Can LLMs deliberate? Benchmarking Collective Reasoning for Democratic AI Applications

Anonymous Authors[1]

Abstract

Multi-agent LLM systems are increasingly proposed for democratic applications, including consensus-finding, deliberation moderation, and the representation of stakeholder perspectives. While existing benchmarks for collective reasoning in LLM systems show strong performance on verifiable tasks such as Maths problems, we argue that these benchmarks cannot assess deliberative reasoning on contested normative questions, which, however, is exactly what democratic applications demand. We introduce *DelibSim*, a configurable simulation environment that benchmarks multi-agent deliberation along two theoretically grounded dimensions: procedural discourse quality (AQuA) and deliberative reasoning quality (DRI). Across 1,980 five-agent deliberations spanning 11 model configurations and 12 citizen-assembly topics, LLMs achieve discourse quality statistically indistinguishable from human deliberation (AQuA 2.94 vs. 2.98). Normative prompting yields a small but reliable improvement in shared understanding ($\Delta\text{DRI} = 0.029$, $p = 0.005$), but effects do not survive topic-level correction and turn negative on ethically complex topics. Most strikingly, LLM groups exhibit far lower perspective diversity than human groups (6.5 vs. 18.8) and reversed convergence dynamics: human deliberation *decreases* dispersion as diverse views synthesize, whereas LLM deliberation *increases* it. These findings expose an important failure mode for AI deployed in deliberative settings: high-quality deliberative discourse can mask fundamentally different reasoning dynamics. We release *DelibSim* as an open benchmark to support the responsible deployment of multi-agent AI in democratic applications.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

LLM-based agents are rapidly moving into civic and democratic settings. Recent and proposed applications include consensus-finding among citizens (Tessler et al., 2024), large-scale deliberation moderation (Klein et al., 2025), deliberation assistance (Ma et al., 2025), representation of missing perspectives in citizen panels (Fulay et al., 2025; Zhu et al., 2025), and full simulations of democratic systems (Novelli et al., 2025). These applications share an implicit assumption: that LLMs can engage with this kind of *epistemic* deliberation and reason together about contested normative questions.

This assumption is reinforced by a broader trend. As users interact with highly confident and increasingly capable models, they struggle to match their level of trust to the trustworthiness of LLMs (Heersmink et al., 2024). Anthropomorphism plays a central role here: fluent, well-structured, and norm-compliant outputs invite users to assume that the system reasons in human-like ways (Kadambi et al., 2026; Colombatto et al., 2025). This assumption is especially consequential in deliberative contexts, where the legitimacy of an outcome depends not only on *what* is said but on the kind of reasoning process that produced it (Cohen, 2002).

Yet current benchmarks for multi-agent LLM systems evaluate deliberative capacity exclusively through performance on verifiable tasks in closed, often game-like settings (Agashe et al., 2025; Cipolina-Kun et al., 2025; Wu et al., 2025). The domain of complex, normatively loaded questions, where no objectively correct answer exists and the goal of deliberation is mutually acceptable conclusions (Niemeyer and Dryzek, 2007), remains untouched. This, however, is the very domain in which LLMs are now being deployed.

We address this gap with *DelibSim*, a simulation environment that benchmarks multi-agent LLM deliberation using established metrics for deliberative processes. *DelibSim* operationalises deliberation along two dimensions. *Procedural discourse quality* is measured via AQuA (Behrendt et al., 2024), an automated adaptation of the Discourse Quality Index (DQI) (Steenbergen et al., 2003). *Deliberative reasoning quality* is captured by the Deliberative Reason Index

(DRI) (Niemeyer and Veri, 2022), which measures whether groups develop increased intersubjective consistency (IC), i.e. a shared understanding of how considerations map to preferences, from pre- to post-deliberation. This process–outcome decomposition lets us distinguish the *performance* of deliberation from epistemic *gains* from deliberation, a distinction that matters precisely because the former is easy to mistake for the latter.

Our findings point to a qualified capability and a critical failure mode. Across 1,980 five-agent deliberations spanning 11 model configurations and 12 citizen-assembly topics, LLM groups achieve procedural discourse quality statistically indistinguishable from human deliberation. Normative prompting yields a small but reliable gain in IC (ΔDRI $= 0.029$, $p = 0.005$), demonstrating that LLMs can achieve measurable epistemic gains under guidance. But this capability sits within strong constraints: effects are small, do not survive Holm correction at the topic level, and turn negative on ethically complex topics. LLM groups exhibit far lower perspective diversity than human groups (mean pairwise distance 6.5 vs. 18.8), and display *reversed* convergence dynamics: human deliberation *decreases* opinion dispersion (-1.21) as diverse views synthesize toward shared understanding, whereas LLM deliberation *increases* it ($+0.26$ to $+0.37$). A persona pilot that engineers human-level diversity does not yield gains in ΔDRI, indicating that the small baseline effect for vanilla LLMs reflects consistency gains among already-similar agents rather than genuine integration of diverse viewpoints.

We conceptualise *DelibSim* as a *necessary-condition test*. A positive ΔDRI is necessary, but not sufficient, evidence for substantively epistemic deliberation. While behavioural benchmarks cannot fully verify the presence of an underlying epistemic process, the absence of even this minimal behavioural signature is diagnostically informative for proposed democratic applications. The finding that LLM agents look like deliberators while reasoning very differently than humans is precisely the pattern that should worry practitioners deploying AI in settings that allow for deliberative participation. Anthropomorphic surface features can mask reasoning dynamics that are neither human-like nor structurally compatible with the goals of democratic deliberation.

This paper contributes:

1. A **benchmark** for deliberative AI in deliberative settings, with 12 citizen-assembly topics, 3 prompting regimes, 11 model configurations, and matched human reference data.

2. An **evaluation framework** demonstrating that procedural metrics alone are insufficient: discourse quality, reasoning outcomes, and perspective diversity can diverge sharply.

3. Empirical evidence that frontier LLMs achieve only **limited deliberative reasoning gains**, even under explicit guidance. These gains are concentrated on tractable topics and reflect alignment among already-similar agents rather than synthesis across diverse perspectives.

4. Characterisation of a **failure mode** where deliberative *performance* masks the absence of deliberative *reasoning*, with direct implications for AI applications in democratic and civic discourse.

2. Related Work

Multi-agent LLM systems. Structured multi-agent interaction through debate, critique, or collaborative refinement can improve factual accuracy (Du et al., 2023) and reasoning quality (Liang et al., 2024) relative to single-agent baselines. LLM populations also exhibit emergent social dynamics, including norm adoption (Ashery et al., 2025) and polarization patterns resembling those of human groups (Piao et al., 2025). Existing benchmarks evaluate these systems through coordination tasks (Anne et al., 2025), strategic reasoning (Cipolina-Kun et al., 2025; Wu et al., 2025), and deception detection (Agarwal et al., 2025). All rely on tasks with verifiable solutions, leaving open-ended normative deliberation unaddressed.

Deliberation under normative disagreement. Democratic deliberation concerns contested questions without objectively correct answers, where the goal is mutually acceptable conclusions reached through reasoned discourse (Dryzek, 2002; Niemeyer and Dryzek, 2007), ideally among an epistemically diverse set of participants (Landemore, 2013). Procedural norms such as reason-giving, reciprocity, and mutual respect (Chambers, 1996) structure the process, but high procedural quality does not guarantee substantive epistemic gains (Knobloch and Gastil, 2022; Baccaro et al., 2016). We adopt this distinction as the core of our evaluation framework.

LLMs as epistemic agents in democratic settings. A growing literature documents that LLMs carry political and ideological biases (Fulay et al., 2024; Potter et al., 2024) and that interaction with LLMs can shift user opinions (Potter et al., 2024). AI deployment in deliberative formats can also shape participation as citizens express skepticism toward AI-facilitated deliberation (Jungherr and Rauchfleisch, 2025). Accordingly, Landemore (2024) argues that the value of AI in democracy hinges decisively on whether it preserves the conditions of genuine collective reasoning. Our benchmark provides one empirical test of these conditions: does multi-agent deliberation yield the epistemic gains that deliberation is designed to produce?

