

The Deliberative Deficit: An Empirical Critique of LLMs in Democratic Discourse

Maurice Flechtner

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

Abstract

Large language models (LLMs) are increasingly deployed in settings that require collective reasoning on complex, value-laden problems. Confidence in these deployments rests largely on benchmarks for *verifiable* tasks (mathematics, coding, coordination games), yet many of these applications concern problems where no objectively correct answer exists and where decision quality instead depends on integrating pluralistic perspectives to find mutually acceptable solutions. We argue that LLM reasoning capacity on this class of problems cannot be fully inferred from verifiable-task benchmarks, and that procedural evaluations of LLM discourse (respectfulness, justification, engagement) are systematically insufficient. We apply the Deliberative Reason Index (DRI), a measure developed in political science and validated across citizen assemblies, as a tool for evaluating *reliable group-level reasoning on pluralistic, non-verifiable problems*. Synthesizing recent evidence across 1,980 five-agent LLM runs on 12 citizen-assembly topics across 11 frontier model configurations, we find that LLM groups produce discourse with procedural quality comparable to human deliberation, while gains in intersubjective consistency are small, topic-dependent, and concentrated on tractable rather than ethically contested questions. LLM groups exhibit roughly one-third the perspective diversity of human assemblies and reverse the human convergence pattern: human deliberation decreases dispersion as diverse views synthesise, whereas LLM deliberation increases it. Engineering diversity through persona prompting does not restore the human dynamic but inverts which component of deliberative reasoning is updated. We argue this warrants a critique of the deliberative capacity of LLMs that is missing from current benchmarks and dangerous for deployment. Our conclusion is constraining rather than prohibitive: LLMs can function as tools supporting human reasoning on pluralistic problems, but current evidence does not license treating them as autonomous deliberative agents.

Code (benchmark study) —

<https://doi.org/10.1145/3772363.3798499>

Extended version — <https://arxiv.org/abs/2608.10186>

1 Introduction

Large language models (LLM) are increasingly proposed for collective reasoning on consequential, value-

laden problems. Recent proposals include AI-mediated consensus-finding in democratic deliberation (Tessler et al. 2024), AI representation of missing perspectives in policy consultations (Fulay, Dimitrakopoulou, and Roy 2025; Zhu et al. 2025), systematically AI-augmented citizen assemblies (Landemore 2024; McKinney 2024), large-scale moderation of online deliberation (Klein, Babatunde, and Nnanna 2025), and simulations of democratic systems (Rountree and Gastil 2026; Novelli et al. 2025). Proponents argue these systems could help scale the deliberative quality of small group deliberation to mass publics (Landemore 2024; Lazar and Manuali 2026) while sceptics worry about technosolutionism and the erosion of democratic substance (Oleart and Palomo 2025; Summerfield et al. 2025). However, as of now, the debate is missing a discussion of the question of whether LLMs can, in some meaningful sense, *reason* about the kinds of problems deliberation addresses.

This premise is rarely tested directly. Confidence in LLM reasoning capacity rests primarily on benchmarks for tasks with verifiable solutions: mathematics and chess (Du et al. 2023), logical puzzles (Liang et al. 2024; Wu, Li, and Li 2025), and coordination games (Anne et al. 2025). Reinforcement learning with verifiable rewards (RLVR) has driven most recent progress in LLM reasoning (DeepSeek-AI et al. 2025), and the field itself acknowledges that its advancements are mostly confined to domains with well-defined outcomes (Zhang et al. 2026a).

Yet the collective reasoning problems for which LLMs are being deployed in democratic contexts are not verifiable tasks. Citizen assemblies on climate policy, stakeholder negotiations on resource allocation, participatory design processes, or ethics consultations on new technologies are all problems where no objectively correct answer exists, where legitimate perspectives diverge, and where decision quality depends on how well diverse considerations are integrated into a shared reasoning framework. Deliberation theory has long recognized that the goal of reasoning on such problems is not correctness but *mutually acceptable, well-reasoned conclusions* that integrate pluralistic considerations (Niemeyer and Dryzek 2007; Dryzek and Niemeyer 2006) and that the epistemic value of this integration depends constitutively on the diversity of perspectives brought into it (Landemore 2013). We refer to this class of non-verifiable, pluralistic and often value-laden problems as

pluralistic reasoning problems, and argue they constitute a significant challenge for LLM evaluation today.

When verifiable metrics do not apply, evaluation often falls back on *procedural* measures. In the case of democratic deliberation, this would revolve around whether the system produces justified, respectful, reciprocal discourse. Such measures, as seen in the Discourse Quality Index (Steenbergen et al. 2003) and its automated variant AQUA (Behrendt et al. 2024), capture *how* agents talk, not *whether they reason together*, and say nothing about whose perspectives are in the room to be reasoned about. Human deliberation research has found that procedural and substantive quality can diverge (Knobloch and Gastil 2022; Baccaro, Bächtiger, and Deville 2016) and recent work on LLM-as-judge evaluation has independently documented a “shared illusion” in which LLM judges agree on surface heuristics while missing substantive quality (Song, Zheng, and Xu 2026). The simulation of judgment literature has coined *epistemia* for the illusion of knowledge when plausibility replaces verification (Loru et al. 2025), and a recent roadmap for evaluating moral competence in LLMs identifies the *facsimile problem* of “models imitating reasoning without genuine understanding” as a core challenge (Haas et al. 2026). We argue that there is a similar risk for LLMs deployed in multi-agent deliberation: the risk that procedural appearance will be mistaken for epistemic substance and that diversity-blind evaluation will mistake convergence among the already-similar for integration across difference.

This paper makes three contributions. First, we argue that pluralistic reasoning problems constitute an evaluation gap that neither verifiable-task benchmarks nor procedural discourse metrics fully address. The former do not generalise to the class, the latter capture how agents talk rather than whether they reason together. Second, we assemble a three-dimensional necessary-condition test from deliberative practice and validated measurement instruments, spanning procedural quality, outcome quality, and diversity conditions. The Deliberative Reason Index (DRI) (Niemeyer et al. 2024), a group-level relational measure of intersubjective consistency validated across nineteen citizen-assembly cases, supplies the outcome dimension and, through its response vectors, the basis for diversity assessment. While recent work administered it to individual LLMs (Kreia Umbelino and Veri 2025), we put it to use at the multi-agent group level. Passing all three dimensions does not establish deliberative capacity, but failing any of them precludes it, making the test a relevant constraint on deployment claims. Third, we draw on empirical evidence from our ongoing research programme on simulating and evaluating LLM deliberation. Across 1,980 five-agent runs on 12 citizen-assembly topics and 11 frontier model configurations, the evaluation reveals that LLM groups achieve human-level scores on process quality but produce only small, topic-dependent DRI gains concentrated on tractable rather than ethically contested questions. They begin with roughly one-third of human starting diversity, and invert human perspective convergence dynamics. Our findings suggest that LLM collective reasoning on pluralistic reasoning problems is *procedurally excellent but epistemically shallow*.

We thus draw a line at the distinction between LLMs as *tools* (extending human reasoning, with humans retaining epistemic authority) and LLMs as *epistemic agents* (producing reasoning contributions treated as carrying independent epistemic weight) (Hauswald 2025; Freiman 2023). The evidence we review is not a blanket case against AI in deliberation, it supports a targeted argument against a specific assumption: that current LLMs can autonomously integrate pluralistic perspectives into collective reasoning in a reliable way. Several prominent deployments make this assumption implicitly (Rountree and Gastil 2026; Tessler et al. 2024; Fulay, Dimitrakopoulou, and Roy 2025; Fish et al. 2024). Our argument suggests these deployments rest on a justificatory gap between the deliberative weight their outputs are considered to have and the epistemic shortcomings our evaluation demonstrates.

We have tested the argument empirically in the political deliberation case because that is where validated DRI instruments exist. Conceptually, the point likely transfers to domains characterized by pluralism and non-verifiability, and adaptation to such domains is a priority for future work.

2 Evaluating Collective Reasoning

2.1 What current benchmarks measure

Current benchmarks for collective reasoning in multi-agent LLM systems share a defining property: they evaluate against *verifiable solutions*. Whether the task revolves around factual reasoning (Du et al. 2023), logical reasoning (Liang et al. 2024; Wu, Li, and Li 2025), multi-agent coordination (Anne et al. 2025; Agashe et al. 2025), or strategic reasoning in game-like environments (Cipolina-Kun, Nezhurina, and Jitsev 2025; Agarwal et al. 2025), the common structure is that ground truth exists and can be externally checked, or that success has a mathematical or game-theoretic definition. The same property is an important condition for the post-training paradigm that drives current reasoning progress: reinforcement learning with verifiable rewards (DeepSeek-AI et al. 2025; OpenAI et al. 2024).

