con* roles. Trained on 840 curated Deliberatorium examples[1], **CIVICPARSE** attains 88.5% accuracy with strong precision (91.1%) and recall (96.5%), substantially outperforming identical prompt-only baselines. Beyond the gains from fine-tuning, we contribute a reproducible extractor–classifier design, a curated dataset, and an evaluation protocol that together cast structured deliberation as a *benchmarkable* task for AI-assisted civic decision-making.

1 Introduction

Public decision-making increasingly unfolds on platforms that invite wide participation, but the resulting discussions are sprawling and unstructured. Without reliable structure, moderators and policymakers struggle to identify core barriers, competing solutions, and points of agreement or dissent. Decades of work on structured deliberation and argument mapping—including IBIS/gIBIS and the Deliberatorium—show that explicit schemas of issues, positions, and arguments improve navigation and traceability [Conklin and Begeman, 1989, Iandoli et al., 2007, Klein, 2012]. These systems help groups reason together by turning conversations into linked, analyzable objects.

A parallel literature documents persistent concerns in open forums: incivility, polarization, and unequal participation. Design constraints and structured interfaces can mitigate these effects by focusing contributions on substance and reducing duplication [Coe et al., 2014, Kennedy et al., 2020, Rega and Marchetti, 2021, Klein and Majdoubi, 2024]. Platforms like Pol.is demonstrate that even lightweight structure, combined with clustering, can surface consensus patterns across polarized groups [Small et al., 2023]. Yet the adoption bottleneck remains: maintaining fine-grained structure has relied on labor-intensive curation, which does not scale to large, open communities.

We contend that a *deliberation-native* formulation is needed: treat structure induction for online deliberation as a *benchmarkable NLP task* with well-defined outputs and evaluation, not merely as an engineering convenience. Argument mining has mapped claims, evidence, and relations in debates [Lawrence and Reed, 2020, Lippi and Torroni, 2018, Karadzhov et al., 2021], and systems

[1] https://deliberatorium.org/show-page?login

like IBM's Project Debater show end-to-end argument pipelines [Slonim et al., 2021]. Modern LLMs achieve competitive zero-shot performance on stance, relevance, and toxicity [Gilardi et al., 2023], but prompt-only approaches in deliberation settings exhibit redundancy and conflate corrective proposals with opposition. Evidence from moderation benchmarks indicates that domain-specialized adaptation improves robustness on community distributions [Zhan et al., 2025, Machlovi et al., 2025, Pietron et al., 2023].

We present **CIVICPARSE** , a two-stage pipeline that maps raw deliberation threads into a structured schema. Stage 1 extracts distinct points; Stage 2 assigns each point a type (*Barrier*, *Solution*, *Metric*) and stance (*Pro*/*Con*). Trained on 840 annotated Deliberatorium examples, **CIVICPARSE** reduces redundancy and misclassification compared to prompt-only baselines. Our contributions are: **(1)** a deliberation-native schema and extractor–classifier pipeline that frame structured moderation as a benchmarkable NLP task; **(2)** a curated dataset of 1,200 annotated items with labeling guidelines for reproducibility; and **(3)** empirical evidence of higher accuracy and recall with fewer structure-to-stance confusions, enabling more reliable large-scale civic analysis.

2 Related Work

Structured deliberation platforms. Formal schemas (IBIS/gIBIS) encode issues, positions, and arguments for collective sensemaking Conklin and Begeman [1989]. The Deliberatorium operationalizes these ideas via attention-mediation and curation workflows for large groups Iandoli et al. [2007], Klein [2012]. Empirical studies document incivility and participation inequities in open forums and show that structured formats reduce toxicity relative to unstructured settings Coe et al. [2014], Kennedy et al. [2020], Rega and Marchetti [2021], Klein and Majdoubi [2024]. Complementary literatures emphasize inclusive design and organizational routines that improve deliberation quality Abdel-Monem et al. [2010], Baek et al. [2012], Møller [2021], Niemeyer et al. [2023], Volkovskii and Filatova [2023], Williams [2010], Wojcieszak [2011], Sunstein [2006], Tapscott and Williams [2006]. Beyond moderation, Pol.is uses clustering and visualization to surface consensus Small et al. [2023], and systems work on discussion trees and facilitation agents explores steering discourse productively Sengoku et al. [2016], Ito et al. [2022]. Human-in-the-loop AI is emerging for argument mapping Anastasiou and De Liddo [2025]. These efforts largely target participation and civility; they seldom address *fine-grained* structure induction in deliberation-native text.