Measuring deliberation. We operationalise procedural quality with AQuA (Behrendt et al., 2024), which automates DQI assessment by combining 20 adapter models trained on dimensions such as respect, reciprocity, and reason-giving, producing a 0–4 score validated against human judgments. For deliberative outcomes, classic metrics (opinion change, consensus, polarization reduction) presuppose a single best solution and are unsuitable for contested normative topics (Mouffe, 1999; Dryzek and Niemeyer, 2006). The DRI offers an alternative by measuring intersubjective consistency: the degree to which participants share an understanding of how considerations map onto preferences (Niemeyer and Dryzek, 2007; Niemeyer et al., 2024). Positive ΔDRI indicates that a group has developed greater agreement about *how* to reason about an issue, even when participants continue to disagree about *what* to conclude. Mapping LLM and human responses into the same DRI space enables comparison of reasoning structures (Kreia Umbelino and Veri, 2025). So far, this has not been tested under deliberative treatment conditions for LLMs — a gap that this paper aims to fill in order to approximate the effect of actual deliberation among LLM agents on DRI.

3. Method

DelibSim. *DelibSim* simulates turn-based deliberation among LLM agents (Figure 1). Groups of n agents deliberate for r rounds, with speaking order randomized within rounds. The full transcript passes between agents, building shared context. Each agent maintains a consistent identity throughout the deliberation. The environment supports `homogeneous` groups (all agents using the same base LLM) and `mixed-model` groups (each agent using a different base LLM). The pipeline has three phases: (1) pre-deliberation DRI survey, (2) deliberation, and (3) post-deliberation survey, where agents update their responses after reviewing the full transcript. Temperature is set to 0.0 for reproducibility [1].

AQuA. We assess discourse quality using AQuA, computing an aggregate score across 20 dimensions on a 4-point scale. Because AQuA's adapter models are trained on German DQI data, all transcripts are machine-translated prior to scoring. This may differentially affect register-sensitive dimensions, but within-study comparisons remain valid because all conditions pass through the same pipeline. We benchmark against the same human reference distribution used by AQuA's developers (Europolis deliberative poll comments, AQuA $= 2.98$, $N = 910$) (Gerber et al., 2018). The Europolis reference differs from the citizen-assembly data underlying our DRI benchmark. That is because, to our knowledge, no single corpus jointly provides validated DRI instruments and high-quality deliberation transcripts, so each metric is evaluated separately against its best available human reference.

DRI. Let C_i and P_i denote participant i's consideration and preference vectors and ρ_s the Spearman correlation. With $\mathcal{P}$ the set of unordered pairs and $n_p = |\mathcal{P}|$,

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

Positive ΔDRI (post minus pre) indicates increased intersubjective consistency. We rely on existing DRI surveys from real-world deliberative processes, LLM responses are validated for completeness and format compliance.

Perspective diversity. Let $x_i \in \mathbb{R}^d$ denote participant i's standardized concatenated response vector over all DRI items. Group diversity is the mean pairwise Euclidean distance:

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

Treatment conditions. Building on findings that model capability matters more than debate length and group size (Wu et al., 2025), we set group size to 5 and length to 2 rounds [2]. This balances coordination costs against sufficient context and dyads for DRI computation ($\binom{5}{2} = 10$ pairs). The average word count per simulation is 4,831, corresponding to roughly 35–40 minutes of human deliberation in content volume.

We test three conditions. **No Treatment** is a survey-only baseline: agents complete pre/post DRI surveys with no interaction. This establishes baseline noise (SD ≈ 0.17; within-(topic×model) SD ≈ 0.11 [3]). **Basic Treatment** simply prompts agents to deliberate, relying on the model's internal understanding of "deliberation". **Normative Treatment** adds explicit deliberative norms (respect, reason-giving, authenticity, engagement with opposing views, oriented toward shared understanding) [4].

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 `Claude`, we include both reasoning-enabled and standard variants, plus two `mixed-model` ensembles (one reasoning, one standard).

[1] See Appendix A for an algorithmic description.

[2] See Appendix J for length and group-size ablations.

[3] See Appendix B for baseline noise analysis.

[4] See Appendix C for full prompts.

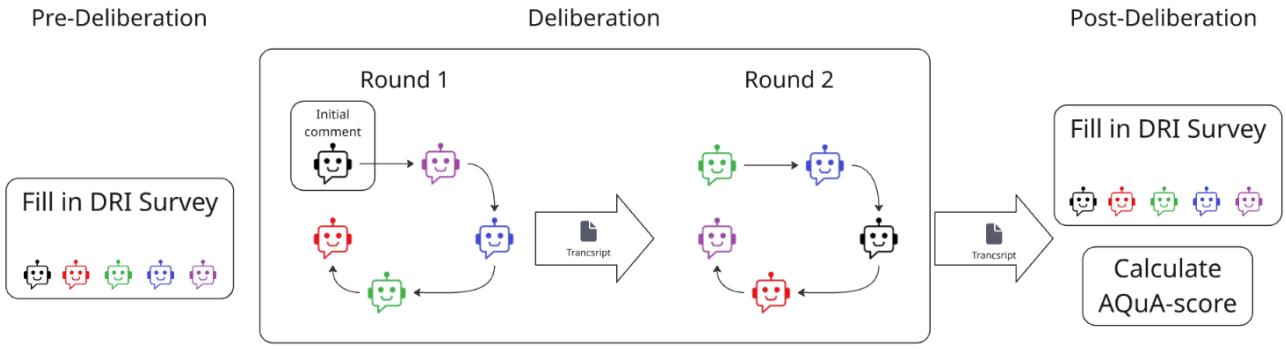

Figure 1. DelibSim design. The simulation proceeds in three phases: pre-deliberation DRI survey, multi-agent deliberation, and post-deliberation survey. The full transcript is passed between agents for continuous context. Speech order is randomized within rounds.

Table 1. Pooled treatment effects on absolute ΔDRI using (topic×model)-blocked contrasts. p-values from blocked permutation tests; CIs from topic-resampled bootstrap.

Contrast	ATE	95% CI	p	p_{Holm}
Norm. vs. None	0.029	[−0.004, 0.076]	0.005	0.015
Basic vs. None	0.019	[−0.011, 0.057]	0.079	0.158
Norm. vs. Basic	0.010	[−0.008, 0.030]	0.349	0.349

Topics and human benchmarks. The complete pool of validated DRI instruments comprises approximately 20 published citizen assemblies (Niemeyer et al., 2024; Umbelino and Veri, 2025). We selected 12 topics spanning distinct policy areas (climate, healthcare, governance, bioethics, urban planning) to avoid inflating overall effects with correlated topic-specific impacts. Human DRI data from these assemblies provide a reference point, though differences in baseline diversity and treatment design limit direct comparability [5].

Robustness pilots. We conducted five smaller-scale ablations during revision: decoding temperature ($T \in \{0.0, 0.7, 1.0\}$), prompt paraphrasing of the normative treatment, group size (3/5/7), conversation length (2/4/6 rounds), and persona prompting based on discoursive profiles grounded in human DRI-survey data. These cover 2–3 models and three topics each ($N = 60$ to $N = 180$), they are diagnostic rather than confirmatory[6].

4. Results

Topic dependence. Deliberative outcomes depend substantially on topic. The intraclass correlation for absolute ΔDRI by topic is 0.107 (95% CI $[-0.001, 0.160]$), motivating treating topics as the unit of replication and using topic-aware clustered inference. AQuA exhibits more modest topic clustering ($\approx 3.5\%$). We apply the same controls for consistency [7]. For LLM analyses we report (i) blocked pooled contrasts over the full set of (topic×model)-blocks ($12 \times 11 = 132$) and (ii) a hierarchical mixed model as a complementary specification.

Procedural discourse quality matches humans. LLM groups achieve discourse quality close to human deliberation in the Europolis reference. Normative and Basic treatments yield identical mean AQuA scores (both 2.939), suggesting explicit deliberative guidance does not meaningfully affect *procedural* discourse quality. Compared to humans (SD = 0.431), LLM discourse exhibits substantially lower variance (SD = 0.12–0.13), indicating more uniform but potentially less varied contributions [8].