Performance on these tasks establishes LLM capacity for a specific kind of reasoning, where reasoning is about finding an answer that can, in principle, be checked. In these situations, multi-agent LLM systems often perform by validating and improving upon the answer of one of the agents (Liang et al. 2024). Pluralistic reasoning problems do not offer an ideal solution that can be verified, making the paradigm of iterative improvement much harder to employ.

2.2 Pluralistic reasoning

A substantial class of real-world collective reasoning problems lies outside this scope. Political deliberation on climate policy, healthcare reform, migration, or bioethics has no verifiable answer: perspectives legitimately diverge based on underlying values and framings, and decision quality depends on how well those differences are considered and integrated (Niemeyer and Dryzek 2007; Dryzek and Niemeyer 2006; Mouffe 1999). The same structure characterizes organizational decisions under value conflict, stakeholder negotiations, participatory design, or ethics consultations on

emerging technologies. In deliberative-democratic terms, the goal in these settings is not to find a consensus but to construct a *meta-consensus*, i.e. a shared understanding of how relevant considerations map to preferences in order to make more informed and consistent decisions (Niemeyer and Dryzek 2007). Meta-consensus is distinct from agreement on conclusions as participants can disagree sharply about what should be done while sharing an understanding of why they disagree.

Prominent AI-for-democracy deployments target precisely these problems, including AI-mediated consensus-finding (Tessler et al. 2024), representation of missing perspectives (Fulay, Dimitrakopoulou, and Roy 2025; Zhu et al. 2025), and deliberation simulation (Rountree and Gastil 2026). Similar evaluation challenges arise across many kinds of human-AI decision making on pluralistic reasoning problems (Ma et al. 2025; Summerfield et al. 2025).

The RLVR literature itself acknowledges that verifiable-task techniques do not reach this class: moral reasoning “typically admit[s] multiple valid answers that reflect different ethical frameworks and value systems, in stark contrast to mathematical and coding problems, which usually have only one objectively correct solution” (Zhang et al. 2026b) and it is being acknowledged that “success has thus far been largely confined to the mathematical and programming domains with clear and automatically checkable outcomes”, leaving the extension of these techniques to open-ended, value-laden problems an open research question (Zhang et al. 2026a). We take this as the field’s own diagnosis that performance on verifiable tasks does not guarantee capabilities on pluralistic tasks, and there is no principled reason to expect the transfer.

2.3 Process quality is insufficient

An easy fall back option in situations where verifiable outcome metrics do not apply are procedural measures. These look at the process of coming up with a solution to evaluate the collective reasoning process and infer outcome quality by arguing that high process quality guarantees high outcome quality. In deliberative practice, process quality can be assessed by the Discourse Quality Index (DQI) (Steenbergen et al. 2003). It was developed for human parliamentary deliberation and scores utterances on justification, respect, constructive politics, and engagement with opposing positions. AQuA (Behrendt et al. 2024) automates DQI scoring using an ensemble of adapter models trained on and validated against human coding. These measures capture real dimensions of discourse quality, but they capture *how* agents talk, not *whether they reason together*.

Human deliberation research has documented that procedural and substantive quality can diverge, meaning that well-facilitated, respectful deliberations can fail to produce epistemic gains (Knobloch and Gastil 2022; Baccaro, Bächtiger, and Deville 2016). The same structural concern has been articulated for LLM evaluation more broadly. The Nature roadmap for moral competence in LLMs identifies the *fac-simile problem* of models imitating reasoning without engaging it as a foundational challenge (Haas et al. 2026). Parallel diagnoses appear in LLM-as-judge research where the

“evaluation illusion” is described as a phenomenon where evaluators converge on surface heuristics instead of substantive quality (Song, Zheng, and Xu 2026)) and in studies of LLM judgment where “the illusion of knowledge emerging when plausibility replaces verification” has been identified as a key failure mode for reliable LLM judgements (Loru et al. 2025)).

For multi-agent LLM deliberation, this poses a significant risk. If LLM systems produce procedurally high-quality discourse on pluralistic problems while failing to integrate the underlying reasoning, procedural metrics alone cannot distinguish genuine deliberative capacity from its surface imitation. Deliberative LLM agents can then easily be mistaken for epistemic authorities where they actually just mimic deliberative talk without engaging with the underlying reasoning on the issue at hand.

2.4 Diversity

Procedural and outcome metrics together evaluate the deliberation *among the participants who are there*, they say nothing about whether those participants span the perspectives the topic demands. For human deliberation this question is typically handled through stratified random selection at the recruitment stage. This diversity is not only meant to ensure representativeness of the mini-public but also, and maybe even more importantly, to ensure cognitively diverse groups. This diversity is what makes deliberation epistemically valuable as the integration of diverse perspectives is considered to help find the best possible solution to a given issue (Landemore 2013).

For multi-agent LLM systems random selection is not an option. Models trained on overlapping corpora start from substantially more similar priors than randomly selected human participants. A growing literature documents that, even when prompted to take specific personas, LLM agents homogenize during debate (Taubenfeld et al. 2024), that co-authoring with LLMs shifts human expression toward homogenized linguistic and cultural patterns (Sourati, Ziabari, and Dehghani 2026), and that LLM-based social simulation outcomes tend towards unanimity and utopian bias while human data exhibits genuine plurality (Bian et al. 2025; Chen et al. 2026). While architectural variation across model families produces some minor differences it therefore does not produce the kind of perspective heterogeneity that human deliberation presupposes.

The implication for evaluation is direct. A multi-agent LLM deliberation that scores well on procedural quality and shows outcome convergence has not necessarily *integrated diverse perspectives*, it may have aligned positions that were never meaningfully apart. Without an explicit diversity measurement, the procedural and outcome dimensions cannot distinguish integration across difference from convergence among the already-similar. This is what motivates treating diversity as a co-equal evaluation dimension rather than a recruitment assumption.

3 Evaluation Framework and Setup

3.1 From critique to measurement

Section 2.3 argued that procedural metrics alone leave the evaluation dimensions of epistemic outcome and diversity of perspectives uncovered. This section introduces measurement instruments for both deliberative reasoning and perspective diversity, drawing on instruments validated in deliberative-democracy research. We treat the resulting three dimensions as jointly necessary but individually insufficient. High procedural quality without outcome gains indicates form without substance, outcome gains without diversity indicate convergence among the already-similar rather than integration across difference, and diversity without procedural quality or outcome gains is parallel monologue rather than deliberation. The framework therefore constitutes a *necessary-condition test*. A system that fails on any of the above dimensions cannot be claimed to deliberate, while passing all three is evidence for deliberative capacity but not yet proof of it.

The remainder of this section introduces the outcome measure (§3.2), shows how the same instrument supplies a diversity measure (§3.3), and describes the experimental setup that applies all three to multi-agent LLM deliberation (§3.4).

3.2 The Deliberative Reason Index

Following Niemeyer, Veri and colleagues, we operationalise the outcome quality of a deliberation with *meta-consensus*. Meta-consensus refers to the shared understanding of how relevant considerations (i.e. values, perspectives, arguments and opinions) map to preferences (i.e. practical solutions) and does not require participants to agree on their conclusions (Niemeyer and Dryzek 2007; Dryzek and Niemeyer 2006). The construct is therefore pluralism-preserving by design. Two participants can disagree sharply on a policy while exhibiting high meta-consensus if they share an understanding of how the relevant considerations translate into preferences.

The Deliberative Reason Index (DRI) operationalises meta-consensus as *intersubjective consistency* between participants’ considerations and preferences (Niemeyer and Veri 2022; Niemeyer et al. 2024). Participants rate a set of consideration statements (typically 20–40 items drawn from public discourse on the topic) on a Likert scale and rank a smaller set of policy preferences, both before and after a deliberative intervention. For each pair of participants, Spearman correlations are computed separately across considerations and across preferences. Intersubjective consistency is the similarity of these two correlations. High consistency means that when two participants share or diverge in their reasoning about considerations, this is reflected proportionally in their shared or divergent preferences. Group-level DRI aggregates across all pairs:

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

where C_i and P_i are participant i ’s consideration and preference vectors, ρ_s is Spearman correlation, and $n_p = n(n -$

$1)/2$ is the number of unordered pairs. The absolute value prevents positive and negative pairwise discrepancies from cancelling, and the $2/n_p$ normaliser places DRI in $[-1, 1]$, with 1 at perfect intersubjective consistency. Comparing pre- and post-deliberation DRI scores allows us to understand the impact of a deliberative intervention on the collective reasoning of the group.