Argument mining and LLM-based induction. Argument mining synthesizes tasks and resources for extracting propositions and relations in multi-party discourse Lawrence and Reed [2020], Lippi and Torroni [2018], Karadzhov et al. [2021]; IBM's Project Debater showcased end-to-end retrieval and generation Slonim et al. [2021]. LLMs extend this line of work: zero-shot models can rival non-experts for stance and toxicity Gilardi et al. [2023], but they remain brittle in domain-specific deliberation. Domain-adapted models outperform generic prompting for moderation and classification on community datasets Zhan et al. [2025], Machlovi et al. [2025], Pietron et al. [2023]. Advances in instruction following and reasoning further improve structured outputs Ouyang et al. [2022], Wei et al. [2022], Huang et al. [2022], Zhao et al. [2023], Weng [2023], Wynter and Yuan [2023], Burnell et al. [2023]. Still, redundancy and label confusions specific to deliberation are under-explored.

Quality evaluation, consensus, and fairness. Recent work proposes interpretable metrics that combine expert/crowd judgments with model predictions to assess deliberation quality Behrendt et al. [2024]. Studies also show that AI can help groups converge on high-approval statements and common ground Bakker et al. [2022], Tessler et al. [2024]. As such systems scale, fairness diagnostics become crucial: dialect-focused audits reveal performance gaps in NLU and moderation pipelines, and cross-dialect benchmarks test generalization beyond standard corpora Gupta et al. [2024, 2025a]. Long-context diagnostics expose multi-hop reasoning failures that affect argument interpretation Gupta et al. [2025b]. Because many evaluators assume structured inputs, robust *induction* of structure is a necessary precondition for equitable analysis.

3 Method

3.1 Task and Schema

We frame structuration as a two-stage pipeline. Stage 1 (*Extraction*) identifies distinct, non-overlapping points from raw threads. Stage 2 (*Classification*) assigns each point a content type and argumentative role. Outputs follow a constrained, line-based schema designed for easy parsing and downstream analysis.

Category	Description
Content Types	
BARRIER	A challenge, obstacle, or limitation that hinders progress.
SOLUTION	A proposed action, idea, or intervention to address a barrier.
METRIC	A criterion, measure, or indicator used to evaluate outcomes.
Argumentative Roles	
PRO	A statement that supports or defends a solution or proposal.
CON	A statement that opposes, criticizes, or raises doubts.

Table 1: Schema categories for structured deliberation. Each extracted point is assigned one TYPE and one ROLE.

3.2 Data and Annotation

We collected **1,200** discussion items from the Deliberatorium. Each item includes the original prompt, theme, and user statements. The **theme** denotes the overall discussion topic (e.g., *research productivity*), while the **prompt** is the specific question guiding contributions (e.g., *what are the barriers to increasing research productivity?*). Annotators produced gold labels in two steps: (i) writing distinct point extractions, and (ii) assigning each point a type and role under written guidelines. A held-out subset was reserved for development and evaluation.

3.3 Models and Training

We fine-tuned GPT-4o models for both stages using supervised training on the annotated data. The extractor was trained to output only `POINT:` lines; the classifier was trained to output `CLASSIFICATION:` and optional `CHILD:` lines. Training used standard next-token prediction with cross-entropy loss. A summary of splits is shown in Table 2.

Stage	Train Size	Dev Size	Test Size
Extractor	840	300	60
Classifier	840	300	60

Table 2: Split on Deliberatorium Dataset.

3.4 Prompt Design

We iteratively refined short, high-precision prompts. The extractor prompt emphasizes conciseness and non-redundancy; the classifier prompt encodes explicit rules to avoid common confusions (e.g., alternative solutions mislabeled as CON). Few-shot examples are minimal for clarity and consistency.

3.5 Inference

At inference, Stage 1 produces candidate points; Stage 2 assigns each point one content type and one argumentative role. Outputs are parsed into structured JSON for downstream evaluation and analysis.

3.6 Evaluation

We evaluate **CIVICPARSE** with GPT-4o models and report accuracy, precision, and recall. We analyze the confusion matrix to diagnose error patterns and to assess whether structural categories are preserved alongside stance. Table 3 defines the metrics and confusion-matrix components used throughout.

