Procedural Parity, Outcome Mismatch: Evaluating Human vs LLM Deliberation

Maurice Flechtner
University of Zurich
Zürich, Switzerland
ETH Zurich
Zürich, Switzerland
maurice.flechtner@zda.uzh.ch

Abstract

As HCI increasingly studies AI-mediated and *human–AI* deliberation in civic contexts, a central open question is whether LLM agents possess the deliberative capabilities that such applications assume. We introduce *DelibSim*, a simulation framework that evaluates LLM-only multi-agent deliberation using two political science measures—AQuA (an automated Discourse Quality Index) for procedural quality and the Deliberative Reason Index (DRI) for epistemic outcomes—comparing both against human baselines from real deliberative processes. Across 1,980 small-group deliberations spanning 11 model configurations and 12 policy topics, LLMs reach near-human procedural quality (AQuA = 2.94 vs. 2.98), but only exhibit small and topic-dependent outcome gains. Normative prompting produces a modest increase in absolute ΔDRI relative to a survey-only baseline (ATE = 0.029, $p_{\text{Holm}} = 0.0146$), while basic prompting yields no robust improvement. Compared to human groups, LLM gains are smaller and display much lower perspective heterogeneity. These results suggest that similar procedural quality can mask fundamentally different outcome dynamics, raising questions about the interchangeability of LLM and human deliberation for civic applications.

CCS Concepts

• **Computing methodologies** → *Agent / discrete models*; Simulation evaluation; Model verification and validation; Natural language generation; • **Applied computing** → **Sociology**.

Keywords

Multi-Agent, LLM, Deliberation, Evaluation, Simulation

1 Introduction

Large Language Models (LLMs) are increasingly applied in deliberative settings where arguments are exchanged and reflected upon in order to coherently discuss a given issue. Current applications propose LLMs in roles as diverse as AI-assistants for deliberation [19], reflection helpers [29, 31], representatives of missing perspectives in policy discussions [7, 32], and providers of voting advice [12, 31]. However, there is little evidence that LLMs possess the epistemic capabilities these roles assume. Deliberation demands justification, reciprocity, and the weighing of competing values across diverse contexts [3, 6, 9]—capabilities that cannot be inferred from isolated reasoning benchmarks. Crucially, empirical evidence shows that *procedural parity does not imply deliberative success*: similar discourse quality can mask fundamentally different outcome and diversity dynamics [1, 14].

To address this gap, we introduce *DelibSim*, a controlled testbed for evaluating LLM-only multi-agent deliberation using metrics from political science. DelibSim is not itself a human-facing deliberation tool. Rather, it evaluates whether LLMs possess the deliberative capabilities that downstream human–AI applications assume. We adapt two indices: AQuA [2], an automated form of the Discourse Quality Index (DQI) [27] for procedural quality, and the Deliberative Reason Index (DRI) [22], measuring intersubjective consistency (IC) for outcome quality. Across 1,980 small-group sessions spanning 11 model configurations and 12 citizen-assembly topics, we find that LLMs can *perform* deliberation procedurally—i.e. producing high-quality discourse—while achieving at most borderline epistemic gains in IC.

We contribute (1) evidence that procedural discourse quality does not guarantee deliberative outcomes in LLM deliberation, providing a process–outcome evaluation lens for civic human–AI systems, and (2) *DelibSim* as a testbed for systematically exploring design parameters such as prompting regimes, model selection, and topic characteristics.

2 Related Work

Online deliberation has been a long-standing interest of the HCI community [10, 15, 16, 18, 25, 26]. Building on this, researchers increasingly consider LLMs for deliberative applications including assistance [19], moderation [13], reflection support [29, 31], and representing missing perspectives [7, 32]. Apart from normative concerns about deploying LLMs in civic deliberation [4] and potential legitimacy costs [11], it has not been established whether LLMs can deliberate in any epistemically meaningful way. *DelibSim* makes deliberative capacity measurable by evaluating multi-agent LLM discussions using established deliberation metrics.

Assessing Deliberation. We define deliberation as reasoned discourse among free and equal individuals [5], in which participants exchange arguments, weigh competing considerations, and may revise their understanding [3, 6, 9]. Procedural conditions—reason-giving, reciprocity, mutual respect—have traditionally been assessed with the DQI [27]. We operationalise these using *AQuA*, an automated DQI adaptation based on pre-trained BERT adapters [2]. However, procedural quality alone is insufficient: high-quality discourse does not reliably translate into stronger deliberative outcomes [14] and may even have detrimental effects [1].

This paper was prepared for PoliSim@CHI 2026: LLM Agent Simulation for Policy, a workshop at CHI 2026 (CHI Conference on Human Factors in Computing Systems), April 16, 2026, Barcelona, Spain.

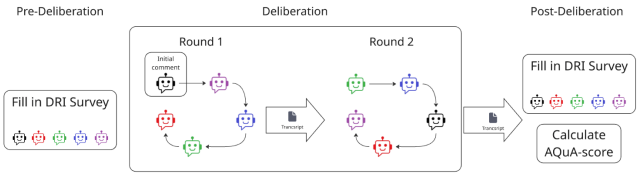


Figure 1: DelibSim design scheme. The simulation proceeds in three phases: (1) Pre-Deliberation DRI survey, (2) multi-agent deliberation, and (3) Post-Deliberation DRI survey. The full deliberation transcript is passed between agents for continuous context. Speech order is randomized within rounds.

The Deliberative Reasoning Index. As the traditional idea of consensus as the ideal outcome of deliberation has been widely criticized [20, 24, 30], DRI [22] operationalises good deliberation outcomes through improvements in IC instead. Theoretically, this relies on the argument that successful deliberation produces *meta-consensus* among participants, i.e. a shared understanding of what is relevant to the issue at hand and how different aspects relate to one another [21, 22]. IC is the empirical signature of this shared understanding: when a group has developed high IC, pairs agree proportionally on considerations (values, arguments) and on preferences (policy options). Importantly, participants can disagree while exhibiting high IC, so long as their disagreements are consistent across considerations and preferences [22]. IC is measured using a two-stage survey. Participants rate consideration statements C and rank preference statements P. DRI is the average pairwise Spearman correlation between C and P agreement, and ΔDRI (post minus pre) is our outcome of interest.[1]

As a group-level relational property [23], DRI does not imply a single correct solution. Instead, it captures whether participants' reasons and preferences cohere in a shared justificatory space. In the context of LLM-deliberation this allows us to measure something that is substantially different from task performance: The ability to exchange and incorporate distinct viewpoints — an important and necessary step in harnessing the epistemic qualities of deliberation [17].