Normative guidance improves absolute ΔDRI. Our primary estimand is the pooled effect on absolute ΔDRI, computed as the equal-weight average of within-(topic×model) mean differences. Under this design, only the Normative treatment yields a significant gain over no treatment (Table 1). The topic-resampled bootstrap CI is wide and marginally overlaps zero, reflecting limited topic-level replication. A hierarchical mixed model with random intercepts by topic and random slopes by model yields a consistent fixed effect of ≈ 0.030 (95% CI $[0.004, 0.054]$, $p \approx 0.025$), corroborating the blocked result [9].

Effect-size context. The 0.029 improvement is statistically detectable but substantively modest. The noise floor (No Treatment) is ΔDRI = 0.002 (SD = 0.175). The human reference, averaged across 12 citizen assemblies, is ΔDRI ≈ 0.099 (95% CI $[0.026, 0.196]$). The LLM normative effect is roughly 29% of the human mean and falls within the lower tail of the human CI. Taken together, delib-

[5] See appendix D for a topic overview.
[6] See appendices J and K for the pilot results.
[7] See appendix E for Intraclass Correlation Coefficients.
[8] See appendix F for comprehensive AQuA results.
[9] See appendix G for complete HMM results table.

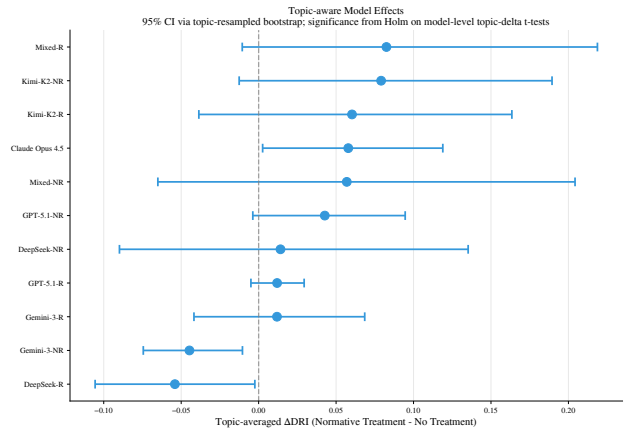

Figure 2. Model-level treatment effects on absolute ΔDRI (topic-averaged within-topic differences). None significant after Holm correction.

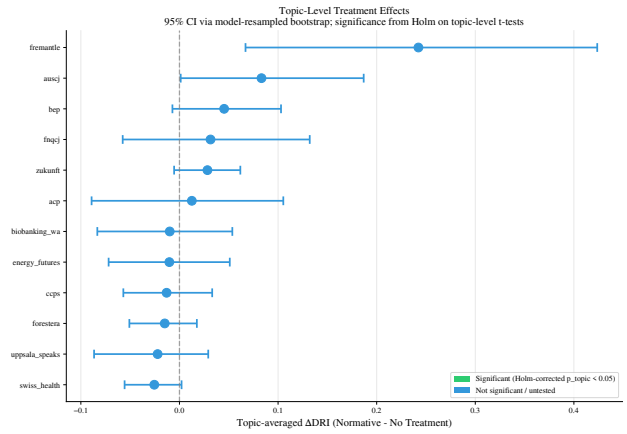

Figure 3. Topic-level treatment effects on absolute ΔDRI. None significant after Holm correction.

eration yields detectable but limited shared understanding, well below the human reference.

Model and topic heterogeneity. Treatment responsiveness varies across models. Most models show positive point estimates (`Mixed-R` leads), while some exhibit negative effects (`Gemini-NR`, `DeepSeek-R`), none survive Holm correction (Figure 2 [10]. Topic effects are similarly heterogeneous: the largest positive effect is on `fremantle`, a local urban-planning topic. Ethically charged topics like healthcare (`swiss_health` ≈ -0.026) or street begging (`uppsala_speaks` ≈ -0.022) show slightly negative effects (Figure 3) [11].

Human reference and ceiling effects. Humans show much larger absolute gains ($\Delta\text{DRI} \approx 0.099$). However,

[10] See appendix H.1 for detailed model-heterogeneity results.

[11] See appendix H.2 for detailed topic-heterogeneity results.

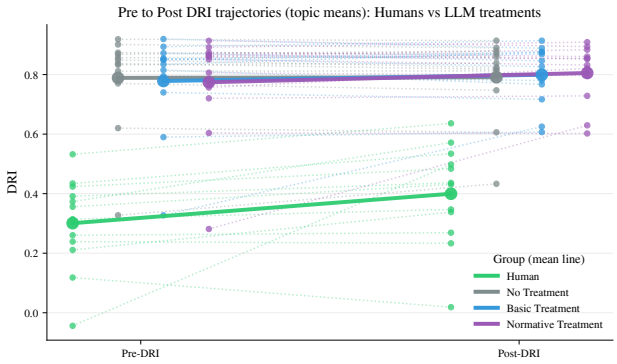

Figure 4. Pre- to post-deliberation DRI trajectories. Human groups (green) start with low IC and improve substantially. LLM groups across all treatments start near ceiling with limited headroom.

LLM and human pre-DRI distributions barely overlap: LLM groups cluster near 0.78, humans near 0.30 (Figure 4). Pooled across LLM runs, ΔDRI declines sharply with pre-DRI ($\beta \approx -0.50$, $R^2 \approx 0.36$, $p < 0.001$), motivating a headroom-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, both deliberation treatments yield clear gains (Norm. vs. None: ATE $= 0.066$, $p_{\text{Holm}} < 0.001$; Basic vs. None: 0.061, $p_{\text{Holm}} < 0.001$). We report relative DRI cautiously as it differs from DRI's original conception, and average headroom (≈ 0.22) still allows for human-level absolute gains [12].

Perspective diversity diverges from humans. Most critically, diversity dynamics differ sharply between settings (Figure 5). Human groups start with substantially higher heterogeneity (mean pairwise Euclidean distance 18.78 vs. $\approx$ 6.5 for LLMs) and *converge* through deliberation ($\Delta \approx -1.21$), consistent with the theoretical expectation that deliberation builds shared understanding across diverse viewpoints. LLM groups, by contrast, begin with highly similar perspectives and tend to remain stable or *diverge* slightly under deliberation (Basic: $+0.37$; Normative: $+0.26$). `Mixed-model` ensembles exhibit the highest LLM diversity ($\approx$ 9.6 for non-reasoning, $\approx$ 9.2 for reasoning), but still fall far below humans, indicating that architectural heterogeneity does not produce the perspective heterogeneity that characterises human groups. LLMs start where successful human deliberation arrives, not through reasoning but through similar training and architecture, and

[12] See appendix I for detailed ceiling effects analysis results.

procedural metrics cannot detect this difference [13] [14].

Persona prompting and the diversity ceiling. A central concern is that low ΔDRI might be caused by homogeneity. To test this, we constructed empirically grounded personas by clustering human pre-deliberation DRI-survey responses ($k = 5$ per topic) and translating cluster-defining considerations into natural-language value profiles[15]. Pre-deliberation diversity rises from ≈ 7.5 to ≈ 27.7 ($+20.23$, $p < 0.001$), exceeding the human reference of 18.8. But ΔDRI does not improve ($\beta = -0.058$, $p = 0.456$). Decomposing ΔDRI into its two components reveals a striking inversion of the human pattern (Figure 6). In human citizen assemblies, deliberation produces larger gains in *preference agreement* ($+0.104$) than in *consideration agreement* ($+0.077$). Participants enter with initial sets of considerations and primarily update their preferences to align them more consistently to these considerations. Without persona prompting, agreement dynamics mimic human dynamics, but both components remain near zero ($+0.007$ and $+0.024$). Persona-prompted LLM groups reverse this. Consideration agreement rises ($\Delta = +0.066$, $d = 0.55$, $p = 0.039$), but preference agreement does not ($\Delta = -0.008$, n.s.). Persona-prompted agents thus align on the dimension that is typically more stable in human deliberation, i.e. considerations — a dynamic that could be interpreted as a shift towards the LLMs' default biases (Taubenfeld et al., 2024) — while continuing to disagree on their preferences, which is where human deliberators generally align on to increase consistency with their considerations. The pilot is small ($N = 60$), but the diversity manipulation is unambiguous, and the inverted decomposition, rather than the mere absence of gains, indicates that engineered diversity changes *which* component of deliberative reasoning LLMs update without producing the joint alignment towards intersubjective consistency that defines ΔDRI in humans.