4 Results

Overall performance. We compare **CIVICPARSE** against an otherwise identical prompt-only pipeline using the same extraction and classification instructions. Prompt-only performance is

Metric	Interpretation
Accuracy	Overall fraction of correct predictions among all items
Precision	Proportion of predicted positives that are actually correct (TP / (TP + FP))
Recall	Proportion of actual positives correctly identified (TP / (TP + FN))
Confusion Matrix Components	
True Positives (TP)	Number of items correctly predicted as positive
False Positives (FP)	Number of items incorrectly predicted as positive (i.e., false alarms)
False Negatives (FN)	Number of items incorrectly predicted as negative (i.e., misses)
Predicted Positives (PP)	Total number of items the model labeled as positive

Table 3: Classification performance metrics used to evaluate **CIVICPARSE** . Precision, recall, and accuracy are derived from the confusion matrix composed of TP, FP, and FN.

substantially lower (overall accuracy ∼0.78), indicating that domain-specific adaptation improves reliability (Table 4). At the *item level*—requiring every point within an item to be correctly extracted and classified—accuracy improves from 0.781 to 0.882 (△0.101), with corresponding gains in precision and recall. At the *point level*—evaluating each extracted proposition independently—accuracy rises from 0.730 to 0.850 (△0.120). In the *binary extraction view*—a yes/no check of whether a point was correctly recovered—recall improves from 0.875 to 0.945 (△0.070) while precision increases slightly from 0.952 to 0.974 (△0.022). Together, the pipeline recovers more relevant points with less noise, yielding outputs that are easier to operationalize in moderation workflows.

Evaluation Setting	Precision	Recall	Accuracy
Item-level (mean $\pm$ std)	0.911 / 0.802 (△0.109)	0.965 / 0.854 (△0.111)	0.882 / 0.781 (△0.101)
Point-level (per point)	0.890 / 0.780 (△0.110)	0.950 / 0.830 (△0.120)	0.850 / 0.730 (△0.120)
Extraction (binary view)	0.974 / 0.952 (△0.022)	0.945 / 0.875 (△0.070)	—

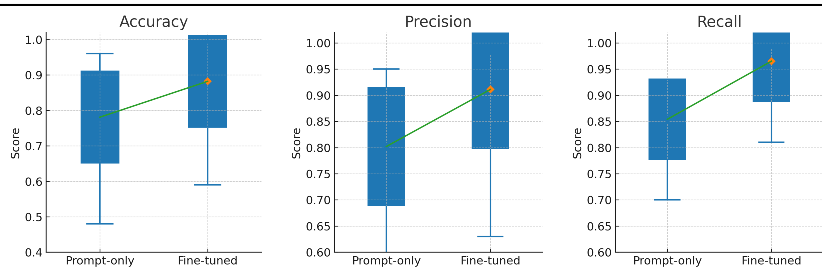

Table 4: **Overall results.** Fine-tuned vs. prompt-only performance across item-, point-, and binary-level evaluation. Top: numerical results. Bottom: graphical comparison (mean = center line, $\pm$1 std = box, min/max = whiskers). Accuracy and recall improvements are significant (Table 5); precision shows a positive but non-significant trend.

Statistical significance. To test robustness, we conduct paired t-tests on per-item metrics over 60 test items. Accuracy improvements are significant ($t(59) = 3.87$, $p = 2.77{\times}10^{-4}$), as are recall improvements ($t(59) = 3.41$, $p = 0.00119$). Precision shows a positive trend that is not significant at 0.05 ($t(59) = 1.80$, $p = 0.077$), indicating consistent gains in overall correctness and coverage.

Metric	Fine-tuned (M$\pm$SD)	Prompt-only (M$\pm$SD)	$t(59), p$
Accuracy	0.882 $\pm$ 0.131	0.788 $\pm$ 0.187	3.87, $2.77{\times}10^{-4}$
Precision	0.911 $\pm$ 0.114	0.876 $\pm$ 0.137	1.80, 0.077
Recall	0.965 $\pm$ 0.078	0.884 $\pm$ 0.171	3.41, 0.00119

Table 5: **Paired t-tests on per-item metrics** (60 test items). Accuracy and recall improvements are significant, while precision shows a positive but non-significant trend.

Extraction view. In the binary view, fine-tuning yields a favorable balance: recall rises by 7 points while precision remains very high. For moderators, this combination means broader coverage of relevant points without a commensurate increase in false positives.

Per-class breakdown. Table 6 shows that nearly all categories improve with fine-tuning. The largest gains occur for SOLUTION (+13) and BARRIER (+8), addressing common prompt-only confusions where proposals or obstacles are mislabeled as PRO. METRIC also improves (+7), reducing drift into argumentative roles. PRO sees moderate gains (+6), and CON—the most challenging class—improves substantially (+8, a 160% increase). These results indicate that the pipeline better preserves *structural* distinctions alongside stance, which is crucial for accurate representation of dissent.