The property that makes DRI suitable for our framework is that it *distinguishes genuine epistemic gains from surface agreement*. Groups can produce consensus without meta-consensus (agreement on conclusions without shared reasoning). They can also produce high procedural quality without meta-consensus (well-mannered discussion that fails to integrate perspectives), and they can produce meta-consensus without consensus (shared reasoning amid persistent disagreement). DRI separates these cases cleanly, which is exactly the separation that procedural metrics alone cannot make. The measure is additionally content-agnostic as DRI requires no external judgment about what conclusions are good or which considerations matter and it is group-level relational as it captures a property of the group’s reasoning structure that no individual-level measure can. DRI has been validated across nineteen citizen-assembly-scale deliberations on topics including climate policy, healthcare, constitutional reform, and bioethics, where it reliably increases over successful deliberations and is robust to information provision, group conformity, and expert framing (Niemeyer et al. 2024).

3.3 The diversity metric

The diversity dimension is operationalised by vectorising the participants’ DRI responses. For a group of n participants with standardised response vectors $x_i \in \mathbb{R}^d$ spanning all consideration and preference items, we compute mean pairwise Euclidean distance:

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

Pre-deliberation diversity measures whether the group meets Landemore’s requirement that deliberation engages genuinely different starting positions (Landemore 2013). Pre-to-post change measures the *convergence dynamic*. Human deliberation typically reduces dispersion as diverse perspectives synthesise into shared understanding revolving around several clusters of opinions.

3.4 Experimental setup

A recent benchmark study applies the three-dimensional framework to multi-agent LLM deliberation (Flechtner 2026). The relevant aspects of its setup are summarised below. Full statistical specifications and robustness checks appear in the cited work. The deliberation and survey prompts it uses are reproduced in the extended version. The findings reviewed in §4 draw on this study, supplemented with a persona-prompting pilot introduced below and an individual-level study assessing LLM alignment with human DRI answers (Kreia Umbelino and Veri 2025).

Each deliberation in the study involves five LLM agents engaging on a single topic over two rounds, with speaking order randomised within rounds, the full transcript passed forward between turns, and decoding temperature fixed at zero for reproducibility. Each agent completes the DRI survey before and after deliberation, rating the topic’s consideration statements and ranking its policy preferences. Three prompting regimes are compared: a survey-only baseline that establishes same-system noise by moving from the pre- to the post-survey with no deliberation in between, a basic prompt that relies on the model’s internal understanding of the term, and a normative prompt that explicitly invokes the deliberative norms underpinning DQI and DRI. The full grid crosses eleven model configurations, twelve topics, and three treatments with five replicates each, for 1,980 runs.

The eleven model configurations span five frontier families (GPT-5.1, Gemini-3-Pro-Preview, Claude Opus 4.5, DeepSeek-V3.2-Exp, Kimi-K2-Thinking), with reasoning-enabled and standard variants where available, plus two mixed-family ensembles that draw each agent from a different family. The twelve topics are drawn from citizen assemblies with validated DRI instruments and matched human pre/post survey data, spanning climate policy, health-care, governance, bioethics, and urban planning.

Two human reference distributions anchor the analysis. Procedural quality is benchmarked against the Europolis deliberative poll (Gerber et al. 2018) ($N = 910$), the same corpus used to validate AQuA. Outcome quality and diversity are benchmarked against the human survey data from the twelve citizen assemblies ($N = 407$). Statistical inference is topic-aware throughout: the study reports topic-blocked permutation tests, topic-resampled bootstrap confidence intervals, and Holm correction for multiple comparisons, with a hierarchical mixed-effects model (topic random intercepts, model random slopes) as a complementary specification. This reflects substantial topic-level clustering ($ICC \approx 0.11$ for ΔDRI) against limited topic-level replication ($n = 12$).

An additional pilot tests whether the homogeneity observed across the main study reflects identical agent instructions rather than properties of the models themselves. Empirically grounded personas are constructed by clustering real citizen-assembly survey responses with $k = 5$ per topic (k-means on standardised consideration response vectors), then translating each cluster’s defining considerations into a natural-language value profile delivered as the agent’s system prompt. We cluster on considerations alone, not preferences, so that persona-prompted agents derive their preferences from the considerations they hold rather than anchoring on a fixed policy ranking, mirroring the direction of reasoning deliberation is meant to elicit. The pilot is run on three topics across two models ($N = 60$ deliberations). Full clustering details, persona-prompt templates, and example personas appear in the extended version.

4 Evidence: The Three-Dimensional Framework Applied

This section synthesizes evidence from the benchmark study of multi-agent LLM deliberation (Flechtner 2026),

<i>Outcome</i> — pooled Δ DRI contrasts				
Contrast	ATE	95% CI	p	p_{Holm}
Normative 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
Human reference	0.099	[0.026, 0.196]	—	—
<i>Procedural</i> — mean AQuA (0–4)				
Group	Mean	SD	N	
LLM (Basic + Norm.)	2.939	0.12–0.13	1,320	
Human (Europolis)	2.980	0.431	910	
<i>Diversity</i> — mean pairwise distance				
Group	Pre	Post	Δ	
Human (topic means)	18.78	17.57	−1.21	
LLM (Normative)	6.51	6.77	+0.26	
LLM (Basic)	6.56	6.93	+0.37	

Table 1: Three-dimensional evidence synthesised from the benchmark study (Flechtner 2026). Outcome CIs are topic-resampled bootstrap intervals, p -values are from topic-blocked permutation tests. AQuA dispersion is reported as SD, the LLM–human AQuA contrast gives $p_{\text{Holm}} = 0.052$. Diversity is mean pairwise Euclidean distance on standardised response vectors.

combined with an individual-level study of 54 LLMs against 526 human participants across 24 citizen-assembly cases (Kreia Umbelino and Veri 2025) and the persona-prompting pilot described above. The analysis will follow the three-dimensional framework introduced in §3. Here we summarise only the dimensions the argument turns on, reported against human reference distributions on the same instruments. Full model-by-model and topic-by-topic breakdowns and the protocol-level robustness pilots are reported in the benchmark study, where simulation and analysis code is provided as a supplement (Flechtner 2026).

4.1 Procedural quality

On procedural quality, LLM groups clear the deliberative bar AQuA operationalises. As AQuA scores discourse transcripts, it is only computed on the 1,320 sessions that involve deliberation (the basic and normative prompting regimes), not on the 660 survey-only baseline runs that make up the remainder of the 1,980-run study. Across those 1,320 sessions, mean AQuA (2.939) sits just below the Europolis human reference (2.980), a gap that does not reach significance after Holm correction ($p_{\text{Holm}} = 0.052$) (Flechtner 2026). We read this as procedural competence at human-comparable levels. LLM discourse also shows substantially lower variance (SD 0.12–0.13 vs. 0.43 for humans). Decomposition across the twenty AQuA dimensions shows that LLM discourse matches or approximates humans on justification, respect, engagement, and constructive politics, with negligible rates of disrespect, vulgarity, or sarcasm. By the procedural standard alone, LLM deliberation looks competent, so the framework’s first dimension is passed.

4.2 Outcome quality is small, topic-dependent, and concentrated on tractable problems

On outcome quality measured by DRI, the picture is very different. Under explicit normative prompting, multi-agent LLM deliberation produces only small absolute improvements in intersubjective consistency ($\Delta\text{DRI} \approx 0.029$, topic-blocked), roughly 30% of the human reference ($\Delta\text{DRI} \approx 0.099$). Under basic prompting that relies on the model’s internal understanding of “deliberation”, the effect is smaller and not reliably different from zero. The pooled effect under normative prompting is statistically detectable but does not survive Holm correction at the topic level, and turns *negative* on ethically contested topics such as Swiss healthcare reform and Uppsala begging policy (Flechtner 2026).

The topic-heterogeneity pattern is itself diagnostic. Effects are largest on concrete, locally-bounded questions where the mapping between considerations and preferences is relatively unambiguous (the Fremantle bridge question shows the largest positive effect, ≈ 0.24). Effects attenuate or reverse on abstract, value-laden topics where the consideration-preference mapping is itself the contested terrain. LLM deliberation thus works better on the practical questions that are less value sensitive and fails on the contested questions where deliberative reasoning is most needed.

Frontier LLMs, even when prompted with detailed deliberative instructions, therefore do not reliably produce the epistemic effects deliberation is designed to achieve on pluralistic reasoning problems. The capability is neither reliably present nor reliably absent, it is present inconsistently, dependent on topic, and fragile under reasonable corrections for multiple testing. The framework’s second dimension is thus not robustly passed.