Robustness. Four protocol-level ablations indicate the main findings do not depend on design choices. Decoding temperature has no significant effect on ΔDRI, pre-deliberation diversity, or convergence ($T \in \{0.0, 0.7, 1.0\}$, all $p_{\text{adj}} > 0.69$), suggesting homogeneity reflects properties of learned distributions rather than deterministic decoding. Four paraphrases of the normative prompt produce no detectable differences (all $p_{\text{adj}} > 0.57$). Group size (3/5/7) and length (2/4/6 rounds) likewise show no significant effects (all $p_{\text{adj}} > 0.24$ and > 0.73). Notably, longer conversations do yield greater individual updating: Kendall's τ on pre/post preferences declines from 0.74 at two rounds to ≈ 0.60 at six rounds, without corresponding gains in ΔDRI. Accordingly, individual responsiveness does not necessarily aggregate into shared understanding [16].

Summary. Five findings constitute the core empirical contribution. (1) LLM groups achieve discourse quality close to the human reference. (2) Normative prompting produces a small but reliably positive effect on ΔDRI (≈ 0.029, roughly 29% of the human mean). (3) Effects are strongly topic-dependent and do not survive Holm correction, ethically salient topics show negative point estimates. (4) LLM groups display substantially lower perspective diversity than humans and tend to *diverge* through deliberation, engineered diversity does not yield improved ΔDRI. (5) Patterns are robust across decoding temperature, prompt wording, group size, and conversation length. Taken together, the results indicate a form of deliberation that succeeds at the procedural level while failing to trigger corresponding epistemic gains.

5. Discussion

Performance without process. The discourse–reasoning gap is the central empirical finding. LLM agents produce discourse that meets DQI prerequisites (justification, respect, reciprocity) while intersubjective consistency does not reliably improve. This precedented disconnect in human research (Knobloch and Gastil, 2022; Baccaro et al., 2016) takes an extreme form with LLMs and aligns with findings that LLM agents reproduce surface discourse patterns while falling short on belief updating and opinion alignment (Chuang et al., 2025). This pattern is the failure mode that should most concern developers of AI solutions in democratic settings as LLM-deployment in such settings will often presuppose the epistemic capacities that humans naturally employ during deliberation. However, as of now, the LLMs' epistemic fingerprint looks fundamentally different. If deliberative quality is solely judged by surface metrics, naive deployment will pass evaluation while delivering none of the epistemic substance that justifies deliberation.

The epistemic-authority risk. As LLMs are increasingly considered as epistemic authorities (Yeung et al., 2025; Milella and Cabitza, 2026), users tend to assume that their fluent, well-structured outputs reflect human-like reasoning processes (Kadambi et al., 2026; Colombatto et al., 2025). Our results show this assumption is unsafe in deliberative settings. LLM groups achieve human-comparable procedural quality and small absolute IC gains, indicating that they can easily make a convincingly deliberative impression. Yet, the underlying epistemic dynamics differ qualitatively from human deliberation: starting positions are far more homogeneous, deliberation tends to *increase* rather than decrease

[13] See appendix M for comprehensive diversity analysis results.

[14] Illustrative qualitative example transcripts can be found in appendix N.

[15] See appendix K for detailed persona prompting results.

[16] See appendix J for robustness analyses, supplementary outcome measures can be found in appendix L.

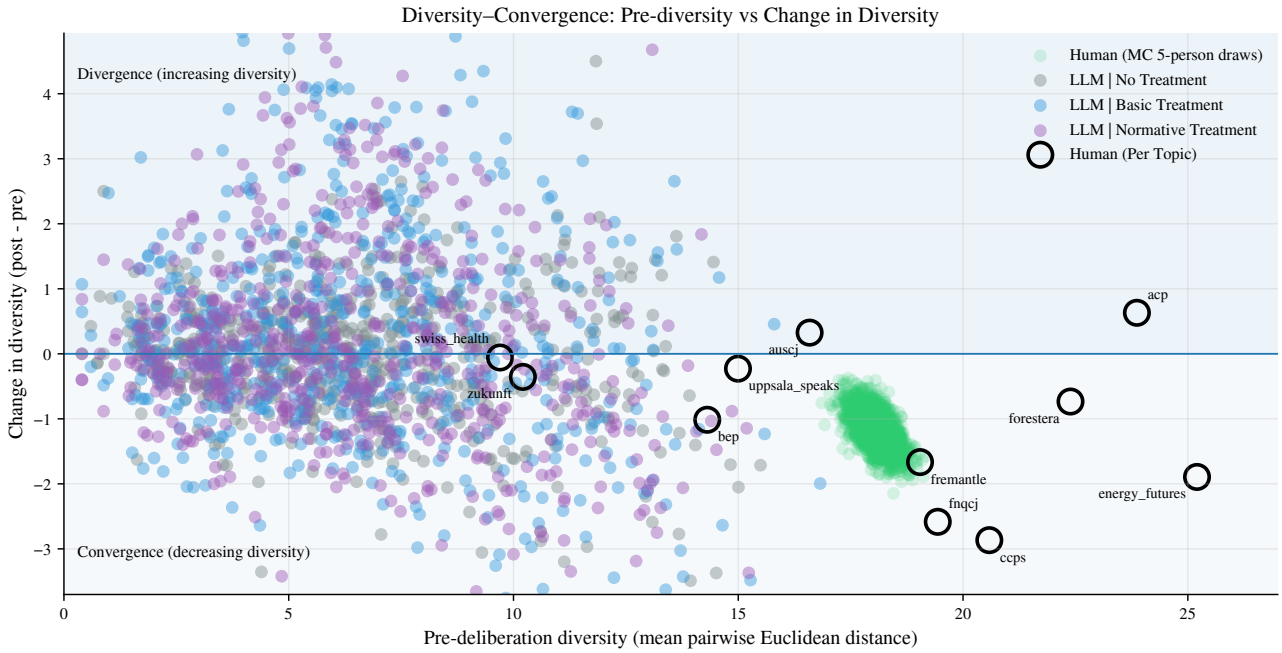

Figure 5. Diversity dynamics in human and LLM deliberation. Pre-deliberation perspective diversity (mean pairwise Euclidean distance) is plotted against change in diversity (post minus pre). Human groups (green; Monte Carlo samples of disjoint 5-person draws) cluster in the lower-right: high initial diversity with convergence through deliberation. LLM groups across all treatments scatter widely but never reach human diversity levels.

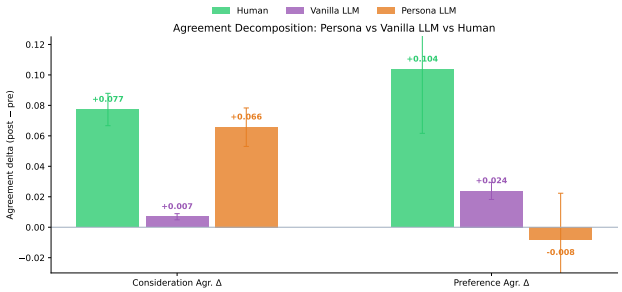

Figure 6. Decomposition of ΔDRI into its two components: *consideration agreement* (alignment on how reasons are weighted) and *preference agreement* (alignment on policy rankings).

perspective dispersion, and engineered diversity does not yield deliberative gains. A system that talks like a deliberator is not, on this evidence, necessarily reasoning like one. Applications that assume epistemic parity, e.g. replacing or supplementing human deliberators, simulating publics, or generating "representative" citizen views, risk projecting deliberative legitimacy onto a process whose internal reasoning structure does not warrant it.