True \ Pred	Barrier	Solution	Metric	Pro	Con	Row total
Barrier	52 / 44 (△8)	0 / 2 (△-2)	0 / 0	9 / 13 (△-4)	1 / 3 (△-2)	62
Solution	80 / 67 (△13)	80 / 72 (△8)	0 / 0	7 / 11 (△-4)	0 / 1 (△-1)	91
Metric	1 / 2 (△-1)	3 / 6 (△-3)	40 / 33 (△7)	1 / 3 (△-2)	0 / 1 (△-1)	45
Pro	3 / 6 (△-3)	2 / 4 (△-2)	4 / 5 (△-1)	37 / 31 (△6)	0 / 0	46
Con	3 / 5 (△-2)	5 / 7 (△-2)	0 / 1 (△-1)	5 / 8 (△-3)	13 / 5 (△8)	26
Col total	63 / 57	90 / 91	44 / 39	60 / 67	14 / 10	**271**

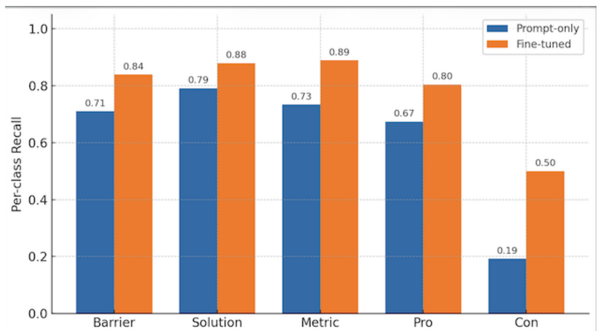

Table 6: **Fine-tuned vs. prompt-only confusion analysis.** Top: confusion matrix with counts (fine-tuned / prompt-only) and improvements in △. Bottom: per-class recall (from confusion matrix), showing consistent gains across all categories, especially CON.

5 Analysis

Error type	Representative example
Structure → stance	"UM6P students are intelligent given admission barriers" labeled as PRO instead of METRIC.
Polarity flip	"Homework is good" marked as CON instead of PRO.
Missed CON	"Paper count alone isn't meaningful" dropped or mapped to SOLUTION.
Hallucination	Spurious PRO created about a "Lydex" anecdote not in the gold data.
Residual bias	BARRIER→PRO drift in evaluative statements.

Table 7: **Representative errors.** Examples highlight how fine-tuning reduces but does not fully eliminate key error modes.

Error analysis. Quantitative gains coincide with qualitative improvements. Prompt-only predictions collapse structural categories into stance labels, especially mapping BARRIER/METRIC to PRO/CON. Fine-tuning preserves structural distinctions more faithfully and reduces polarity flips. CON remains hardest to recall, though recall improves materially—important for representing dissent. Prompt-only systems hallucinate unsupported points; fine-tuning curbs these false positives. A mild residual PRO bias persists when evaluative language accompanies proposals.

6 Conclusion

We recast structured moderation for online deliberation as a benchmarkable NLP task with civic importance and instantiate it via **CIVICPARSE** , a deliberation-native extractor–classifier pipeline.

Relative to an identical prompt-only setup, **CivicParse** yields consistent gains in accuracy and recall and reduces structural-to-stance collapses, polarity flips, and hallucinations. The strongest benefits appear in disentangling structural categories from stance, which supports fairer, more reliable large-scale analysis and moderation. Remaining challenges include improving Con recall and reducing residual Pro bias, as well as assessing transfer to other civic platforms. By providing a schema, dataset, and evaluation protocol, we aim to catalyze research on AI-assisted collective decision-making.

7 Limitations

Dataset size and scope. Training uses 840 annotated items from a single platform, limiting generalization across discourse styles and communities.

Model reliance on annotations. Performance depends on high-quality labels; ambiguities in guidelines (e.g., corrective proposals vs. opposition) introduce noise that constrains ceiling performance.

Residual stance bias. The model over-predicts Pro in some evaluative contexts, which can under-represent dissent.

Challenges with opposition. Con remains the weakest category; even with fine-tuning, oppositional statements can be reframed or missed.

Transferability. Evaluation focuses on the Deliberatorium; transfer to other civic platforms and less formal domains remains to be established.