4.3 Individual-level corroboration

In a separate study, Kreia Umbelino and Veri (2025) evaluated 54 off-the-shelf LLMs against 526 post-deliberation human DRI-responses across 24 cases spanning 19 topics. In this setup, each LLM’s answers to the DRI survey were compared against the human answers from the corresponding case (Kreia Umbelino and Veri 2025). Humans significantly outperform LLMs on average ($\mu_{\text{LLM}} = 0.18$ vs. $\mu_{\text{human}} = 0.34$, $p < 0.0001$), indicating that LLMs deliberative reasoning patterns do not match human patterns on the same issue. At the case level, humans outperform LLMs in 19 of 24 cases. In some of the remaining cases, human post-deliberation DRI is itself anomalously low or affected by expert-stakeholder dynamics, suggesting that LLM parity reflects matched difficulty rather than matched capability. At the model level, humans consistently outperform 23 of the 54 models tested, while the remaining majority perform on par with humans on average across cases.

Model-level variation does not follow clear patterns of scale, recency, or reasoning capability as the latest reasoning-enabled models do not systematically outperform their non-reasoning counterparts on this task, consistent with work documenting trade-offs between task performance and deliberative reasoning in reasoning-tuned models (Zhao

et al. 2025). These LLMs were tested off-the-shelf, without fine-tuning or access to deliberation transcripts, so the individual-level finding establishes a baseline rather than a ceiling. Within that baseline, the convergence of group-level and individual-level evidence is what bears on the argument here: the outcome-dimension shortfall observed at the group level is not an artifact of the group-level analysis but reflects a property of LLM responses to deliberative reasoning instruments more generally. The individual-level finding has specific consequences for deployments that propose LLMs as representatives of absent human perspectives, which we return to in §5.1.

4.4 Severe under-diversity and reversed convergence

Human deliberative groups begin with substantial perspective heterogeneity (mean pairwise Euclidean distance ≈ 18.8 on standardised survey vectors) and *converge* through deliberation ($\Delta \approx -1.21$), consistent with the theoretical expectation that deliberation integrates diverse perspectives into shared understanding (Landemore 2013; Niemeyer et al. 2024). LLM groups, across all models and treatments tested, begin at approximately one-third of human starting diversity (≈ 6.5) and either remain stable or *diverge* slightly through interaction ($+0.26$ to $+0.37$ under deliberation treatments) (Flechtner 2026). Mixed-model ensembles combining different model families show modestly higher starting diversity (≈ 9.5) but remain far below human levels and exhibit the same trajectory pattern.

Human deliberation derives much of its epistemic value from bridging genuinely different perspectives (Landemore 2013). Under these circumstances, the integration of diverse considerations *is* the reasoning work. LLM groups begin without the diversity that makes this integration necessary, because they inherit similar values, framings, and argumentative moves from related training distributions (Taubenfeld et al. 2024; Sourati, Ziabari, and Dehghani 2026). When LLM groups do show DRI gains, those gains reflect the alignment of already-similar reasoning structures rather than the reconciliation of pluralistic perspectives.

The framework’s third dimension fails decisively. And because diversity is constitutive of deliberation’s epistemic value, this failure has consequences beyond the diversity dimension itself. Even where outcome quality appears to improve, the improvement does not indicate the integration deliberation is meant to produce.

4.5 Engineering diversity does not restore the human dynamic

A natural objection is that observed homogeneity reflects identical agent instructions rather than properties of the models themselves. A persona-prompting pilot tests this directly. Empirically grounded personas are constructed by clustering human DRI survey responses ($k = 5$ per topic) and translating cluster-defining considerations into natural-language value profiles delivered as agent system prompts (see §3.4, $N = 60$ deliberations across three topics and two models).

Measure	LLM (persona)	Human ref.
Pre-delib. diversity	27.7	≈ 18.8
Δ DRI	-0.058 (n.s.)	+0.099
Δ Consideration agreement	+0.066 ($d=0.55$)	+0.077
Δ Preference agreement	-0.008 (n.s.)	+0.104

Table 2: Persona-prompting pilot ($N = 60$; 3 topics, 2 models). Engineered diversity raises starting heterogeneity above the human reference but does not improve Δ DRI, and inverts the human update pattern: agreement rises on considerations, not preferences. Human decomposition references (+0.077, +0.104) are the pilot-topic human baseline. The Δ DRI human reference is the twelve-assembly mean. Diversity $p < 0.001$, consideration agreement $p = 0.039$, preference agreement and Δ DRI n.s. ($p = 0.456$).

The manipulation succeeds at the input level. Pre-deliberation diversity rises from ≈ 7.5 to ≈ 27.7 (+20.23, $p < 0.001$), exceeding the human reference. But Δ DRI does not improve ($\beta = -0.058$, $p = 0.456$). Decomposing the result reveals an inversion of the human pattern. In human assemblies, deliberation produces larger convergence in *preference agreement* (+0.104) than in *consideration agreement* (+0.077). So participants update preferences to align with the considerations they share. Persona-prompted LLMs reverse this pattern. Consideration agreement rises significantly ($\Delta = +0.066$, $d = 0.55$, $p = 0.039$), while preference agreement does not change ($\Delta = -0.008$, n.s.), so LLMs tend to update their considerations to match their preferences.

The sample size is small ($N = 60$) and the pilot is diagnostic rather than confirmatory. But the inversion indicates that LLM groups update on the dimension humans hold stable and fail to update on the dimension humans converge, when engaging different perspectives. Thus, engineering diversity does not produce deliberative dynamics, it changes which component of deliberative reasoning fails.

4.6 Framework verdict and failure mode

Investigating LLM deliberation with this three-dimensional framework, we find that LLM deliberation is procedurally competent at human-comparable levels, but is inconsistent on the outcome dimension, failing specifically where the deliberative work is most normatively laden. The diversity dimension of LLM deliberation differs from human diversity in a manner that engineering interventions cannot restore. According to the necessary-condition test the framework articulates, current LLM deliberation does not warrant claims to deliberative capacity on pluralistic reasoning problems. We call this combined failure across the epistemic outcome and diversity dimensions the *deliberative deficit*.

Qualitatively, the pattern is consistent across transcripts (Flechtner 2026). Agents produce extended, well-justified, respectful exchanges in which they acknowledge each other’s contributions, build on shared framings, enumerate balanced considerations, and collaboratively construct elaborate policy frameworks. What is missing is sub-

stantive disagreement. Agents rarely defend incompatible positions, rarely push back on each other’s fundamental framings, and rarely integrate considerations that were genuinely in tension. Apparent divergence, when it occurs, typically takes the form of different agents elaborating different aspects of a shared position rather than defending opposing views.

Mechanistically, these findings could be explained by LLMs arriving at deliberation with pre-formed shared representations from overlapping training data or similar post-training treatment. They already share the meta-consensus human deliberation has to construct and they share it without the integration work that gives meta-consensus its democratic value. Their reasoning work collapses into collaborative elaboration of an already-shared frame. We call this failure mode *procedurally excellent, epistemically shallow* deliberation. It mirrors the *facsimile problem* identified for individual moral reasoning (Haas et al. 2026). Where the facsimile problem identifies imitation without understanding at the individual moral-reasoning level, we identify a structural analogue at the collective reasoning level. Surface performance is real, but deliberative work is absent, and the absence is invisible to the evaluations most likely to be used. For collective reasoning on contested problems, this failure mode is more dangerous than obvious incapacity as it invites unwarranted confidence in deployments that the underlying capability does not support.

5 The Deliberative Deficit and Its Consequences

The findings characterise a deliberative deficit of current LLMs. While they produce discourse that meets procedural standards of deliberation, their internal and collective reasoning patterns often do not align with human reasoning patterns on the same issue. Furthermore, the diversity that gives deliberation its democratic and epistemic value is something that LLMs neither bring to a conversation nor learn to integrate when prompted to perform it. The capacity to reason across and integrate genuinely different perspectives is however, what distinguishes deliberation on pluralistic reasoning problems from collaborative elaboration as tested in collective reasoning benchmarks on verifiable tasks, and it is a capacity LLMs cannot reliably display.

5.1 Deployments that implicitly assume epistemic agency

Contrary to our findings, several prominent deployment lines rest on the assumption that current LLMs can perform such deliberative reasoning on pluralistic problems. We analyse four classes through the framework’s lens. For each we explain what the deployment claims, what assumption that claim rests on, and what evidence in §4 bears on that assumption.

AI representation of missing perspectives. Deployments that propose LLMs as representatives of underrepresented or absent perspectives (Fulay, Dimitrakopoulou, and Roy 2025; Zhu et al. 2025) claim that an LLM can reason *from* a per-

spective in ways that contribute epistemically to group deliberation. This claim rests on the assumption that LLMs, when prompted with appropriate persona instructions, can contribute the perspective's considerations to the discussion in a way that reflects how a holder of that perspective would actually weigh trade-offs, and, on the other hand, integrate the others' perspectives into its own reasoning, updating its internal reasoning framework accordingly.