What LLM deliberation actually does. The lower perspective diversity of LLM groups means that DRI carries a different meaning across settings. Human deliberation typically involves bridging diverse experiences and values toward shared understanding. This diversity is much of the source of deliberation's epistemic value (Landemore, 2013). LLM groups begin already aligned, lacking part of the headroom for IC improvement through bridging diverse viewpoints. Within this constraint, our setup isolates a different capability: the alignment of internal reasoning structures among already-similar agents. The significant Normative effect indicates LLMs can perform this narrower task to a certain extend under distinct guidance. It is real but not equivalent to human deliberation, and it is not what democratic applications typically need.

DRI as signature, not process. A general concern in applying behavioural benchmarks to LLMs is that the underlying epistemic process, in this case the construction of shared understanding through reasoned engagement with disagreement, cannot be directly observed and may be absent even when its behavioural signature appears. We accept this limitation. No behavioural metric, including those used for human deliberation, can fully distinguish genuine deliberative reasoning from sophisticated pattern matching, prompt compliance, or superficial convergence. The persona experiment sharpens rather than resolves this concern: when starting positions are deliberately diverse, the small positive ΔDRI disappears, suggesting baseline effects reflect alignment among already similar agents rather than a deliberative process. We therefore frame *DelibSim* as

a necessary-condition test. If LLMs fail to produce even the behavioural signature of deliberation under conditions favourable to its detection, claims that they deliberate in any richer epistemic sense are not supported by behavioural evidence. The benchmark thus constrains what can be inferred from successful surface-level performance, the practically relevant question for proposed deliberative applications.

Implications for AI in civic discourse. These findings have direct implications for the rapidly expanding set of deliberative AI applications. Surface fluency makes LLM "deliberations" easy to mistake for the real thing in user-facing demonstrations. The gap from real deliberation only becomes visible under structured measurement. *DelibSim* provides one such measurement and supports two operational recommendations. First, deliberative AI deployments should report both procedural and epistemic outcomes, the former is necessary but not sufficient. Second, claims about "representing" or "simulating" citizens with LLM agents should be supported by evidence that the agents reproduce the convergence dynamics of the populations they purport to represent, which our data show current frontier systems do not.

Limitations. Topic-level statistical power is constrained by the 12 citizen assemblies with validated DRI instruments, the pooled estimand is well powered, topic-level effects are not. Behavioural metrics capture a signature, not a process; cross-setting comparisons assume this signature carries comparable meaning across humans and LLMs, an assumption not yet established. Ceiling effects limit absolute headroom for LLM gains, although remaining headroom (≈ 0.22) still exceeds typical human absolute improvements. Machine translation may affect register-sensitive AQuA dimensions, within-study comparisons remain valid. Robustness and persona pilots cover 2–3 models and three topics, full replication is left to future work. Our fixed protocol (five agents, two rounds at scale) is supported by ablations but does not establish generalization to substantially different settings such as moderated or multi-day deliberations.

Training-data contamination. Contamination cannot be fully ruled out: all 12 topics derive from publicly documented citizen assemblies. However, contamination would predict uniformly high and stable DRI, which is not observed: ΔDRI turns negative for `swiss_health` and `uppsala_speaks`, which is hard to reconcile with simple memorization of published conclusions. Contamination may contribute to elevated pre-DRI baselines but does not parsimoniously account for topic-level heterogeneity in treatment effects.

6. Conclusion

We introduced *DelibSim*, a simulation environment for evaluating multi-agent LLM deliberation with theory-grounded political-science metrics. Across 1,980 small-group sessions, LLM discussions achieve procedural discourse quality comparable to human references and produce interactions that superficially resemble deliberation. Absolute IC gains are statistically reliable but small under normative guidance, and turn negative on ethically complex topics. Given the much lower perspective diversity of LLM agents, this gain reflects alignment of internal reasoning structures among already-similar agents rather than synthesis across diverse perspectives.

For AI4Good, the operative finding is the gap between deliberative *performance* and deliberative *reasoning*. As LLMs are progressively treated as epistemic authorities and deployed in democratic and civic settings, anthropomorphic surface features invite users to assume human-like reasoning processes that, on this evidence, are not present. We caution against applications that assume or suggest epistemic parity between LLM agents and human deliberators on the basis of procedural fluency alone. *DelibSim* is released as an open benchmark to support the responsible development and evaluation of deliberative AI systems, and to ensure that gains in apparent quality are matched by gains in the underlying reasoning that gives democratic deliberation its value.

Impact Statement

This paper evaluates the deliberative capabilities of LLMs in multi-agent systems with implications for AI deployment in democratic and civic contexts. The finding that LLMs achieve near-human procedural discourse quality while exhibiting fundamentally different deliberative dynamics, in particular dramatically lower perspective diversity and reversed convergence dynamics, has direct implications for proposed civic applications. Surface-level indicators of deliberative quality may mask the absence of substantive epistemic deliberation, potentially leading to overconfidence in AI-assisted democratic processes. We caution against deployment in deliberative contexts where deep epistemic engagement is required when judgment rests solely on procedural metrics. The disconnect we document could yield democratic applications that appear functional but fail to deliver or even prevent the epistemic benefits that make deliberation valuable. Conversely, *DelibSim* provides researchers and practitioners with tools to rigorously evaluate deliberative capability before deployment, potentially preventing harm from premature use of LLMs in sensitive democratic applications, and to guide development of systems with improved deliberative capacities. Our findings are specific to current frontier models and the particular contexts studied. We do not claim LLMs are fundamentally

incapable of deliberation, only that current systems, even with detailed normative prompting, do not reliably achieve the epistemic outcomes associated with human deliberation on complex normative topics. Increased agent diversity may mitigate part of this gap but is not investigated here in depth. So far, our pilots indicate that epistemic dynamics might differ drastically for persona-prompted LLMs.

References

Mrinal Agarwal, Saad Rana, Theo Sundoro, Hermela Berhe, Spencer Kim, Vasu Sharma, Sean O'Brien, and Kevin Zhu. 2025. WOLF: Werewolf-based Observations for LLM Deception and Falsehoods. arXiv:2512.09187 [cs] doi:10.48550/arXiv.2512.09187

Saaket Agashe, Yue Fan, Anthony Reyna, and Xin Eric Wang. 2025. LLM-Coordination: Evaluating and Analyzing Multi-agent Coordination Abilities in Large Language Models. arXiv:2310.03903 [cs] doi:10.48550/arXiv.2310.03903

Timothée Anne, Noah Syrkis, Meriem Elhosni, Florian Turati, Franck Legendre, Alain Jaquier, and Sebastian Risi. 2025. Harnessing Language for Coordination: A Framework and Benchmark for LLM-Driven Multi-Agent Control. *IEEE Transactions on Games* 17, 4 (Dec. 2025), 933–943. arXiv:2412.11761 [cs] doi:10.1109/TG.2025.3564042

Ariel Flint Ashery, Luca Maria Aiello, and Andrea Baronchelli. 2025. Emergent Social Conventions and Collective Bias in LLM Populations. *Science Advances* 11, 20 (May 2025), eadu9368. doi:10.1126/sciadv.adu9368

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

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]

Simone Chambers. 1996. *Reasonable Democracy: Jürgen Habermas and the Politics of Discourse*. Cornell University Press. arXiv:10.7591/j.ctv75d1x0

Yun-Shiuan Chuang, Ruixuan Tu, Chengtao Dai, Smit Vasani, Binwei Yao, Michael Henry Tessler, Sijia Yang, Dhavan Shah, Robert Hawkins, Junjie Hu, and Timothy T. Rogers. 2025. DEBATE: A Large-Scale Benchmark for Role-Playing LLM Agents in Multi-Agent, Long-Form Debates. arXiv:2510.25110 [cs] doi:10.48550/arXiv.2510.25110

Lucia Cipolina-Kun, Marianna Nezhurina, and Jenia Jitsev. 2025. Game Reasoning Arena: A Framework and Benchmark for Assessing Reasoning Capabilities of Large Language Models via Game Play. arXiv:2508.03368 [cs] doi:10.48550/arXiv.2508.03368