Reproducibility Statement

We will release the 1,200-item annotated dataset and labeling guidelines upon acceptance. Preprocessing scripts, training code, and evaluation pipelines (Python) will be made publicly available, together with exact train/dev/test splits and metric definitions. Our experiments rely on widely accessible GPT-4o APIs and standard fine-tuning workflows; all training and evaluation outputs used in this study have been archived for verification. These resources are intended to support independent replication and extension.

References

T. Abdel-Monem, S. Bingham, J. Marincic, and A. Tomkins. Deliberation and diversity: Perceptions of small group discussions by race and ethnicity. *Small Group Research*, 41:746–776, 2010.

Lucas Anastasiou and Anna De Liddo. BCause: Human-AI collaboration to improve hybrid mapping and ideation in argumentation-grounded deliberation. *arXiv preprint arXiv:2505.03584*, 2025.

Y. Baek, M. Wojcieszak, and M. Carpini. Online versus face-to-face deliberation: Who? why? what? with what effects? *New Media & Society*, 14:363–383, 2012.

Michiel A. Bakker, Martin J. Chadwick, Hannah R. Sheahan, Michael Henry Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matthew M. Botvinick, and Christopher Summerfield. Fine-tuning language models to find agreement among humans with diverse preferences. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 13029–13042, 2022.

Maike Behrendt, Stefan S. Wagner, Marc Ziegele, Lena Wilms, Anke Stoll, Dominique Heinbach, and Stefan Harmeling. AQuA: Combining Experts' and Non-Experts' Views To Assess Deliberation Quality in Online Discussions Using LLMs. In *Proceedings of the 1st Workshop on Language-driven Deliberation Technology (DELITE)*, pages 1–12, 2024.

R. Burnell, H. Hao, A. Conway, and J. Orallo. Revealing the structure of language model capabilities. *arXiv*, abs/2306.10062, 2023.

K. Coe, K. Kenski, and S.A. Rains. Online and uncivil? patterns and determinants of incivility in newspaper website comments. *Journal of Communication*, 64(4):658–679, 2014.

J. Conklin and M. L. Begeman. gibis: A tool for all reasons. *Journal of the American Society for Information Science*, 40(3):200–213, 1989.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120, 2023. doi: 10.1073/pnas.2305016120.

Abhay Gupta, Philip Meng, Ece Yurtseven, Sean O'Brien, and Kevin Zhu. Aavenue: Detecting llm biases on nlu tasks in aave via a novel benchmark, 2024. URL https://arxiv.org/abs/2408.14845.

Abhay Gupta, Jacob Cheung, Philip Meng, Shayan Sayyed, Austen Liao, Kevin Zhu, and Sean O'Brien. Endive: A cross-dialect benchmark for fairness and performance in large language models, 2025a. URL https://arxiv.org/abs/2504.07100.

Abhay Gupta, Michael Lu, Kevin Zhu, Sean O'Brien, and Vasu Sharma. Novelhopqa: Diagnosing multi-hop reasoning failures in long narrative contexts, 2025b. URL https://arxiv.org/abs/2506.02000.

J. Huang, S. Gu, L. Hou, Y. Wu, X. Wang, H. Yu, and J. Han. Large language models can self-improve. *arXiv*, abs/2210.11610, 2022.

L. Iandoli, M. Klein, and G. Zollo. Can we exploit collective intelligence for collaborative deliberation? the case of the climate change collaboratorium. In *Proceedings of the International Conference on Digital Ecosystems and Technologies*, pages 127–132, 2007.

T. Ito, R. Hadfi, and S. Suzuki. An agent that facilitates crowd discussion. *Group Decision and Negotiation*, 31:621–647, 2022.

G. Karadzhov, T. Stafford, and A. Vlachos. Delidata: A dataset for deliberation in multi-party problem solving. *arXiv*, 2021.

R. Kennedy, A. Sokhey, C. Abernathy, K. Esterling, D. Lazer, A. Lee, W. Minozzi, and M. Neblo. Demographics and (equal?) voice: Assessing participation in online deliberative sessions. *Political Studies*, 69:66–88, 2020.

M. Klein. Enabling large-scale deliberation using attention-mediation metrics. *Computer Supported Cooperative Work*, 21(4–5):449–473, 2012.

M. Klein and N. Majdoubi. The medium is the message: toxicity declines in structured vs unstructured online deliberations. *World Wide Web*, 27(1):31, 2024.