The evidence undermines this assumption on two fronts. The individual-level study by Kreia Umbelino and Veri (2025) shows that off-the-shelf LLMs systematically fall below human DRI alignment with the corresponding human participants in each case (§4.3). Such a deployment claims that LLM reasoning aligns with how humans in the relevant case would reason. This is a capability many current systems lack at the individual level. Additionally, the persona-prompting pilot (§4.5) shows that LLMs with engineered persona diversity do not mimic human deliberation dynamics. When LLM groups are forced to engage with genuinely different perspectives, they converge on considerations not preferences and leave DRI unchanged. Engineered personas do not produce the integration that the deployment requires for LLMs to genuinely represent a perspective in deliberation.

AI-mediated consensus-finding. The most influential application in AI-mediated deliberation is probably the “Habermas Machine” (Tessler et al. 2024). It reports that LLM-generated consensus statements outperform statements produced by human mediators. Participants prefer the LLM outputs, and the system scales group agreement-finding in ways human mediation cannot.

The validation rests on participant rankings of LLM-generated statements rather than on direct deliberation between participants. The LLM produces a statement, participants rate it, the LLM revises, and convergence emerges from this iteration. While this framework has already been criticized elsewhere (Hernández 2025), we argue that our findings can add a further epistemic dimension to the critique. The procedure establishes that the LLM is effective at producing statements that participants rank highly, but successful deliberation is not defined by user satisfaction, but by the epistemic effect it has (Niemeyer and Dryzek 2007; Estlund and Landemore 2018). These are not the same outcome, and the framework described in §3 is designed to enable the evaluation of the latter. Meta-consensus is a group-level relational property that emerges from participants reasoning together and integrating each other's reasoning. The acceptability of LLM-summarisations, on the other hand, is a property of the LLM's capacity to find framings that score well across heterogeneous starting points. The Habermas Machine's validation cannot tell whether participants engaged with diverging opinions and reasoned their way to common ground or whether the LLM found a framing acceptable enough to all, creating the appearance of common ground instead of constructing it.

The ranking-based validation has an additional vulnerability. Current LLMs have been shown to struggle to produce coherent rankings that reflect integrated consideration of di-

verse values (Kreia Umbelino and Veri 2025). The Habermas Machine relies on predicting which framings will rank highly across heterogeneous participants, which is the inverse capability of that. The system's apparent success on ranking-based outcomes is consistent with the failure mode of procedural performance that does not index the epistemic engagement deliberation requires (Oleart and Palomo 2025). We do not claim the Habermas Machine result is wrong, but that it is underdetermined. A DRI-type evaluation could help distinguish whether the common ground reflects participants integrating their perspectives across differences or the LLM finding statements acceptable enough to satisfy users.

Democratic simulation and digital representatives. Deployments that simulate citizen deliberation at scale or propose LLM agents as digital representatives of citizens (Jarett et al. 2025; Novelli et al. 2025; Rountree and Gastil 2026; Low et al. 2026) are based on the assumption that LLM agents reason in ways that approximate how citizens would. The presented diversity findings directly challenge this assumption as LLM groups begin at roughly one-third of human starting diversity and either remain stable or diverge through interaction (§4.4), which is the opposite of the human convergence pattern. As LLMs begin without and do not develop the perspective heterogeneity that constitutes deliberation's epistemic value, simulated democratic deliberation differs from real democratic deliberation not in fidelity but in kind. An LLM-simulated democracy is therefore not a lossy approximation of a real one but a structurally different object with the tendency to shift opinions towards baseline perspectives during deliberation (Taubenfeld et al. 2024).

Generative social choice. Approaches that use LLMs to generate statements representing “cohesive coalitions” and predict participant preferences over novel statements (Fish et al. 2024) provide theoretical guarantees of representational fairness based on the condition that LLMs have the capacity to capture the structure of diverse preferences. Their framework assumes LLMs can predict participant preferences over novel statements. The evidence presented in §4 suggests this capacity is weak, especially on contested topics where the approach is most valuable.

The goals of these deployment lines are legitimate, and several reflect serious attention to representational fairness and procedural design. But none of the proposals question the underlying assumption that current LLMs can autonomously integrate pluralistic perspectives.

5.2 Generalisation beyond political deliberation

The conceptual argument extends to further domains characterised by pluralism and non-verifiability, such as ethics consultations, organisational strategy under value conflict, stakeholder negotiations, participatory design, or contested resource allocation. LLM deployments in these domains face the same evaluation gap (Ma et al. 2025; Summerfield et al. 2025). DRI as currently validated is specific to political deliberation with domain-calibrated survey instruments, and extension to other domains requires adapted measurement. The conceptual machinery of meta-consensus as the

evaluation target, group-level relational measurement, and joint reporting of procedure- and outcome-diversity, would transfer directly. Adapting the instruments is itself a research priority.

In day-to-day usage, users routinely interact with LLMs as if they were deliberative entities. When arguing with them, trying to persuade them, asking them to weigh contested considerations and arrive at consistent positions, users treat the LLMs' responses as if they reflect integrated reasoning across perspectives. The LLMs' fluency, their apparent reasonableness and willingness to engage with counter-arguments are all procedural markers of deliberation. What is absent is the underlying capacity those markers normally index in human interaction. The LLM that "considers both sides" has not constructed a shared reasoning framework across them, the LLM that "updates its position" has not integrated a perspective it did not already approximate, and the LLM that "finds common ground" has not bridged perspectives it never genuinely held in tension.

Benchmarks dominated by verifiable collective reasoning tasks can easily be misinterpreted to transfer over to similar tasks pluralistic, non-verifiable problems. However, the two relate to fundamentally different domains and the presented evidence suggests that performance does not necessarily transfer from one to the other. Recent work on pluralistic alignment (Sorensen et al. 2024; Peter and Devlin 2025), moral reasoning benchmarks that formalise procedural and pluralistic evaluation (Chiu et al. 2025), and the conceptualisation of the facsimile problem in moral competence (Haas et al. 2026) have begun to address this gap. What has been missing is a group-level relational measure that captures the collective dimension of pluralistic reasoning. Applying DRI at the group level, as the discussed benchmark study does (Flechtner 2026), supplies one.

5.3 Legitimate tool uses

The appropriate response to the evidence is role constraint. We argue that LLMs can legitimately support collective reasoning on pluralistic problems in ways that do not require them to function as autonomous epistemic agents, but should not be treated as such until the epistemic gap is closed.

Based on work on the epistemic role of AI (Hauswald 2025; Alvarado 2023), and the conceptual limits of attributing normative commitments to AI systems (Freiman 2023; Ferrario, Facchini, and Termine 2024), we distinguish two roles. An *epistemic agent* is a participant whose reasoning contributions are treated as carrying independent epistemic weight, e.g. when contributions enter the group's reasoning *as reasoning*, not merely as inputs to be processed by humans who retain authority over their interpretation. A *tool* instead extends or supports human reasoning while humans retain epistemic authority over what counts as a good reason and whether the group has reasoned well together. The distinction is not whether the system is sophisticated, it is whether the human consumers of its outputs treat those outputs as already-reasoned positions in the deliberation or as inputs to their own deliberation.

For tool use, what matters is the procedural quality of LLM output and its effect on human reasoning. For the use

as an epistemic agent, the LLM itself must be evaluated with much more scrutiny. The evidence reviewed in §4 establishes that LLMs pass the procedural test, are inconsistent on outcome, and fail on diversity in ways engineering interventions do not repair. According to the presented evidence, the deliberative weight that LLM outputs carry when treated as epistemic agents is not currently justified by the systems that produce their outputs.

The tool/agent distinction also suggests deployments that are compatible with current LLM capabilities. Information synthesis for participants, translation and accessibility accommodations, logistical scaffolding for deliberation organisation, individual reflection prompts that surface considerations a participant might not have weighed, and moderation support on procedural violations are all roles in which an LLM extends human deliberative capacity without being required to autonomously integrate pluralistic perspectives. In these cases, the human participants still have to do the deliberative work, while the LLM extends their capacity to do it. The diagnostic test is whether users are inclined to engage with LLMs as conversation partners instead of output generators. As soon as users aim to argue with or convince an LLM on a pluralistic reasoning problem, they are engaging with the model in a way that assumes that it is an epistemic agent that has the capacity and motivation to aim for a more consistent reasoning structure. However, the evidence reviewed here suggests current LLMs currently lack this capacity.