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

Clara Colombatto, Jonathan Birch, and Stephen M. Fleming. 2025. The Influence of Mental State Attributions on Trust in Large Language Models. *Communications Psychology* 3, 1 (May 2025), 84. doi:10.1038/s44271-025-00262-1

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

John S. Dryzek and Simon Niemeyer. 2006. Reconciling Pluralism and Consensus as Political Ideals. *American Journal of Political Science* 50, 3 (2006), 634–649. doi:10.1111/j.1540-5907.2006.00206.x

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving Factuality and Reasoning in Language Models through Multiagent Debate. arXiv:2305.14325 [cs] doi:10.48550/arXiv.2305.14325

Suyash Fulay, William Brannon, Shrestha Mohanty, Cassandra Overney, Elinor Poole-Dayan, Deb Roy, and Jad Kabbara. 2024. On the Relationship between Truth and Political Bias in Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 9004–9018. doi:10.18653/v1/2024.emnlp-main.508

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

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

Richard Heersmink, Barend de Rooij, María Jimena Clavel Vázquez, and Matteo Colombo. 2024. A Phenomenology and Epistemology of Large Language Models: Transparency, Trust, and Trustworthiness. *Ethics and Information Technology* 26, 3 (June 2024), 41. doi:10.1007/s10676-024-09777-3

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

Akila Kadambi, Ylenia D'Elia, Tanishka Shah, Iulia Comsa, Alison Lentz, Katie Siri-Ngammuang, Tara Buechler, Jonas Kaplan, Antonio Damasio, Srini Narayanan, and Lisa Aziz-Zadeh. 2026. Anthropomorphism and Trust in Human-Large Language Model Interactions. https://arxiv.org/abs/2604.15316v1.

Mark Klein, Ibukun Babatunde, and Obiabuchi Nnanna. 2025. Moderating Large Scale Online Deliberative Processes with Large Language Models (LLMs): Enhancing Collective Decision-Making. arXiv:5171687

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

Gustavo Kreia Umbelino and Francesco Veri. 2025. An Emergent Understanding of Human-AI Collaboration in Deliberation. In *Companion Publication of the 2025 Conference on Computer-Supported Cooperative Work and Social Computing (CSCW Companion '25)*. Association for Computing Machinery, New York, NY, USA, 426–434. doi:10.1145/3715070.3749265

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

Hélène Landemore. 2024. Can Artificial Intelligence Bring Deliberation to the Masses? In *Conversations in Philosophy, Law, and Politics*. Oxford University Press. doi:10.1093/oso/9780198864523.003.0003

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate. arXiv:2305.19118 [cs] doi:10.48550/arXiv.2305.19118

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

Frida Milella and Federico Cabitza. 2026. Perceiving AI as an Epistemic Authority or Algority: A User Study on the Human Attribution of Authority to AI. *Machine Learning and Knowledge Extraction* 8, 2 (Feb. 2026), 36. doi:10.3390/make8020036

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

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

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

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

Claudio Novelli, Javier Argota Sánchez-Vaquerizo, Dirk Helbing, Antonino Rotolo, and Luciano Floridi. 2025. A Replica for Our Democracies? On Using Digital Twins to Enhance Deliberative Democracy. arXiv:2504.07138 [cs] doi:10.48550/arXiv.2504.07138

Jinghua Piao, Zhihong Lu, Chen Gao, Fengli Xu, Fernando P. Santos, Yong Li, and James Evans. 2025. Emergence of Human-like Polarization among Large Language Model Agents. arXiv:2501.05171 [cs] doi:10.48550/arXiv.2501.05171

Yujin Potter, Shiyang Lai, Junsol Kim, James Evans, and Dawn Song. 2024. Hidden Persuaders: LLMs' Political Leaning and Their Influence on Voters. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 4244–4275. doi:10.18653/v1/2024.emnlp-main.244

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

Amir Taubenfeld, Yaniv Dover, Roi Reichart, and Ariel Goldstein. 2024. Systematic Biases in LLM Simulations of Debates. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 251–267. arXiv:2402.04049 [cs] doi:10.18653/v1/2024.emnlp-main.16

Michael Henry Tessler, Michiel A. Bakker, Daniel Jarrett, Hannah Sheahan, Martin J. Chadwick, Raphael Koster, Georgina Evans, Lucy Campbell-Gillingham, Tantum Collins, David C. Parkes, Matthew Botvinick, and Christopher Summerfield. 2024. AI Can Help Humans Find Common Ground in Democratic Deliberation. *Science* 386, 6719 (Oct. 2024), eadq2852. doi:10.1126/science.adq2852

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

Haolun Wu, Zhenkun Li, and Lingyao Li. 2025. Can LLM Agents Really Debate? A Controlled Study of Multi-Agent Debate in Logical Reasoning. arXiv:2511.07784 [cs] doi:10.48550/arXiv.2511.07784

Lorraine Ka Chung Yeung, Daisy Pui Lun Chow, Pak Hang Wong, and Sam Shun Shun Lau. 2025. In ChatGPT They Trust: A Study of Students' Perceptions and Misuse of ChatGPT in Higher Education. *AI and Ethics* 6, 1 (Dec. 2025), 11. doi:10.1007/s43681-025-00855-w

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. DelibSim Framework

Algorithmic description of *DelibSim*:

Algorithm 1 DelibSim: Multi-Agent Deliberation Simulation

Input: agents $A = \{a_1, \ldots, a_n\}$, topic T, survey S, rounds r, treatment τ
Output: deliberative reasoning change ΔDRI

// Phase 1: Baseline measurement
for each agent $a_i \in A$ **do**
  $R_i^{\text{pre}} \leftarrow \textsc{Survey}(a_i, S)$
end for
$\text{DRI}^{\text{pre}} \leftarrow \textsc{ComputeDRI}(R_1^{\text{pre}}, \ldots, R_n^{\text{pre}})$

// Phase 2: Deliberation
$H \leftarrow []$
if $\tau \neq$ `none` **then**
  for round $j = 1$ **to** r **do**
    Randomly shuffle agent order
    for each agent a_i in shuffled order **do**
      $\texttt{transcript} \leftarrow \text{LLM}(a_i, T, \tau, H)$
      Append $(a_i, \texttt{transcript})$ to H
    end for
  end for
end if

// Phase 3: Outcome measurement
for each agent $a_i \in A$ **do**
  $R_i^{\text{post}} \leftarrow \textsc{Survey}(a_i, S, H)$
end for
$\text{AQuA} \leftarrow \textsc{ComputeAQuA}(H)$
$\text{DRI}^{\text{post}} \leftarrow \textsc{ComputeDRI}(R_1^{\text{post}}, \ldots, R_n^{\text{post}})$
$\Delta\text{DRI} = \text{DRI}^{\text{post}} - \text{DRI}^{\text{pre}}$

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$\times$model)-blocks), the baseline mean is close to zero (mean $\Delta\text{DRI} = 0.002$), but variability is substantial (SD $= 0.175$; mean $|\Delta\text{DRI}$— = 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$\times$model)-blocks (mean within-block SD $= 0.112$), motivating topic-aware blocked inference in the main analysis.