DRI has been applied across 19 human deliberative forums, with control groups confirming that improvements require deliberation content rather than repeated survey exposure [23]. Concurrent work by Umbelino and Veri [28] benchmarks 54 LLMs against human DRI baselines, finding that most LLMs exhibit lower individual-level DRI than post-deliberation humans—complementary to our group-level analysis of DRI *changes* through LLM deliberation.

3 Method

The DelibSim Framework. DelibSim simulates deliberation among LLM agents in a turn-based manner. As illustrated in Figure 1, Groups of 5 agents deliberate for 2 rounds, with speaking order randomized within rounds. The full transcript passes between agents, building shared context. This design aims to balance group size against coordination costs while ensuring sufficient context and dyads for DRI computation ($\binom{5}{2} = 10$ pairs per group). Each agent receives the same system prompt and an updated conversation history, when prompted. The framework supports two model configurations: *homogeneous* groups with all agents using the same base LLM and *mixed-model* groups where each agent uses a different base LLM. This allows us to both examine differences between models and the impact of increased diversity in *mixed-model* contexts.

Treatment Conditions. We evaluate three conditions. The *No Treatment* condition serves as a survey-only baseline: groups complete pre- and post-deliberation DRI surveys with no discussion, establishing baseline ΔDRI variation. This controls for survey-induced artifacts such as deterministic decoding—analogous to control groups in human DRI validation studies (e.g., Uppsala Speaks, AusCJ) that similarly show no DRI improvement without deliberation [23]. Baseline noise remains substantial even at temperature = 0.0 (SD ≈ 0.17; within-block SD ≈ 0.11), reflecting residual instability in survey outputs.[2] The *Basic Treatment* provides minimal intervention: agents are instructed to deliberate, relying on the model's internal understanding of "deliberation." The *Normative Treatment* adds explicit deliberative guidance, explicitly mentioning respect, reason-giving, authenticity, engagement with opposing viewpoints to approximate a normatively structured deliberation. This design tests whether deliberation improves IC beyond baseline noise and whether explicit prompting yields additional benefits.[3]

Models. We evaluate 11 model configurations across 5 frontier families: GPT-5.1, Gemini-3-Pro-Preview, DeepSeek-V3.2-Exp, Kimi-K2-Thinking, and Claude Opus 4.5. Apart from the Claude model, we include both reasoning-enabled and standard variants. Additionally, we add two mixed-model ensembles with one agent per family each. One with reasoning-enabled and one with standard variants.

Unit of Analysis and Inference. Each observation is a *run*: a 5-agent group deliberating on one policy topic under one model configuration and one condition. Across 12 topics and 11 model configurations, we generate 1,980 runs (660 per condition) arranged in 132 topic × model blocks. Each run consists of three phases: (1) pre-deliberation DRI survey, where each agent independently rates considerations and ranks preferences via a structured prompt (responses are parsed and validated for completeness, valid ranges, and unique rankings), (2) deliberation, where agents engage in structured turn-based discussion, and (3) post-deliberation DRI survey, where agents update their responses after reviewing the full transcript. Temperature is set to 0.0 to maximize reproducibility.

Estimand and Dependence-Aware Inference. Our primary estimand is the topic × model-blocked average treatment effect (ATE) on ΔDRI: for each block b, we compute the mean outcome difference $\bar{Y}_{b,1} - \bar{Y}_{b,0}$ between two treatment conditions and then average these differences with equal weight across the 132 blocks. Treatment effects are tested via blocked permutation tests that shuffle condition labels *within* each topic × model block while preserving group sizes. Confidence intervals use a topic-resampled bootstrap (resampling the 12 topics with replacement and recomputing the blocked ATE)

[1] See Appendix A for calculation details.

[2] See Appendix B for details.

[3] See Appendix C for full prompts.

Table 1: Pooled treatment effects on absolute and relative ΔDRI using topic×model-blocked contrasts. Point estimates are averages over (topic×model) blocks; p-values come from blocked permutation tests preserving group sizes within blocks; confidence intervals use topic-resampled bootstrap.

Outcome	Contrast	ATE	95% CI	p	p_{Holm}
Absolute ΔDRI	Normative vs. No Treatment	0.029	[−0.004, 0.076]	0.0049	0.0146
	Basic vs. No Treatment	0.019	[−0.011, 0.057]	0.0791	0.1583
	Normative vs. Basic	0.010	[−0.008, 0.030]	0.3493	0.3493
Relative ΔDRI	Normative vs. No Treatment	0.066	[0.032, 0.107]	< 0.001	< 0.001
	Basic vs. No Treatment	0.061	[0.024, 0.100]	< 0.001	< 0.001
	Normative vs. Basic	0.005	[−0.027, 0.037]	0.723	0.723

Table 2: Hierarchical mixed-effects model for absolute ΔDRI across all LLM runs ($N = 1980$). Fixed effects are reported relative to the No Treatment baseline. The model includes random intercepts, treatment slopes by model and a topic random intercept (variance component). 95% CIs are parametric bootstrap percentile intervals ($B = 2000$).

Fixed effect	Estimate	SE	p	95% CI (bootstrap)
Intercept (No Treatment)	0.002	0.012	0.861	[−0.022, 0.026]
Basic vs. No Treatment	0.019	0.015	0.213	[−0.012, 0.047]
Normative vs. No Treatment	0.029	0.013	0.025	[0.004, 0.054]

to reflect limited topic-level replication and topic clustering ($n = 12$). We apply Holm correction within each outcome family of contrasts. As a complementary robustness specification, we fit a hierarchical mixed-effects model to all runs with a topic random intercept and model-level random treatment slopes. Uncertainty is quantified with a parametric bootstrap (B=2000).

AQuA. We assess discourse quality using AQuA [2], computing a weighted aggregate across 20 dimensions (justification, respect, engagement, etc.) on a 4-point scale. AQuA is trained on German data. We follow the English→German translation pipeline prescribed by its developers for English-language application [2, Section 3.5]. Within-LLM comparisons are unaffected by translation since all conditions undergo identical processing. The same is true for the human baseline. Here, we use the Europolis dataset [8] —the same benchmark used by AQuA's developers for validation (AQuA = 2.98, $N = 910$ comments). It consists of english transcripts from a deliberative poll, that undergo exactly the same treatment as LLm transcripts. However, LLM-generated text is systematically more structured than natural discussion, which may inflate certain AQuA dimensions. The near-identical aggregate scores should therefore be interpreted as approximate parity.[4]

DRI. DRI is measured through pre/post surveys where agents rate consideration statements and rank policy preferences. All LLM answers are checked for consistency. DRI captures IC in how considerations relate to preferences. Positive ΔDRI (post minus pre) indicates successful deliberation. We hold the DRI survey instrument (wording, scale anchors, output format) constant across conditions, so condition contrasts in ΔDRI are not attributable to survey prompt differences. However, absolute DRI levels may be sensitive to alternative survey prompt formats, which we do not evaluate here.