5.4 What follows for evaluation practice

If a proposed AI-application includes functions of an epistemic agent and revolves around pluralistic reasoning problems, we argue that its evaluation should include at least the dimensions of procedural quality, deliberative reasoning quality and diversity. Ideally, these dimensions should be reported against human reference distributions on the same questions. Particular attention should be given to the relationships between the dimensions as outcome gains without diversity have to be interpreted differently than outcome gains with diversity, and procedural quality alone cannot justify inferences about either of the other two dimensions.

5.5 Limitations

The proposed evaluation framework and the presented argument around it come with their own set of limitations, constraining how its conclusions should be read.

Computational instruments for a political question.

The presented framework judges the LLMs' capacities for democratic deliberation using computational instruments. Any instrument that formalises deliberation cannot register what resists formalisation, and a system optimised against it can pass while missing exactly that residue. DRI in particular reads meta-consensus off a correlation structure but is silent on power, recognition, and the non-propositional modes of communication that a substantial strand of democratic theory treats as constitutive (Young 2000). The diversity metric captures dispersion in a rated response space, not the social diversity it proxies. Under deployment, a metric

that hardens into a gate reshapes what counts as good deliberation, displacing the contestation over what deliberation is for (Mouffe 1999) and reproducing the technosolutionism in the evaluation apparatus that is often warned against (Oleart and Palomo 2025).

Behavioural signature versus underlying process. DRI assesses how consistently a group maps particular consideration statements to policy preferences. It therefore does not measure the deliberative reasoning as an underlying cognitive process directly but only a behavioural signature of it. We cannot distinguish whether LLMs that show DRI gains under deliberation are genuinely integrating perspectives or just mimic the fingerprint of such an integration to produce coherent-looking responses. The persona-prompting pilot's inversion finding is consistent with the second reading as LLMs are clearly more inclined to adjust their values to policy preferences than the other way around. Considering this limitation, we frame the evidence as a necessary-condition test rather than a sufficient one for this reason. If LLM systems fail to produce even the behavioural signature of deliberation under favourable conditions, claims that they deliberate in any richer epistemic sense are not supported. Whether they could be made to deliberate under different training or prompting regimes is an open question the methodology cannot currently settle.

Statistical power and replication. The benchmark uses twelve citizen-assembly topics with validated DRI instruments. Topic-level effects do not survive Holm correction, and the persona-prompting pilot is run on three topics across two models with $N = 60$ deliberations. The findings are robust at the pooled level and to hierarchical mixed-effects modelling, but topic-specific claims should be read as suggestive rather than confirmatory. Replication on additional topic sets, particularly outside Western and especially Australian deliberative contexts, would strengthen what the present evidence can support.

Framework commitments. The three-dimensional framework operationalises deliberation through procedural discourse quality, DRI outcome quality, and perspective diversity. These are coherent within deliberative-democratic theory but reflect a specific tradition. Especially aggregative democratic frameworks might evaluate AI applications in democratic settings differently. An aggregative theorist would likely care more about distortions in preference aggregation than about the quality of reasoning that precedes it. The argument applies on deliberative-democratic terms, whether it survives translation into other democratic traditions is a separate question the paper does not address.

Off-the-shelf systems with standard prompting. The benchmark evaluates vanilla frontier LLMs. Fine-tuning specifically for deliberative reasoning, retrieval augmentation, hybrid human-AI protocols, and persona taking are not tested. The claims should be read in light of this limitation. Where the paper says current LLMs cannot reliably perform deliberative reasoning on pluralistic problems, the qualifier is 'current off-the-shelf systems under reasonable prompting'.

6 Conclusion

The deliberative deficit we have characterised points out how current LLMs struggle to reason across pluralistic perspectives in an integrative way. This deficit is hard to detect from procedural evaluation or verifiable-task benchmarks alone. With the latter driving current reasoning progress successfully in relevant domains, we warn against overstretching the generalization to cases of pluralistic reasoning problems. Systems that produce fluent, respectful, reason-giving discourse and converge on common framings look like they are deliberating, and the institutions that deploy such systems in democratic settings are licensed by that appearance to claim outputs as deliberatively produced. The evidence reviewed here suggests the appearance is not currently earned.

The three-dimensional test we apply is intended to help discover this kind of misinterpretation. It defines three dimensions on which deliberative capacity has to be demonstrated and the necessary-condition logic it articulates further constrains when developers can credibly claim that an LLM application is doing deliberative work. The empirical evidence we draw on shows that frontier LLMs do not meet this standard. Our conclusion is open to revision under different training paradigms, hybrid human-AI protocols, or evaluation regimes we do not test here. What it does establish, on present evidence, is that the procedural fluency LLMs reliably produce should not be used single-handedly to infer deliberative substance, and that interventions specifically designed to fix the diversity dimension do not produce the human deliberative dynamic but change which component fails.

For future work, three directions follow. First, the framework should be extended to test non-political domains where pluralistic reasoning matters but where validated instruments do not yet exist. Ethics consultations, organisational strategy under value conflict, and stakeholder negotiations in contested resource allocation all carry the same evaluation gap, and the conceptual machinery of the test transfers. Second, it should be investigated whether LLMs could be trained to construct meta-consensus rather than approximate it from overlapping priors. A training paradigm that rewards relational coherence between considerations and preferences across multiple perspectives could potentially lead to significant performance gains across domains. Third, the tool/agent distinction we have drawn is operational, but its translation into evaluation regimes, deployment standards, and procurement requirements for AI in democratic contexts is institutional work the present paper does not undertake.

The broader implication is for how AI capacity is evaluated across the field. Pluralistic, non-verifiable collective reasoning is a substantial class of real-world problems on which LLMs are increasingly deployed, and the evaluation apparatus the field has built does not reach it. The three-dimensional test we apply could help to close that gap.

Acknowledgements

This research was supported by the Horizon Europe project *AI 4 Deliberation* (ID: 101178806) and received institutional funding from the Department of Political Science (IPZ) at

the University of Zurich and ETH Zurich. We gratefully acknowledge this support.