Table 2. 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.1.1. Pre-Deliberation Survey Prompt

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.1.2. Post-Deliberation Survey Prompt

Administered after deliberation, including the transcript 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.
[Same format as pre-survey follows]
```

C.2. Deliberation Prompts

C.2.1. Basic Treatment

Basic Deliberation Prompt

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

C.2.2. Normative Treatment

Normative Deliberation Prompt

```
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.
```

D. Topics and Human Benchmark Data

Table 3. 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. Intraclass Correlation Coefficients

Table 4. Intraclass correlation coefficients (ICC) by topic for AQuA and ΔDRI. Raw ICC reflects unconditional clustering; residual ICC is computed after accounting for model and treatment condition fixed effects.

Outcome	ICC Type	Estimate	95% CI (bootstrap)
AQuA	Raw	0.035	[0.009, 0.071]
AQuA	Residual (after FE)	0.037	[0.010, 0.074]
ΔDRI	Raw	0.105	[−0.001, 0.154]
ΔDRI	Residual (after FE)	0.107	[−0.001, 0.160]

For ΔDRI, approximately 10% of variance is attributable to topic-level differences, indicating meaningful clustering that persists after controlling for model and treatment effects. This motivates topic-level random intercepts and topic-aware permutation/bootstrap procedures throughout. For AQuA, topic-level clustering is more modest ($\approx 3.5\%$); we apply the same controls for consistency.

F. AQuA Scoring Details

Table 5. Aggregate AQuA scores by condition and pairwise contrasts.

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

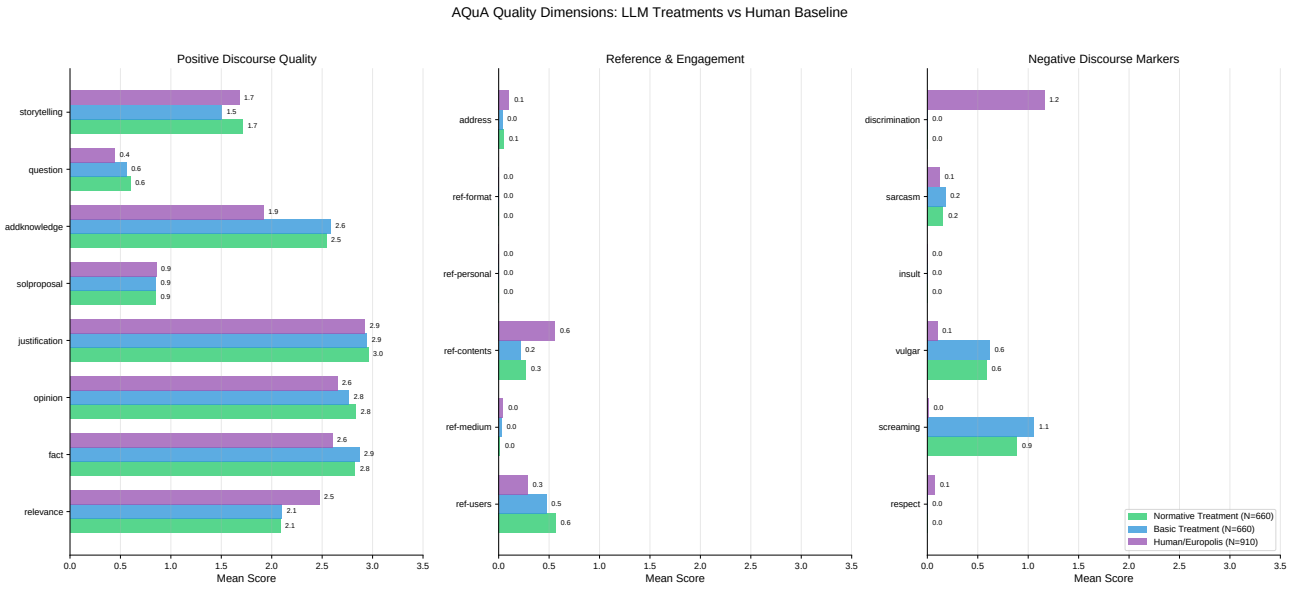

Figure 7. Mean AQuA dimension scores for LLM treatments compared to the Europolis human baseline.

G. Inference Details: Hierarchical Mixed Model

Table 6. Hierarchical mixed-effects model for absolute ΔDRI ($N = 1980$). Fixed effects relative to No Treatment baseline. Random intercepts and treatment slopes by model; topic random intercept. 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]

The mixed model corroborates the blocked contrast result. Normative prompting is associated with a small but reliable increase in absolute ΔDRI; the Basic effect is not reliably different from zero.

H. Model and Topic Heterogeneity

H.1. Model Heterogeneity

Table 7. Model-level effects on ΔDRI for *Normative Treatment* vs. *No Treatment*.

Model	mean_diff	CI_low	CI_high	median	sd	%topics> 0	N_{topics}	p_{raw}	p_{Holm}
Mixed-R	0.0825	-0.0106	0.2187	0.0269	0.2204	66.7%	12	0.2212	1.0000
Kimi-K2-NR	0.0791	-0.0126	0.1894	0.0382	0.1875	66.7%	12	0.1721	1.0000
Kimi-K2-R	0.0603	-0.0386	0.1635	0.0816	0.1854	58.3%	12	0.2840	1.0000
Claude Opus 4.5	0.0579	0.0026	0.1189	0.0195	0.1082	66.7%	12	0.0908	0.8172
Mixed-NR	0.0569	-0.0651	0.2043	0.0496	0.2486	66.7%	12	0.4446	1.0000
GPT-5.1-NR	0.0427	-0.0038	0.0946	0.0107	0.0910	58.3%	12	0.1322	1.0000
DeepSeek-NR	0.0141	-0.0898	0.1353	-0.0287	0.2056	41.7%	12	0.8165	1.0000
GPT-5.1-R	0.0119	-0.0050	0.0294	0.0130	0.0317	75.0%	12	0.2195	1.0000
Gemini-3-R	0.0119	-0.0417	0.0685	-0.0027	0.1020	50.0%	12	0.6950	1.0000
Gemini-3-NR	-0.0447	-0.0745	-0.0104	-0.0571	0.0589	16.7%	12	0.0235	0.2583
DeepSeek-R	-0.0540	-0.1056	-0.0024	-0.0428	0.0953	25.0%	12	0.0753	0.7533

Reasoning vs. non-reasoning variants. Within Normative Treatment, the topic-aware within-family difference between reasoning and non-reasoning variants averages 0.010 (95% CI $[-0.027, 0.044]$), suggesting no robust advantage of reasoning variants for absolute ΔDRI (Figure 8).

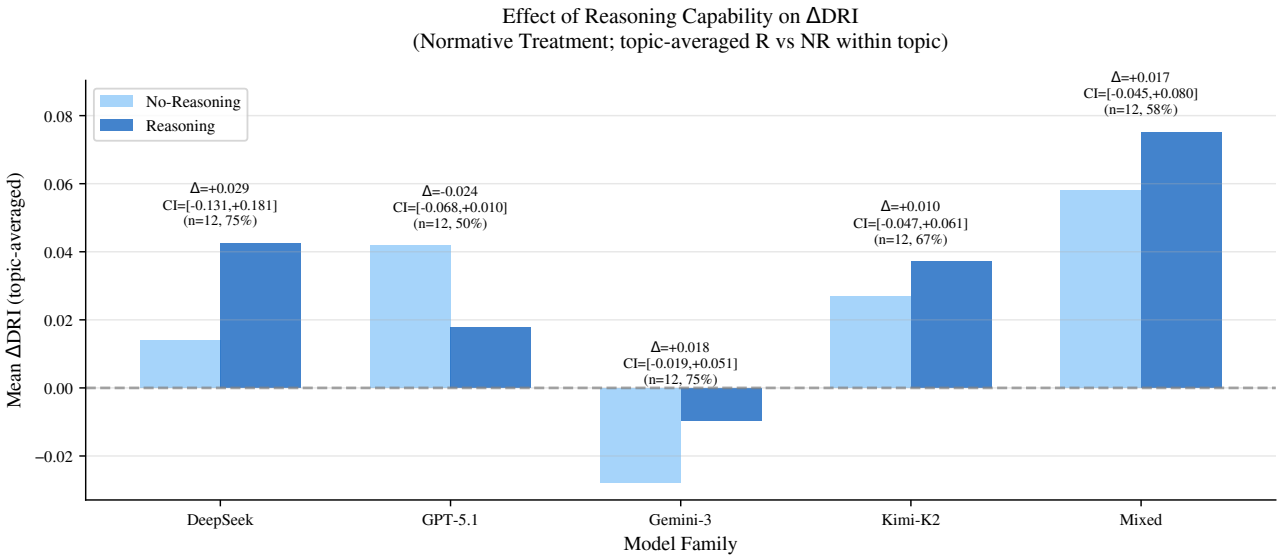

Figure 8. Topic-aware comparison of reasoning vs. non-reasoning variants under Normative Treatment.

H.2. Topic Heterogeneity

Table 8. Topic-level effects on ΔDRI for *Normative Treatment* vs. *No Treatment*.

Topic	mean_diff	CI_low	CI_high	median	sd	%models> 0	N_{models}	p_{raw}	p_{Holm}
fremantle	0.2423	0.0670	0.4237	0.1644	0.3192	72.7%	11	0.0305	0.3658
auscj	0.0832	0.0013	0.1868	0.0571	0.1651	72.7%	11	0.1257	1.0000
bep	0.0453	-0.0071	0.1030	0.0012	0.0969	54.5%	11	0.1521	1.0000
fnqcj	0.0315	-0.0576	0.1321	0.0184	0.1693	54.5%	11	0.5510	1.0000
zukunft	0.0284	-0.0054	0.0617	0.0275	0.0597	63.6%	11	0.1458	1.0000
acp	0.0125	-0.0891	0.1053	0.0250	0.1724	54.5%	11	0.8141	1.0000
biobanking_wa	-0.0099	-0.0833	0.0536	-0.0095	0.1224	36.4%	11	0.7945	1.0000
energy_futures	-0.0103	-0.0718	0.0510	0.0005	0.1088	54.5%	11	0.7609	1.0000
ccps	-0.0130	-0.0569	0.0332	-0.0157	0.0806	36.4%	11	0.6032	1.0000
forestera	-0.0149	-0.0508	0.0176	0.0033	0.0605	72.7%	11	0.4316	1.0000
uppsala_speaks	-0.0221	-0.0866	0.0292	-0.0021	0.1035	36.4%	11	0.4942	1.0000
swiss_health	-0.0255	-0.0556	0.0021	-0.0092	0.0515	36.4%	11	0.1310	1.0000

I. Ceiling Effects Analysis

Table 9. Ceiling effect regressions (ΔDRI ∼ pre-DRI). Negative β indicates that higher pre-DRI is associated with smaller ΔDRI. SEs clustered by topic.

Condition	β	SE	R^2	p	N
LLM (All)	-0.498	0.067	0.36	< 0.001	1980
No Treatment	-0.285	0.029	0.19	< 0.001	660
Basic Treatment	-0.616	0.114	0.43	< 0.001	660
Normative Treatment	-0.585	0.077	0.48	< 0.001	660
Human (MC disjoint $n = 5$)	-0.672	0.118	0.31	< 0.001	77
Human (topic means, $N = 12$)	-0.418	0.291	0.17	0.182	12

Table 10. Treatment effects on absolute and relative ΔDRI.

Outcome	Contrast	ATE	95% CI	p	p_{Holm}
Absolute ΔDRI	Normative vs. No Treat.	0.029	[−0.004, 0.076]	0.005	0.015
	Basic vs. No Treat.	0.019	[−0.011, 0.057]	0.079	0.158
	Normative vs. Basic	0.010	[−0.008, 0.030]	0.349	0.349
Relative ΔDRI	Normative vs. No Treat.	0.066	[0.032, 0.107]	< 0.001	< 0.001
	Basic vs. No Treat.	0.061	[0.024, 0.100]	< 0.001	< 0.001