[4]See Appendix F for full AQuA analysis and discussion of register effects.

Topics and Human Benchmarks. We use 12 policy topics from real citizen assemblies covering climate policy, healthcare, governance, bioethics, and urban planning.[5] Human DRI data from these actual deliberative events enable direct comparison. Human samples range from 11 to 63 participants per topic, with a mean of 33.92 participants and DRI computed from pre- and post-deliberation surveys administered before and after deliberative interventions. As we do not have transcripts of these deliberative processes to assess process quality through AQuA, our human baseline stems from transcripts of a deliberative poll available in the Europolis dataset [8].

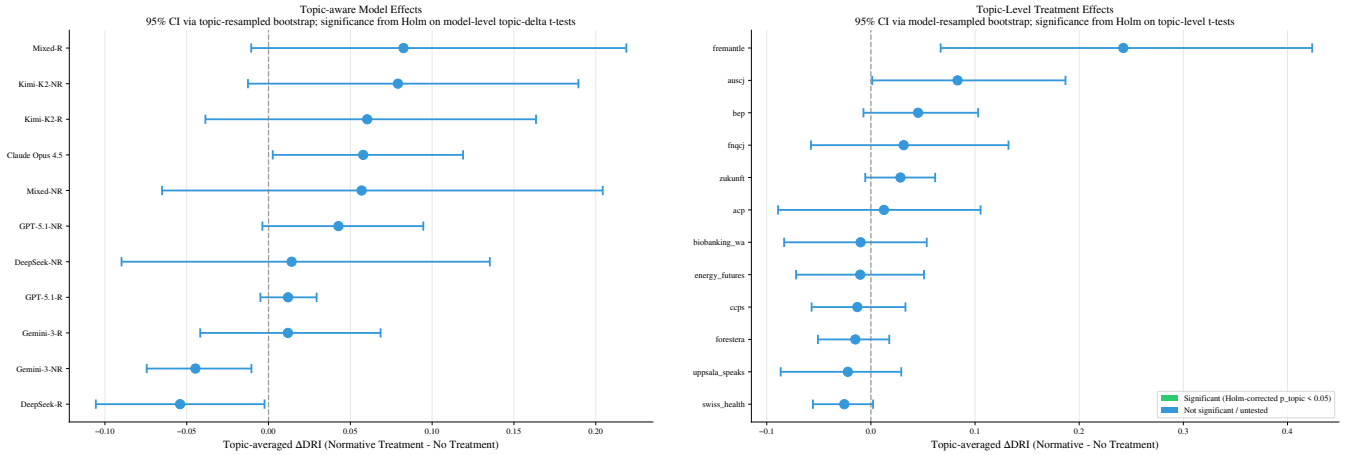
Opinion Diversity. We operationalise opinion diversity as the mean pairwise Euclidean distance between participants' standardized survey response vectors, measuring perspective heterogeneity within groups.[6]

4 Results

Procedural Discourse Quality Is Comparable to Humans. LLM groups achieve discourse quality scores close to human deliberation. Normative and basic treatments yield identical mean AQuA scores (both 2.939), suggesting that explicit deliberative guidance does not meaningfully affect *procedural* quality. Compared to humans (SD = 0.431), LLM discourse exhibits substantially lower variance (SD = 0.12–0.13), indicating more consistent but potentially less diverse contributions.[7]

Topic Dependence Dominates. Deliberative outcomes exhibit substantial dependence on policy topic. The ICC for absolute ΔDRI by topic is 0.107 (95% CI ≈ [−0.001, 0.160]), indicating meaningful clustering at the topic level and motivating topic-aware clustered inference throughout.

[5]See Appendix D for full topic overview.
[6]See Appendix E for exact definition.
[7]See Appendix F for full analysis.



(a) Model-level treatment effects **(b) Topic-level treatment effects**

Figure 2: Heterogeneity in treatment effects on absolute ΔDRI. (a) Model-level effects averaged across topics. (b) Topic-level effects averaged across models. Error bars show bootstrap confidence intervals.

Main Outcome: Normative Guidance Improves Absolute ΔDRI. Only the normative treatment yields a small but significant increase in absolute ΔDRI relative to no treatment. The bootstrap CI is wide and marginally overlaps zero, reflecting limited topic-level replication.[8] We additionally report relative ΔDRI as a robustness check under strong baseline differences (Table 1).

Hierarchical mixed model corroboration. A hierarchical mixed model with random intercepts by topic and random slopes by model fitted to all LLM runs yields a consistent pattern: the normative treatment has a positive fixed effect of approximately 0.030 on absolute ΔDRI (95% bootstrap CI $\approx [0.004, 0.054]$, $p \approx 0.025$), whereas the basic treatment is smaller and not reliably different from zero (Table 2).

4.1 Model and Topic Heterogeneity

Model heterogeneity. Treatment responsiveness varies substantially across models, but no model family shows a robustly large positive effect once topic structure and multiple comparisons are accounted for. Some models show positive point estimates (e.g., Claude Opus 4.5), while others exhibit negative effects (notably Gemini and DeepSeek variants). Reasoning-enabled vs. standard variants show no robust advantage (mean difference 0.010, 95% CI $[-0.027, 0.044]$), paralleling Umbelino and Veri [28], who found no clear link between chain-of-thought capability and deliberative performance across 54 models.

Topic heterogeneity. Treatment effects are strongly topic dependent. Positive effects cluster around concrete, local topics (Fremantle bridge: ≈ 0.24; Bloomfield Track) where the consideration-preference mapping is relatively straightforward. Near-zero or negative effects appear on more abstract, value-laden topics (Swiss Health ≈ −0.026; Australian governance; genome editing). This pattern is consistent with findings from human DRI research, where issue *complexity* attenuates DRI improvement unless offset by group building [23]—a mechanism LLM deliberation lacks. Notably, Fremantle shows the largest positive effect in both human (ΔDRI = 0.127) and LLM deliberation, suggesting that when issue structure is amenable and unambiguous, LLM deliberation can produce meaningful IC gains.