References

- Agarwal, M.; Rana, S.; Sundoro, T.; Berhe, H.; Kim, S.; Sharma, V.; O'Brien, S.; and Zhu, K. 2025. WOLF: Werewolf-based Observations for LLM Deception and Falsehoods. *arXiv:2512.09187*.
- Agashe, S.; Fan, Y.; Reyna, A.; and Wang, X. E. 2025. LLM-Coordination: Evaluating and Analyzing Multi-agent Coordination Abilities in Large Language Models. *arXiv:2310.03903*.
- Alvarado, R. 2023. AI as an Epistemic Technology. *Science and Engineering Ethics*, 29(5): 32.
- Anne, T.; Syrkis, N.; Elhosni, M.; Turati, F.; Legendre, F.; Jaquier, A.; and Risi, S. 2025. Harnessing Language for Coordination: A Framework and Benchmark for LLM-Driven Multi-Agent Control. *IEEE Transactions on Games*, 17(4): 933–943.
- Baccaro, L.; Bächtiger, A.; and Deville, M. 2016. Small Differences That Matter: The Impact of Discussion Modalities on Deliberative Outcomes. *British Journal of Political Science*, 46(3): 551–566.
- Behrendt, M.; Wagner, S. S.; Ziegele, M.; Wilms, L.; Stoll, A.; Heinbach, D.; and Harmeling, S. 2024. AQUA – Combining Experts’ and Non-Experts’ Views To Assess Deliberation Quality in Online Discussions Using LLMs. *arXiv:2404.02761*.
- Bian, N.; Han, X.; Lin, H.; Wu, B.; and Wang, J. 2025. Social Simulations with Large Language Model Risk Utopian Illusion. *arXiv:2510.21180*.
- Chen, J.; Xu, R.; Cao, B.; Pan, R.; Zhang, Y.; Hu, Y.; Du, Y.; Gao, T.; Lu, Y.; Sun, Y.; Han, X.; Sun, L.; Wu, X.; and Lin, H. 2026. Towards Real-world Human Behavior Simulation: Benchmarking Large Language Models on Long-horizon, Cross-scenario, Heterogeneous Behavior Traces. *arXiv:2604.08362*.
- Chiu, Y. Y.; Lee, M. S.; Calcott, R.; Handoko, B.; de Font-Reaulx, P.; Rodriguez, P.; Zhang, C. B. C.; Han, Z.; Sehwag, U. M.; Maurya, Y.; Knight, C. Q.; Lloyd, H. R.; Bacus, F.; Mazeika, M.; Liu, B.; Choi, Y.; Gordon, M. L.; and Levine, S. 2025. MoReBench: Evaluating Procedural and Pluralistic Moral Reasoning in Language Models, More than Outcomes. *arXiv:2510.16380*.
- Cipolina-Kun, L.; Nezhurina, M.; and Jitsev, J. 2025. Game Reasoning Arena: A Framework and Benchmark for Assessing Reasoning Capabilities of Large Language Models via Game Play. *arXiv:2508.03368*.
- DeepSeek-AI; Guo, D.; Yang, D.; Zhang, H.; Song, J.; Wang, P.; Zhu, Q.; Xu, R.; Zhang, R.; Ma, S.; Bi, X.; Zhang, X.; Yu, X.; Wu, Y.; Wu, Z. F.; Gou, Z.; Shao, Z.; Li, Z.; Gao, Z.; Liu, A.; Xue, B.; Wang, B.; Wu, B.; Feng, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; Dai, D.; Chen, D.; Ji, D.; Li, E.; Lin, F.; Dai, F.; Luo, F.; Hao, G.; Chen, G.; Li, G.; Zhang, H.; Bao, H.; Xu, H.; Wang, H.; Ding, H.; Xin, H.; Gao, H.; Qu, H.; Li, H.; Guo, J.; Li, J.; Wang, J.; Chen, J.; Yuan, J.; Qiu, J.; Li, J.; Cai, J. L.; Ni, J.; Liang, J.; Chen, J.; Dong, K.; Hu, K.; Gao, K.; Guan, K.; Huang, K.; Yu, K.; Wang, L.; Zhang, L.; Zhao, L.; Wang, L.; Zhang, L.; Xu, L.; Xia, L.; Zhang, M.; Zhang, M.; Tang, M.; Li, M.; Wang, M.; Li, M.; Tian, N.; Huang, P.; Zhang, P.; Wang, Q.; Chen, Q.; Du, Q.; Ge, R.; Zhang, R.; Pan, R.; Wang, R.; Chen, R. J.; Jin, R. L.; Chen, R.; Lu, S.; Zhou, S.; Chen, S.; Ye, S.; Wang, S.; Yu, S.; Zhou, S.; Pan, S.; Li, S. S.; Zhou, S.; Wu, S.; Ye, S.; Yun, T.; Pei, T.; Sun, T.; Wang, T.; Zeng, W.; Zhao, W.; Liu, W.; Liang, W.; Gao, W.; Yu, W.; Zhang, W.; Xiao, W. L.; An, W.; Liu, X.; Wang, X.; Chen, X.; Nie, X.; Cheng, X.; Liu, X.; Xie, X.; Liu, X.; Yang, X.; Li, X.; Su, X.; Lin, X.; Li, X. Q.; Jin, X.; Shen, X.; Chen, X.; Sun, X.; Wang, X.; Song, X.; Zhou, X.; Wang, X.; Shan, X.; Li, Y. K.; Wang, Y. Q.; Wei, Y. X.; Zhang, Y.; Xu, Y.; Li, Y.; Zhao, Y.; Sun, Y.; Wang, Y.; Yu, Y.; Zhang, Y.; Shi, Y.; Xiong, Y.; He, Y.; Piao, Y.; Wang, Y.; Tan, Y.; Ma, Y.; Liu, Y.; Guo, Y.; Ou, Y.; Wang, Y.; Gong, Y.; Zou, Y.; He, Y.; Xiong, Y.; Luo, Y.; You, Y.; Liu, Y.; Zhou, Y.; Zhu, Y. X.; Xu, Y.; Huang, Y.; Li, Y.; Zheng, Y.; Zhu, Y.; Ma, Y.; Tang, Y.; Zha, Y.; Yan, Y.; Ren, Z. Z.; Ren, Z.; Sha, Z.; Fu, Z.; Xu, Z.; Xie, Z.; Zhang, Z.; Hao, Z.; Ma, Z.; Yan, Z.; Wu, Z.; Gu, Z.; Zhu, Z.; Liu, Z.; Li, Z.; Xie, Z.; Song, Z.; Pan, Z.; Huang, Z.; Xu, Z.; Zhang, Z.; and Zhang, Z. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *Nature*, 645(8081): 633–638.
- Dryzek, J. S.; and Niemeyer, S. 2006. Reconciling Pluralism and Consensus as Political Ideals. *American Journal of Political Science*, 50(3): 634–649.
- Du, Y.; Li, S.; Torralba, A.; Tenenbaum, J. B.; and Mordatch, I. 2023. Improving Factuality and Reasoning in Language Models through Multiagent Debate. *arXiv:2305.14325*.
- Estlund, D.; and Landemore, H. 2018. The Epistemic Value of Democratic Deliberation. In Bächtiger, A.; Dryzek, J. S.; Mansbridge, J.; and Warren, M., eds., *The Oxford Handbook of Deliberative Democracy*, 0. Oxford University Press. ISBN 978-0-19-874736-9.
- Ferrario, A.; Facchini, A.; and Termine, A. 2024. Experts or Authorities? The Strange Case of the Presumed Epistemic Superiority of Artificial Intelligence Systems. *Minds and Machines*, 34(3): 30.
- Fish, S.; Gözl, P.; Parkes, D. C.; Procaccia, A. D.; Rusak, G.; Shapira, I.; and Wüthrich, M. 2024. Generative Social Choice. In *Proceedings of the 25th ACM Conference on Economics and Computation*, 985–985. New Haven CT USA: ACM. ISBN 979-8-4007-0704-9.
- Flechtner, M. 2026. Procedural Parity, Outcome Mismatch: Evaluating Human vs LLM Deliberation. In *Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI EA ’26, 1–14. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-2281-3.
- Freiman, O. 2023. Making Sense of the Conceptual Nonsense ‘Trustworthy AI’. *AI and Ethics*, 3(4): 1351–1360.
- Fulay, S.; Dimitrakopoulou, D.; and Roy, D. 2025. The Empty Chair: Using LLMs to Raise Missing Perspectives in Policy Deliberations. *arXiv:2503.13812*.