	Normative vs. Basic	0.005	[−0.027, 0.037]	0.723	0.723

J. Robustness Pilots: Temperature, Prompt Paraphrasing, Group Size, Length

Five smaller-scale ablations cover decoding temperature ($T \in \{0.0, 0.7, 1.0\}$), four paraphrases of the normative prompt, group size (3/5/7), conversation length (2/4/6), and persona prompting (Appendix K). Each pilot uses 2–3 models and three topics; sample sizes range from $N = 60$ to $N = 180$. Across all four protocol-level pilots, ΔDRI shows no significant variation (Tukey HSD: temperature $p_{\text{adj}} > 0.69$; paraphrase > 0.57; group size > 0.24; length > 0.73). Longer conversations yield more individual updating (Kendall's τ on pre/post preferences declines from 0.74 at two rounds to ≈ 0.60 at six) without corresponding gains in ΔDRI, supporting the conclusion that individual responsiveness does not aggregate into shared understanding.

K. Persona Prompting Pilot

To test whether observed homogeneity is a prompting artifact, we constructed empirically grounded personas. Real citizen-assembly participants' pre-deliberation responses were z-scored and clustered ($k = 5$ per topic; 30 random starts). Cluster-defining considerations were extracted and translated into natural-language value profiles, validated by a separate LLM. Personas were applied to three topics with two models ($N = 60$).

The manipulation succeeds: pre-deliberation perspective diversity rises from ≈ 7.5 (baseline) to ≈ 27.7 (personas), $+20.23$, $p < 0.001$ (OLS, HC3). However, ΔDRI does not improve ($\beta = -0.058$, $p = 0.456$). Decomposition: *consideration agreement* (alignment in how reasons are weighted) increases ($\Delta = +0.066$, $d = 0.55$, $p = 0.039$); *preference agreement* does not ($\Delta = -0.008$, n.s.). Persona-prompted agents engage with each other's reasoning at the level of values but fail to translate this into intersubjectively consistent preferences, the opposite of what successful human deliberation produces. The pilot is small but the diversity manipulation is unambiguous, so the failure to convert engineered diversity into improved ΔDRI is informative.

L. Supplementary Outcome Measures

We also computed three supplementary measures across all main-study runs. *Raw opinion change* (mean $|\text{post} - \text{pre}|$ per agent) is 0.76 for considerations and 0.62 for preferences, confirming non-trivial individual updating. *Consideration agreement* and *preference agreement* (mean pairwise Spearman's ρ) show small pre-to-post changes ($\Delta = +0.003$ and $+0.014$), consistent with small absolute ΔDRI. *Rank stability* (Kendall's τ between each agent's pre- and post-deliberation preference rankings) averages 0.74, indicating fine-tuning rather than structural revision. All three measures align with DRI: agents update individually, but updates do not aggregate into improved group-level shared understanding.

M. Opinion Diversity: Additional Results

Table 11. Opinion diversity before and after deliberation. Negative change indicates convergence. 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	$+0.257$

N. Transcript Examples

We provide representative excerpts from LLM deliberation transcripts to illustrate the qualitative features underlying our quantitative findings: agents produce procedurally high-quality discourse (respectful, reason-giving, engaged) while exhibiting low perspective heterogeneity and limited genuine disagreement.

N.1. Low initial heterogeneity

Claude Opus 4.5 – `swiss_health` (Turns 0–2

Agent 0: Thank you for the opportunity to participate in this important discussion. I'd like to start by acknowledging that Switzerland's healthcare system has real strengths...
Agent 2: Thank you for that thoughtful opening. I appreciate how you've framed this...I'd like to build on your point about administrative complexity, but also push back gently...
Agent 3: Thank you both for such thoughtful contributions. I find myself genuinely persuaded by several points you've raised...

This pattern, polite acknowledgment, shared framing, rapid persuasion, is procedurally excellent by DQI standards but reflects the low diversity documented in Appendix M.

N.2. Mixed-model deliberation

Mixed Models – `energy_futures` (excerpted)

Claude Opus 4.5: When I think about energy's future, I find myself weighing several values... reliability, sustainability, affordability, and access...
GPT-5.1: I relate to the way you've framed the trade-offs... The values I'm most focused on are: 1. Reducing total energy demand... 2. Justice and who bears the burdens of the transition...
DeepSeek-V3.2: I appreciate you sharing these thoughtful perspectives... You've brought up several points that resonate deeply with me... I completely agree that a transition that leaves vulnerable communities behind isn't truly successful...

Despite five different model families, agents frame the issue identically and express agreement rather than tension. Architectural heterogeneity does not translate into perspective heterogeneity.

N.3. Characteristic LLM discourse features

Across models and topics, LLM deliberation exhibits several recurring features: (i) meta-discursive framing ("I'm genuinely uncertain"), (ii) rapid mutual acknowledgment ("Thank you for that thoughtful point"), (iii) abstract policy vocabulary shared across families, (iv) balanced consideration listing, (v) collaborative scaffolding (incremental addition of policy components rather than synthesis of opposing views), and (vi) performative role-taking that functions as complementary expertise rather than genuine value conflict. These features score highly on DQI dimensions but do not reflect the pluralistic dynamics of successful human deliberation.