4.2 Human Benchmark

Compared to human deliberation, LLM groups show substantially smaller absolute gains in IC. Humans improve by ΔDRI ≈ 0.099 (95% CI $\approx [0.026, 0.196]$), whereas LLM groups improve by 0.031 (Normative), 0.021 (Basic), and 0.002 (No Treatment). At the same time, conditioning on baseline reveals different results. In a Monte-Carlo ANCOVA (post-DRI on pre-DRI with disjoint 5-person human groupings), the LLM indicator is strongly positive ($\beta_{\text{LLM}} \approx 0.30$, 95% CI $\approx [0.18, 0.41]$), implying higher post-deliberation consistency for a fixed pre-DRI. This can be explained by baseline differences: LLMs start from much higher pre-DRI (≈ 0.78) than humans (≈ 0.30), consistent with initial agreement rather than deliberation-induced understanding.

Ceiling Effects. We find strong ceiling effects. ΔDRI declines sharply with pre-DRI (pooled LLM $\beta \approx -0.50$, $R^2 \approx 0.36$, $p < 0.001$), motivating a head-/floorroom-adjusted outcome:

$$\Delta\text{DRI}_{\text{rel}} = \begin{cases} \dfrac{\Delta\text{DRI}}{1 - \text{pre-DRI}} & \text{if } \Delta\text{DRI} \geq 0, \\ \dfrac{\Delta\text{DRI}}{1 + \text{pre-DRI}} & \text{if } \Delta\text{DRI} < 0. \end{cases}$$

Under this normalization, treatment effects are clearly detectable (Normative vs. No Treatment: ATE = 0.066, $p_{\text{Holm}} < 0.001$; Basic vs. No Treatment: ATE = 0.061, $p_{\text{Holm}} < 0.001$), indicating that, relative

[8] Permutation p-values test the sharp null under the blocked design, whereas the bootstrap CI reflects uncertainty under limited topic-level replication ($n = 12$). These can disagree when the effect is small and topic heterogeneity is substantial.

to available headroom, LLMs can make significant epistemic gains from deliberation.[9]

Perspective Diversity. Finally, diversity dynamics differ sharply from humans. Human groups start much more heterogeneous (Euclidean distance 18.78 vs. ≈6.5) and tend to *converge* through deliberation ($\Delta \approx -1.21$). LLM groups begin with highly similar perspectives and tend to remain stable or *diverge* slightly (Basic: +0.37; Normative: +0.26).

5 Discussion

LLM groups reached near-human procedural discourse quality but did not reliably reproduce human-like outcome dynamics: absolute ΔDRI gains were small, topic-dependent, and only significant under normative prompting. We conclude that LLMs can *perform* deliberation without consistently achieving its characteristic outcome. This process–outcome disconnect has precedent in human deliberation research [1, 14] but takes an extreme form with LLMs. When normalizing by available headroom, treatment effects become significant ($p < 0.001$) and LLMs show relative improvement (0.195) exceeding humans (0.138). However, the theoretical motivation for DRI concerns *absolute* improvements in shared understanding, not proportional gains from already-high baselines.

Shared Representations Without Deliberation. The DRI literature provides a lens for interpreting these patterns. In human deliberation, DRI improvements arise when diverse participants construct a *shared representational framework*—a mutual logic connecting considerations to preferences—through engagement with different perspectives [17, 22]. This requires both *meta-consensus* (agreement on what considerations are relevant) and *integration* (incorporating diverse viewpoints into reasoning) [21, 23]. LLMs, trained on overlapping data, arrive with pre-formed shared representations: they already share meta-consensus and exhibit high baseline IC (≈0.78 vs. human ≈0.30). With substantially lower opinion diversity (∼6.5 vs. ∼18.8), the diversity-driven representational updating that produces human DRI gains is largely absent. LLMs begin where humans hope to arrive through deliberation, but this coherence emerges from training similarity, not from genuine engagement with different perspectives.

DRI Construct Validity in LLM Contexts. A concern is whether the LLMs' DRI improvements reflect deliberative reasoning or response consistency under deterministic decoding. As we cannot ultimately distinguish whether DRI improvements reflect genuine representational updating or pattern-matching to transcript content for LLMs, we treat ΔDRI as internally valid for our fixed instrument and publish full prompts to support replication and future robustness checks. However, the No Treatment baseline ($\Delta\text{DRI} \approx 0.002$, $\text{SD} \approx 0.175$) provides some insights into this question: It establishes that repeating the survey without deliberation does not systematically shift IC, even though decoding and survey format are identical across conditions. Even at temperature=0, within-block $\text{SD} \approx 0.11$ shows that context (transcript vs. no transcript) drives meaningful variation in reason-preference coherence, indicating that LLM responses are not trivially deterministic.

[9] See Table 1 for results.

Limitations. Several limitations accompany our conclusions. The topic level inference with $n = 12$ provides limited statistical power. Our fixed protocol (5 agents, 2 rounds, no persona diversity, temperature = 0) may not generalize to longer deliberations, different group sizes, or more diverse agent configurations. The ceiling effect complicates human–LLM comparability. We do not test measurement invariance of DRI under alternative survey prompts (e.g., different scale framing, item order, or output formats), so we cannot fully exclude the possibility of prompt-induced DRI convergence. The AQuA-based human comparison involves a register asymmetry: LLM text is more structured than natural discussion, potentially inflating rationality-related dimensions. Finally, because outcomes vary strongly by topic the pooled treatment effect should not be interpreted as uniformly generalizing to all topics. Rather, it describes average performance across a deliberately broad topic set.

Future Work. Future work should investigate: (1) the impact of increased perspective diversity (e.g., through persona prompting) on deliberative outcomes, (2) which topics, prompts and protocols facilitate LLM deliberation, and (3) hybrid human–LLM deliberation where LLMs encounter genuine disagreement.

6 Conclusion

We introduced *DelibSim*, a simulation framework for evaluating multi-agent LLM deliberation using theory-grounded political science metrics. Across 1,980 small-group sessions spanning 11 model configurations and 12 policy topics, LLM groups produce discourse that matches human deliberation in procedural quality, yet absolute improvements in intersubjective consistency remain small, topic-dependent, and only reliably positive under normative prompting. On complex and value-laden topics, gains attenuate and can even reverse. While relative improvements are detectable once ceiling effects are accounted for, these reflect coherence within already homogeneous agent populations rather than the diversity-driven representational updating characteristic of human deliberation.

Many civic applications infer epistemic contribution from the presence of respectful, reasoned dialogue. Our findings show that, similarly to human deliberation, such surface-level deliberative performance does not guarantee epistemic gains in shared understanding among LLM agents. Systems that appear deliberative may still fail to generate the meta-consensus that gives deliberation its democratic value. At the same time, we do not find evidence that LLMs are incapable of deliberative reasoning altogether. Rather, their performance is strongly conditioned by topic structure, prompting regime, and baseline diversity. This suggests that deliberative capacity is not an intrinsic model property but emerges from the interaction between model characteristics and process design.