J. Lawrence and C. Reed. Argument mining: A survey. *Computational Linguistics*, 45:765–818, 2020.

M. Lippi and P. Torroni. Argumentation mining, 2018.

Naseem Machlovi, Maryam Saleki, Innocent Ababio, and Ruhul Amin. Towards safer ai moderation: Evaluating llm moderators through a unified benchmark dataset and advocating a human-first approach. *arXiv preprint arXiv:2508.07063*, 2025.

A. Møller. Deliberation and deliberative organizational routines in frontline decision-making. *Journal of Public Administration Research and Theory*, 2021.

S. Niemeyer, F. Veri, J. Dryzek, and A. Bächtiger. How deliberation happens: Enabling deliberative reason. *American Political Science Review*, 2023.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

M. Pietron, R. Olszowski, and J. Gomułka. Efficient argument classification with compact language models and chatgpt-4 refinements. Technical report, AGH University of Krakow, 2023.

R. Rega and R. Marchetti. The strategic use of incivility in contemporary politics. the case of the 2018 italian general election on facebook. *Communication Review*, 24(2):107–132, 2021.

A. Sengoku, T. Ito, K. Takahashi, S. Shiramatsu, T. Ito, E. Hideshima, and K. Fujita. Discussion tree for managing large-scale internet-based discussions. In *Collective Intelligence*, 2016.

Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, Lilach Edelstein, Liat Ein-Dor, Roni Friedman-Melamed, Assaf Gavron, Ariel Gera, Martin Gleize, Shai Gretz, Dan Gutfreund, Alon Halfon, Daniel Hershcovich, Ron Hoory, Yufang Hou, Shay Hummel, Michal Jacovi, Charles Jochim, Yoav Kantor, Yoav Katz, David Konopnicki, Zvi Kons, Lili Kotlerman, Dalia Krieger, Dan Lahav, Tamar Lavee, Ran Levy, Naftali Liberman, Yosi Mass, Amir Menczel, Shachar Mirkin, Guy Moshkowich, Shila Ofek-Koifman, Matan Orbach, Ella Rabinovich, Ruty Rinott, Slava Shechtman, Dafna Sheinwald, Eyal Shnarch, Ilya Shnayderman, Aya Soffer, Artem Spector, Benjamin Sznajder, Assaf Toledo, and IBM Research AI. An autonomous debating system. *Nature*, 591(7850):379–384, 2021. doi: 10.1038/s41586-021-03215-w.

C.T. Small, I. Vendrov, E. Durmus, H. Homaei, E. Barry, J. Cornebise, T. Suzman, D. Ganguli, and C. Megill. Opportunities and risks of llms for scalable deliberation with polis. Technical report, The Computational Democracy Project and Anthropic, 2023.

C. R. Sunstein. *Infotopia: How Many Minds Produce Knowledge*. Oxford University Press, 2006.

D. Tapscott and A. D. Williams. *Wikinomics: How Mass Collaboration Changes Everything*. Portfolio, 2006.

Michael Henry Tessler, Michiel A. Bakker, Daniel Jarrett, Hannah Sheahan, Martin Chadwick, Raphael Koster, Georgina Evans, Lucy Campbell-Gillingham, Teddy Collins, David C. Parkes, Matthew Botvinick, and Christopher Summerfield. Ai can help humans find common ground in democratic deliberation. *Science*, 386(6719):eadq2852, 2024. doi: 10.1126/science.adq2852.

D. Volkovskii and O. Filatova. Low civility and high incivility in russian online deliberation. *KOME – An International Journal of Pure Communication Inquiry*, 11(1):95–109, 2023.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou. Chain-of-thought prompting elicits reasoning in large language models. *arXiv*, abs/2201.11903, 2022.

Lilian Weng. Prompt engineering. `https://lilianweng.github.io/posts/2023-03-15-prompt-engineering/`, 2023.

S. N. Williams. A twenty-first century citizens polis: Introducing a democratic experiment in electronic citizen participation in science and technology decision-making. *Public Understanding of Science*, 19(5):528–544, 2010.

M. Wojcieszak. Deliberation and attitude polarization. *Journal of Communication*, 61:596–617, 2011.

A. Wynter and T. Yuan. I wish to have an argument: Argumentative reasoning in large language models. *arXiv*, abs/2309.16938, 2023.

Xianyang Zhan, Agam Goyal, Yilun Chen, Eshwar Chandrasekharan, and Koustuv Saha. SLM-Mod: Small Language Models Surpass LLMs at Content Moderation. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 8774–8790, 2025.

W. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J. Nie, and J. Wen. A survey of large language models. *arXiv*, 2023.