Beyond empirical findings, *DelibSim* contributes an evaluation scaffold for policy-relevant simulations. It provides a controlled environment for systematically testing prompting strategies, model ensembles, and various manipulations before deploying LLMs in civic contexts. In this sense, we position deliberation simulation not as a replacement for human deliberation, but as a diagnostic instrument for assessing the epistemic affordances and limits of AI-mediated participation.

Future work should investigate whether introducing genuine perspective heterogeneity, through persona diversification, mixed-model ensembles, or hybrid human–AI groups, enables the diversity-driven dynamics that underlie human DRI gains. More broadly, evaluating AI systems intended for democratic settings requires outcome-sensitive metrics grounded in deliberative theory. Measuring discourse quality alone is insufficient.

Under the conditions tested, LLMs can convincingly *perform* deliberation without consistently achieving its epistemic core. Distinguishing between these two remains essential for responsible design of civic AI systems.

Acknowledgments

This research was supported by the Horizon Europe project *AI 4 Deliberation* and received institutional funding from the Department of Political Science (IPZ) at the University of Zurich and ETH Zurich. We gratefully acknowledge this support.

Generative AI tools were used during the coding process to assist with code writing and debugging and during the writing provess to improve phrasing and clarity of the manuscript. All AI-assisted outputs were reviewed, tested, and edited by the author, who takes full responsibility for the content, analyses, and conclusions presented in this paper.

References

[1] Lucio Baccaro, André Bächtiger, and Marion Deville. 2016. Small Differences That Matter: The Impact of Discussion Modalities on Deliberative Outcomes. *British Journal of Political Science* 46, 3 (July 2016), 551–566. doi:10.1017/S0007123414000167

[2] Maike Behrendt, Stefan Sylvius Wagner, Marc Ziegele, Lena Wilms, Anke Stoll, Dominique Heinbach, and Stefan Harmeling. 2024. AQuA – Combining Experts' and Non-Experts' Views To Assess Deliberation Quality in Online Discussions Using LLMs. arXiv:2404.02761 [cs]

[3] Simone Chambers. 2003. Deliberative Democratic Theory. *Annual Review of Political Science* 6, 1 (June 2003), 307–326. doi:10.1146/annurev.polisci.6.121901.085538

[4] Mark Coeckelbergh. 2025. Llms, Truth, and Democracy: An Overview of Risks. *Science and Engineering Ethics* 31, 1 (2025), 1–13. doi:10.1007/s11948-025-00529-0

[5] Joshua Cohen. 2002. Deliberation and Democratic Legitimacy. In *Debates in Contemporary Political Philosophy: An Anthology*, Derek Matravers and Jonathan E. Pike (Eds.). Routledge.

[6] John S. Dryzek. 2002. *Deliberative Democracy and Beyond: Liberals, Critics, Contestations*. Oxford University Press.

[7] Suyash Fulay, Dimitra Dimitrakopoulou, and Deb Roy. 2025. The Empty Chair: Using LLMs to Raise Missing Perspectives in Policy Deliberations. arXiv:2503.13812 [cs] doi:10.48550/arXiv.2503.13812

[8] Marlène Gerber, André Bächtiger, Susumu Shikano, Simon Reber, and Samuel Rohr. 2018. Deliberative Abilities and Influence in a Transnational Deliberative Poll (EuroPolis). *British Journal of Political Science* 48, 4 (Oct. 2018), 1093–1118. doi:10.1017/S0007123416000144

[9] Robert E. Goodin. 2003. *Reflective Democracy* (1 ed.). Oxford University PressOxford. doi:10.1093/0199256179.001.0001

[10] Ian G. Johnson, Alistair MacDonald, Jo Briggs, Jennifer Manuel, Karen Salt, Emma Flynn, and John Vines. 2017. Community Conversational: Supporting and Capturing Deliberative Talk in Local Consultation Processes. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. Association for Computing Machinery, New York, NY, USA, 2320–2333. doi:10.1145/3025453.3025559

[11] Andreas Jungherr and Adrian Rauchfleisch. 2025. Artificial Intelligence in Deliberation: The AI Penalty and the Emergence of a New Deliberative Divide. arXiv:2503.07690 [cs] doi:10.48550/arXiv.2503.07690

[12] Naomi Kamoen and Christine Liebrecht. 2022. I Need a CAVAA: How Conversational Agent Voting Advice Applications (CAVAAs) Affect Users' Political Knowledge and Tool Experience. *Frontiers in Artificial Intelligence* 5 (May 2022), 835505. doi:10.3389/frai.2022.835505

[13] Mark Klein, Ibukun Babatunde, and Obiabuchi Nnanna. 2025. Moderating Large Scale Online Deliberative Processes with Large Language Models (LLMs): Enhancing Collective Decision-Making. social science research network:5171687

[14] Katherine Knobloch and John Gastil. 2022. How Deliberative Experiences Shape Subjective Outcomes: A Study of Fifteen Minipublics from 2010-2018. *Journal of Deliberative Democracy* 18, 1 (March 2022). doi:10.16997/jdd.942

[15] Travis Kriplean, Jonathan Morgan, Deen Freelon, Alan Borning, and Lance Bennett. 2012. Supporting Reflective Public Thought with Considerit. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*. ACM, Seattle Washington USA, 265–274. doi:10.1145/2145204.2145249

[16] Tzu-Sheng Kuo, Quan Ze Chen, Amy X. Zhang, Jane Hsieh, Haiyi Zhu, and Kenneth Holstein. 2025. PolicyCraft: Supporting Collaborative and Participatory Policy Design through Case-Grounded Deliberation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–24. doi:10.1145/3706598.3713865

[17] Hélène Landemore. 2013. Deliberation, Cognitive Diversity, and Democratic Inclusiveness: An Epistemic Argument for the Random Selection of Representatives. *Synthese* 190, 7 (May 2013), 1209–1231. doi:10.1007/s11229-012-0062-6

[18] Kristine J. Lu, Gustavo K. Umbelino, Spencer E. Carlson, and Matthew W. Easterday. 2025. DeliberationWorks: A Deliberation System for Developing Capacities in Civic Organizing. *Proc. ACM Hum.-Comput. Interact.* 9, 2 (May 2025), CSCW060:1–CSCW060:29. doi:10.1145/3710958

[19] Shuai Ma, Qiaoyi Chen, Xinru Wang, Chengbo Zheng, Zhenhui Peng, Ming Yin, and Xiaojuan Ma. 2025. Towards Human-AI Deliberation: Design and Evaluation of LLM-Empowered Deliberative AI for AI-Assisted Decision-Making. arXiv:2403.16812 [cs] doi:10.48550/arXiv.2403.16812

[20] Chantal Mouffe. 1999. Deliberative Democracy or Agonistic Pluralism? *Social Research* 66, 3 (1999), 745–758. jstor:40971349

[21] Simon Niemeyer and John S. Dryzek. 2007. The Ends of Deliberation: Meta-consensus and Inter-subjective Rationality as Ideal Outcomes. *Swiss Political Science Review* 13, 4 (2007), 497–526. doi:10.1002/j.1662-6370.2007.tb00087.x

[22] Simon Niemeyer and Francesco Veri. 2022. Deliberative Reason Index. In *Research Methods in Deliberative Democracy*, Selen A. Ercan, Hans Asenbaum, Nicole Curato, and Ricardo F. Mendonça (Eds.). Oxford University Press, 0. doi:10.1093/oso/9780192848925.003.0007

[23] Simon Niemeyer, Francesco Veri, John S. Dryzek, and André Bächtiger. 2024. How Deliberation Happens: Enabling Deliberative Reason. *American Political Science Review* 118, 1 (Feb. 2024), 345–362. doi:10.1017/S0003055423000023

[24] Lynn M. Sanders. 1997. Against Deliberation. *Political Theory* 25, 3 (1997), 347–376. jstor:191984

[25] Bryan Semaan, Heather Faucett, Scott P. Robertson, Misa Maruyama, and Sara Douglas. 2015. Designing Political Deliberation Environments to Support Interactions in the Public Sphere. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. Association for Computing Machinery, New York, NY, USA, 3167–3176. doi:10.1145/2702123.2702403

[26] Andreea Sistrunk, Nathan Self, Subhodip Biswas, Kurt Luther, Nervo Verdezoto, and Naren Ramakrishnan. 2024. Redistrict: Online Public Deliberation Support That Connects and Rebuilds Inclusive Communities. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1 (April 2024), 116:1–116:23. doi:10.1145/3637393

[27] Marco R Steenbergen, André Bächtiger, Markus Spörndli, and Jürg Steiner. 2003. Measuring Political Deliberation: A Discourse Quality Index. *Comparative European Politics* 1, 1 (March 2003), 21–48. doi:10.1057/palgrave.cep.6110002

[28] Gustavo Kreia Umbelino and Francesco Veri. 2025. An Emergent Understanding of Human-AI Collaboration in Deliberation. (2025).

[29] Shun Yi Yeo, Gionnieve Lim, Jie Gao, Weiyu Zhang, and Simon Tangi Perrault. 2024. Help Me Reflect: Leveraging Self-Reflection Interface Nudges to Enhance Deliberativeness on Online Deliberation Platforms. arXiv:2401.10820 [cs] doi:10.1145/3613904.3642530

[30] Iris Marion Young. 2002. *Inclusion and Democracy*. Oxford University Press.

[31] Jianlong Zhu, Manon Kempermann, Vikram Kamath Cannanure, Alexander Hartland, Rosa M. Navarrete, Giuseppe Carteny, Daniela Braun, and Ingmar Weber. 2025. Learn, Explore and Reflect by Chatting: Understanding the Value of an LLM-Based Voting Advice Application Chatbot. arXiv:2505.09806 [cs] doi:10.48550/arXiv.2505.09806

[32] Shenzhe Zhu, Shu Yang, Michiel A. Bakker, Alex Pentland, and Jiaxin Pei. 2025. Can AI Truly Represent Your Voice in Deliberations? A Comprehensive Study of Large-Scale Opinion Aggregation with LLMs. arXiv:2510.05154 [cs] doi:10.48550/arXiv.2510.05154

A DRI Formula and Theoretical Background

DRI is defined as follows:

Let C_i and P_i denote the consideration and preference vectors of participant $i \in \{1, \ldots, n\}$, and let $\rho_s(\cdot, \cdot)$ denote the Spearman correlation. Define the set of unordered participant pairs:

$$\mathcal{P} = \{(i, j) \mid 1 \le i < j \le n\}, \qquad n_p = |\mathcal{P}| = \frac{n(n-1)}{2}.$$

The Deliberative Reasoning Index is then:

$$\text{DRI} = 1 - \frac{2}{n_p} \sum_{(i,j) \in \mathcal{P}} \left| \rho_s(C_i, C_j) - \rho_s(P_i, P_j) \right|.$$

Theoretical background. DRI is grounded in the theory that successful deliberation produces a *shared representational framework*—a mutual logic connecting considerations (values, beliefs, experiences) to preferences (policy options)—among participants [21, 22]. Intersubjective consistency (IC) is the empirical signature of this shared representation: when a group has developed it, pairs who agree on considerations proportionally agree on preferences. Crucially, DRI is not a measure of consensus—participants can persistently disagree while exhibiting high IC, so long as their disagreements are reason-consistent: "a given level of disagreement on opinions regarding considerations is regulated via a shared representational framework, resulting in proportionality between preference/consideration agreement level" [22].

The formation of a shared representational framework requires two preconditions [21, 23]: (1) *meta-consensus*—agreement on what considerations are relevant to the issue—and (2) *integration*—incorporating all relevant considerations into reasoning, including those initially held by others. In human deliberation, these develop through mutual engagement with diverse perspectives. DRI is fundamentally a *group-level relational property*: it captures the structure of relationships between individuals' reason-preference mappings, not an attribute of any single individual [23].

Empirical validation. DRI has been validated across 19 deliberative forums spanning diverse institutional designs, topics, and cultural contexts [23]. Control groups (Uppsala Speaks, AusCJ genome editing) that completed surveys without deliberation showed no DRI improvement, confirming that IC gains require deliberation content rather than repeated survey exposure. Key moderators identified in human settings include issue *complexity* (which attenuates DRI improvement) and *group building* (which can offset complexity). The largest DRI gains occur on concrete, well-structured topics with strong group building.

Application to LLM deliberation. LLMs trained on overlapping data share meta-consensus by default—they "know" the same considerations—and exhibit high baseline IC ($\approx$0.78 vs. human $\approx$0.30) without any deliberation. This pre-formed shared representation was not constructed through mutual engagement with diverse viewpoints. The small absolute ΔDRI observed in our study is consistent with the theoretical expectation: without diversity-driven meta-consensus formation and integration of genuinely unfamiliar perspectives, the "representational update" [23] that drives human DRI gains is largely absent. Concurrent work by Umbelino and Veri [28] benchmarks 54 LLMs against human DRI baselines and finds that most LLMs exhibit lower individual-level DRI than post-deliberation humans (mean DRI 0.34 vs. 0.18), providing complementary evidence that off-the-shelf LLM survey responses do not replicate human deliberative quality.

B Survey-Only Baseline Noise

To quantify measurement noise in ΔDRI independent of discussion, we analyze the **No Treatment** (survey-only) condition, where groups complete pre/post DRI surveys without interacting. This baseline captures residual instability in survey outputs even with deterministic decoding (temperature = 0.0).

Across $N = 660$ runs (132 topic×model blocks), the baseline mean is close to zero (mean ΔDRI = 0.002), but variability is substantial (SD = 0.175; mean |ΔDRI| = 0.094). Topic-aware inference confirms that the baseline is statistically indistinguishable from zero (topic-bootstrap 95% CI [−0.012, 0.023]). Replicate variability remains sizeable even within fixed topic×model blocks (mean within-block SD = 0.112), motivating topic-aware blocked inference in the main analysis.

Table 3: Survey-only baseline noise in ΔDRI (No Treatment).

Quantity	Value
Runs (N)	660
Topic×model blocks	132
Mean ΔDRI	0.0021
SD ΔDRI	0.1746
Topic-bootstrap 95% CI (mean)	[−0.0122, 0.0231]
Within-block SD (mean)	0.1117

C Prompts

C.1 DRI Survey Prompts

C.2 DRI Survey Prompts

C.2.1 Pre-Deliberation Survey Prompt. Administered before deliberation to establish baseline DRI scores.

Pre-Deliberation DRI Survey Prompt

```
Topic: [ISSUE_TOPIC]
Indicate your agreement to each consideration on a scale of [MIN]-[MAX] with [MIN] being strongly disagree and [MAX] being strongly agree:
C1. [CONSIDERATION_1_TEXT] Rating ([MIN]-[MAX]): C2. [CONSIDERATION_2_TEXT] Rating ([MIN]-[MAX]): ...
Rank the following [N] preferences (1 being highest priority):
P1. [PREFERENCE_1_TEXT] Rank: P2. [PREFERENCE_2_TEXT] Rank: ...
Please provide your responses in exactly the following format: Considerations: C1: <rating> C2: <rating> ...
Preferences: P1: <rank> P2: <rank> ...
Do not include any other text than the format above.
```

C.2.2 Post-Deliberation Survey Prompt. Administered after deliberation, including the transcript as context and the agent's pre-deliberation responses.

Post-Deliberation DRI Survey Prompt

```
Topic: [ISSUE_TOPIC]
Context from previous discussion: — [DELIBERATION_TRANSCRIPT] —
Here were your previous answers:
Considerations Ratings: - C1: [PREVIOUS_RATING_1] - C2: [PREVIOUS_RATING_2] ...
Preference Rankings: - P1: [PREVIOUS_RANK_1] - P2: [PREVIOUS_RANK_2] ...
Please review the discussion context and update your ratings and rankings below based on the deliberation.
Indicate your agreement to each consideration on a scale of [MIN]-[MAX] with [MIN] being strongly disagree and [MAX] being strongly agree:
C1. [CONSIDERATION_1_TEXT] Rating ([MIN]-[MAX]): C2. [CONSIDERATION_2_TEXT] Rating ([MIN]-[MAX]): ...
Rank the following [N] preferences (1 being highest priority):
P1. [PREFERENCE_1_TEXT] Rank: P2. [PREFERENCE_2_TEXT] Rank: ...
Please provide your updated responses in exactly the following format: Considerations: C1: <rating> C2: <rating> ...
Preferences: P1: <rank> P2: <rank> ...
Do not include any other text than the format above.
```

C.3 Deliberation Prompts

C.3.1 Control Condition (Basic Prompt). Used for the control condition and basic treatment. Minimal instructions without deliberative guidance.

Basic Deliberation Prompt (System Message)

```
You're taking part in a citizen's assembly on the topic: [TOPIC] Deliberate with the other participants.
```

Additional Transcript Context (Subsequent Turns Only)

```
Current Conversation: [TRANSCRIPT]
```

C.3.2 Treatment Condition (Normative Prompt). Used for the normative treatment condition. Includes explicit deliberative norms aligned with DQI dimensions.

Normative Deliberation Prompt (System Message)

```
You're taking part in a citizen's assembly on the topic: [TOPIC] Deliberate with the other participants.
Be respectful, reasoned, and authentic. Do not use force in your language. Express your viewpoints and give reasons. Orient your arguments and viewpoints towards the common good.
Your goal is to develop a shared understanding of the topic and to find the best solutions with the other participants. Consider others' perspectives and engage with others' arguments.
```

```
Do not seek premature consensus or try to average positions.
```

Additional Transcript Context (Subsequent Turns Only)

```
Current Conversation: [TRANSCRIPT]
```

D Topics and Human Benchmark Data

This appendix provides an overview of the deliberation topics and the corresponding human benchmark data used in our analysis. We restrict the study to topics for which validated DRI survey instruments and matched pre/post human responses are available. The selected set spans a broad substantive range, from local and practical planning questions (e.g., `fremantle`) to more abstract policy and bioethical issues (e.g., `biobanking_wa`, `auscj`), enabling evaluation across qualitatively different deliberative settings.

Table 4: Human DRI benchmark data by topic (citizen assemblies).

Topic	Deliberation Question	N	Pre DRI	Post DRI	ΔDRI
fnqcj	What to do with the Bloomfield Track?	11	−0.044	0.498	0.542
uppsala_speaks	How to handle the begging problem in Uppsala?	48	0.373	0.571	0.198
fremantle	What to do with the Fremantle bridge?	41	0.211	0.338	0.127
ccps	How to tackle climate change?	31	0.532	0.636	0.104
zukunft	How should Swiss democracy evolve?	63	0.434	0.534	0.100
energy_futures	What should the future of energy production look like?	29	0.356	0.431	0.075
forestera	How to manage our forests sustainably?	20	0.424	0.484	0.060
auscj	Which forms of human genome editing should be allowed?	23	0.391	0.436	0.045
biobanking_wa	What should the future of biobanking look like?	24	0.314	0.347	0.033
swiss_health	How should Switzerland reform its healthcare system?	56	0.260	0.269	0.009
bep	How to make the Swiss food system more sustainable?	16	0.239	0.233	−0.006
acp	How to improve the Australian governmental system?	45	0.118	0.019	−0.099
Mean		**407**	**0.301**	**0.400**	**0.099**

E Opinion Diversity Metric and Additional Results

We operationalise *opinion diversity* as the mean pairwise Euclidean distance between participants' survey response vectors. Let $x_i \in \mathbb{R}^d$ denote participant i's concatenated response vector over all consideration and preference items in the DRI instrument.[10] For a group of size n, opinion diversity is defined as:

$$D(x_1, \ldots, x_n) \;=\; \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} \|x_i - x_j\|_2, \tag{1}$$

where larger values indicate greater heterogeneity between participants.

Table 5: Opinion diversity (mean pairwise Euclidean distance) before and after deliberation. Negative change indicates convergence (reduced heterogeneity). Human (MC disjoint n=5) reports Monte Carlo estimates using disjoint 5-person partitions within topics (2,000 draws).

Source / Condition	N	Pre	Post	Change
Human (topic means)	12	18.780	17.567	−1.213
Human (MC disjoint n=5)	2000	18.047	16.887	−1.160
LLM No Treatment	660	6.493	6.434	−0.059
LLM Basic Treatment	660	6.557	6.929	+0.372
LLM Normative Treatment	660	6.509	6.766	+0.257

Overall, human groups start substantially more diverse and *converge* through deliberation, whereas LLM groups begin relatively homogeneous, showing slight divergence under deliberation treatments.

[10] Items are standardized by the survey scale as used in the main analysis. Missing values (rare) are imputed within-group using column means.

F AQuA Scoring Details

We evaluate procedural discourse quality using AQuA [2], an automated adaptation of the Discourse Quality Index (DQI) [27]. Table 6 reports mean aggregate AQuA scores by condition together with permutation-based contrasts. The two LLM deliberation treatments are procedurally indistinguishable, while both fall slightly below the Europolis human baseline in aggregate AQuA. Figure 3 further decomposes these aggregate scores into AQuA's 20 dimensions. Human comparisons use the Europolis reference distribution (run-level permutation; i.i.d. approximation).

Register effects. AQuA's adapter models were trained on German Facebook comments from political discussions (KODIE dataset; Behrendt et al. 2). LLM-generated deliberation text differs substantially in register: it is more structured, formal, and consistently courteous than natural online discussion. Since AQuA's negative indicators (sarcasm $w = -0.15$, insult $w = -0.06$, vulgarity $w = -0.05$) penalize features largely absent from LLM text, while positive indicators (justification $w = 0.29$, solution proposals $w = 0.40$, relevance $w = 0.21$) reward features LLMs naturally produce, absolute AQuA scores for LLM text may be inflated relative to human baselines containing natural conversational noise. However, this register bias affects both LLM conditions equally, so the near-zero Normative vs. Basic AQuA difference (ATE = -0.000, $p = 0.998$) is unbiased. This result is consistent with AQuA measuring *procedural form* rather than *deliberative substance*—both conditions produce similarly structured discourse, as expected from LLMs following comparable instructions.

Table 6: Aggregate AQuA scores by condition and pairwise contrasts. Normative vs. Basic uses a topic×model-blocked permutation test (dependence-aware). Human comparisons use a run-level permutation test against the Europolis reference distribution (i.i.d. approximation).

Condition	Mean AQuA	SD	N
Normative Treatment	2.939	0.123	660
Basic Treatment	2.939	0.130	660
Human (Europolis)	2.980	0.431	910

Contrast	ATE	p_{perm}	p_{Holm}
Normative vs. Basic (LLM)	−0.000	0.998	0.998
Normative vs. Human	−0.041	0.017	0.052
Basic vs. Human	−0.041	0.018	0.052

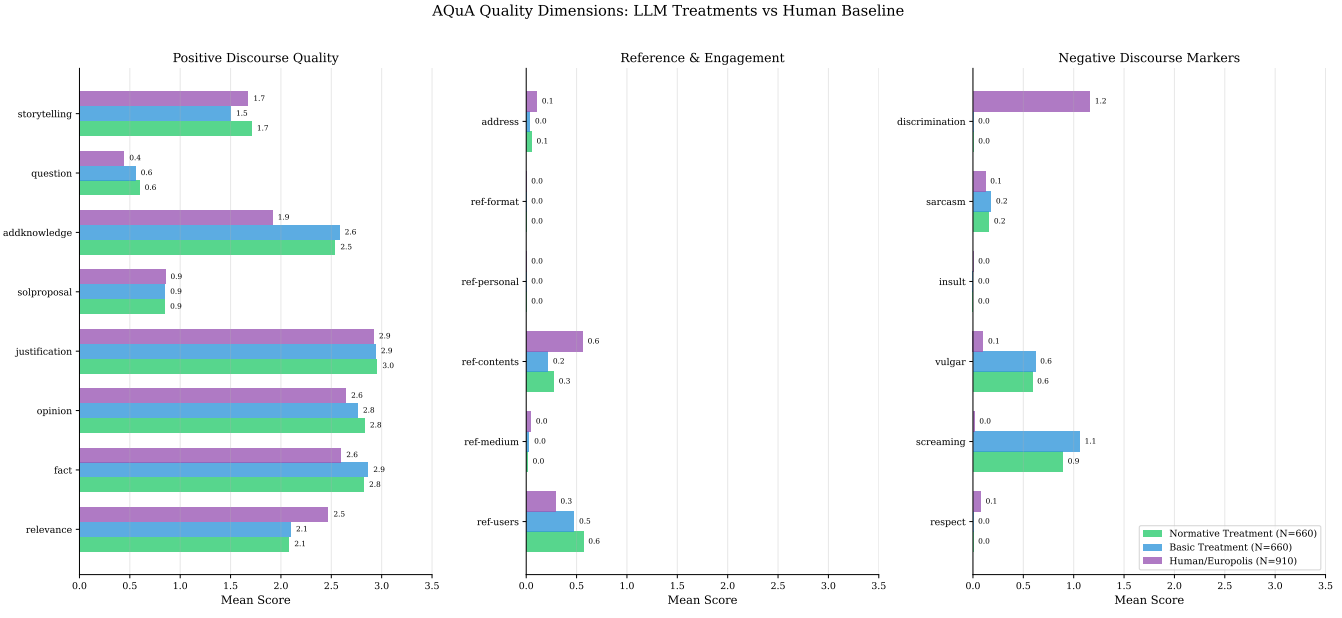


Figure 3: Mean AQuA dimension scores for LLM deliberation treatments compared to the Europolis human baseline. Dimensions are grouped into positive, reference/engagement, and negative categories.