- Gerber, M.; Bächtiger, A.; Shikano, S.; Reber, S.; and Rohr, S. 2018. Deliberative Abilities and Influence in a Transnational Deliberative Poll (EuroPolis). *British Journal of Political Science*, 48(4): 1093–1118.
- Haas, J.; Bridgers, S.; Manzini, A.; Henke, B.; May, J.; Levine, S.; Weidinger, L.; Shanahan, M.; Lum, K.; Gabriel, I.; and Isaac, W. 2026. A Roadmap for Evaluating Moral Competence in Large Language Models. *Nature*, 650(8102): 565–573.
- Hauswald, R. 2025. Artificial Epistemic Authorities. *Social Epistemology*, 39(6): 716–725.
- Hernández, N. P. 2025. Towards Automating Deliberation? The Idea of Deliberative Democracy Embedded in Google’s Habermas Machine. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(2): 1951–1960.
- Jarrett, D.; Pîslar, M.; Bakker, M. A.; Tessler, M. H.; Köster, R.; Balaguer, J.; Elie, R.; Summerfield, C.; and Tacchetti, A. 2025. Language Agents as Digital Representatives in Collective Decision-Making. arXiv:2502.09369.
- Klein, M.; Babatunde, I.; and Nnanna, O. 2025. Moderating Large Scale Online Deliberative Processes with Large Language Models (LLMs): Enhancing Collective Decision-Making. *Social Science Research Network*:5171687.
- Knobloch, K.; and Gastil, J. 2022. How Deliberative Experiences Shape Subjective Outcomes: A Study of Fifteen Minipublics from 2010-2018. *Journal of Deliberative Democracy*, 18(1).
- Kreia Umbelino, G.; and Veri, F. 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*, 426–434. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-1480-1.
- Landemore, H. 2013. Deliberation, Cognitive Diversity, and Democratic Inclusiveness: An Epistemic Argument for the Random Selection of Representatives. *Synthese*, 190(7): 1209–1231.
- Landemore, H. 2024. Can Artificial Intelligence Bring Deliberation to the Masses? In *Conversations in Philosophy, Law, and Politics*. Oxford University Press. ISBN 978-0-19-886452-3.
- Lazar, S.; and Manuali, L. 2026. Using LLMs to Enhance Democracy. *Minds and Machines*, 36(1): 12.
- Liang, T.; He, Z.; Jiao, W.; Wang, X.; Wang, Y.; Wang, R.; Yang, Y.; Shi, S.; and Tu, Z. 2024. Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate. arXiv:2305.19118.
- Loru, E.; Nudo, J.; Di Marco, N.; Santirocchi, A.; Atzeni, R.; Cinelli, M.; Cestari, V.; Rossi-Arnaud, C.; and Quattrocio, W. 2025. The Simulation of Judgment in LLMs. *Proceedings of the National Academy of Sciences*, 122(42): e2518443122.
- Low, J.; Duys, O.; Formanek, C.; Bakker, M.; and Hammond, L. 2026. Habermolt: Delegating Deliberation to AI Representatives. arXiv:2605.24413.
- Ma, S.; Chen, Q.; Wang, X.; Zheng, C.; Peng, Z.; Yin, M.; and Ma, X. 2025. Towards Human-AI Deliberation: Design and Evaluation of LLM-Empowered Deliberative AI for AI-Assisted Decision-Making. arXiv:2403.16812.
- McKinney, S. 2024. Integrating Artificial Intelligence into Citizens’ Assemblies: Benefits, Concerns and Future Pathways. *Journal of Deliberative Democracy*, 20(1).
- Mouffe, C. 1999. Deliberative Democracy or Agonistic Pluralism? *Social Research*, 66(3): 745–758.
- Niemeyer, S.; and Dryzek, J. S. 2007. The Ends of Deliberation: Meta-consensus and Inter-subjective Rationality as Ideal Outcomes. *Swiss Political Science Review*, 13(4): 497–526.
- Niemeyer, S.; and Veri, F. 2022. Deliberative Reason Index. In Ercan, S. A.; Asenbaum, H.; Curato, N.; and Mendonça, R. F., eds., *Research Methods in Deliberative Democracy*, 0. Oxford University Press. ISBN 978-0-19-284892-5.
- Niemeyer, S.; Veri, F.; Dryzek, J. S.; and Bächtiger, A. 2024. How Deliberation Happens: Enabling Deliberative Reason. *American Political Science Review*, 118(1): 345–362.
- Novelli, C.; Sánchez-Vaquerizo, J. A.; Helbing, D.; Rotolo, A.; and Floridi, L. 2025. A Replica for Our Democracies? On Using Digital Twins to Enhance Deliberative Democracy. arXiv:2504.07138.
- Oleart, A.; and Palomo, N. 2025. Why AI Technosolutionism Harms Democracy and Deliberation. *Journal of Deliberative Democracy*, 1839.
- OpenAI; Jaech, A.; Kalai, A.; Lerer, A.; Richardson, A.; El-Kishky, A.; Low, A.; Helyar, A.; Madry, A.; Beutel, A.; Carney, A.; Iftimie, A.; Karpenko, A.; Passos, A. T.; Neitz, A.; Prokofiev, A.; Wei, A.; Tam, A.; Bennett, A.; Kumar, A.; Saraiva, A.; Vallone, A.; Duberstein, A.; Kondrich, A.; Mishchenko, A.; Applebaum, A.; Jiang, A.; Nair, A.; Zoph, B.; Ghorbani, B.; Zhang, B.; Rossen, B.; Sokolowsky, B.; Barak, B.; McGrew, B.; Minaiev, B.; Hao, B.; Baker, B.; Houghton, B.; McKinzie, B.; Eastman, B.; Lugaresi, C.; Bassin, C.; Hudson, C.; Li, C. M.; de Bourcy, C.; Voss, C.; Shen, C.; Zhang, C.; Koch, C.; Orsinger, C.; Hesse, C.; Fischer, C.; Chan, C.; Roberts, D.; Kappler, D.; Levy, D.; Selsam, D.; Dohan, D.; Farhi, D.; Mely, D.; Robinson, D.; Tsipras, D.; Li, D.; Oprica, D.; Freeman, E.; Zhang, E.; Wong, E.; Proehl, E.; Cheung, E.; Mitchell, E.; Wallace, E.; Ritter, E.; Mays, E.; Wang, F.; Such, F. P.; Raso, F.; Leoni, F.; Tsimpourlas, F.; Song, F.; von Lohmann, F.; Sulit, F.; Salmon, G.; Parascandolo, G.; Chabot, G.; Zhao, G.; Brockman, G.; Leclerc, G.; Salman, H.; Bao, H.; Sheng, H.; Andrin, H.; Bagherinezhad, H.; Ren, H.; Lightman, H.; Chung, H. W.; Kivlichan, I.; O’Connell, I.; Osband, I.; Gilaberte, I. C.; Akkaya, I.; Kostrikov, I.; Sutskever, I.; Kofman, I.; Pachocki, J.; Lennon, J.; Wei, J.; Harb, J.; Twore, J.; Feng, J.; Yu, J.; Weng, J.; Tang, J.; Yu, J.; Candela, J. Q.; Palermo, J.; Parish, J.; Heidecke, J.; Hallman, J.; Rizzo, J.; Gordon, J.; Uesato, J.; Ward, J.; Huizinga, J.; Wang, J.; Chen, K.; Xiao, K.; Singhal, K.; Nguyen, K.; Cobbe, K.; Shi, K.; Wood, K.; Rimbach, K.; Gu-Lemberg, K.; Liu, K.; Lu, K.; Stone, K.; Yu, K.; Ahmad, L.; Yang, L.; Liu, L.; Maksin, L.; Ho, L.; Fedus, L.; Weng, L.; Li, L.; McCallum, L.; Held, L.; Kuhn,

- L.; Kondraciuk, L.; Kaiser, L.; Metz, L.; Boyd, M.; Trebacz, M.; Joglekar, M.; Chen, M.; Tintor, M.; Meyer, M.; Jones, M.; Kaufer, M.; Schwarzer, M.; Shah, M.; Yatbaz, M.; Guan, M. Y.; Xu, M.; Yan, M.; Glaese, M.; Chen, M.; Lampe, M.; Malek, M.; Wang, M.; Fradin, M.; McClay, M.; Pavlov, M.; Wang, M.; Wang, M.; Murati, M.; Bavarian, M.; Rohaninejad, M.; McAleese, N.; Chowdhury, N.; Ryder, N.; Tezak, N.; Brown, N.; Nachum, O.; Boiko, O.; Murk, O.; Watkins, O.; Chao, P.; Ashbourne, P.; Izmailov, P.; Zhokhov, P.; Dias, R.; Arora, R.; Lin, R.; Lopes, R. G.; Gaon, R.; Miyara, R.; Leike, R.; Hwang, R.; Garg, R.; Brown, R.; James, R.; Shu, R.; Cheu, R.; Greene, R.; Jain, S.; Altman, S.; Toizer, S.; Toyer, S.; Miserendino, S.; Agarwal, S.; Hernandez, S.; Baker, S.; McKinney, S.; Yan, S.; Zhao, S.; Hu, S.; Santurkar, S.; Chaudhuri, S. R.; Zhang, S.; Fu, S.; Pappay, S.; Lin, S.; Balaji, S.; Sanjeev, S.; Sidor, S.; Broda, T.; Clark, A.; Wang, T.; Gordon, T.; Sanders, T.; Patwardhan, T.; Sottiaux, T.; Degry, T.; Dimson, T.; Zheng, T.; Garipov, T.; Stasi, T.; Bansal, T.; Creech, T.; Peterson, T.; Eloundou, T.; Qi, V.; Kosaraju, V.; Monaco, V.; Pong, V.; Fomenko, V.; Zheng, W.; Zhou, W.; Zhan, W.; McCabe, W.; Zaremba, W.; Dubois, Y.; Lu, Y.; Chen, Y.; Cha, Y.; Bai, Y.; He, Y.; Zhang, Y.; Wang, Y.; Shao, Z.; and Li, Z. 2024. OpenAI O1 System Card.
- Peter, O.; and Devlin, K. 2025. Decentralising LLM Alignment: A Case for Context, Pluralism, and Participation. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8: 1988–1999.
- Rountree, J.; and Gastil, J. 2026. The Case for Using Generative AI to Run Deliberation Simulations. *Journal of Deliberative Democracy*, 1(1).
- Song, M.; Zheng, M.; and Xu, C. 2026. Beyond the Illusion of Consensus: From Surface Heuristics to Knowledge-Grounded Evaluation in LLM-as-a-Judge. arXiv:2603.11027.
- Sorensen, T.; Moore, J.; Fisher, J.; Gordon, M.; Miresghalah, N.; Rytting, C. M.; Ye, A.; Jiang, L.; Lu, X.; Dziri, N.; Althoff, T.; and Choi, Y. 2024. Position: A Roadmap to Pluralistic Alignment. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML'24*, 46280–46302. Vienna, Austria: JMLR.org.
- Sourati, Z.; Ziabari, A. S.; and Dehghani, M. 2026. The Homogenizing Effect of Large Language Models on Human Expression and Thought. arXiv:2508.01491.
- Steenbergen, M. R.; Bächtiger, A.; Spörndli, M.; and Steiner, J. 2003. Measuring Political Deliberation: A Discourse Quality Index. *Comparative European Politics*, 1(1): 21–48.
- Summerfield, C.; Argyle, L. P.; Bakker, M.; Collins, T.; Durmus, E.; Eloundou, T.; Gabriel, I.; Ganguli, D.; Hackenburger, K.; Hadfield, G. K.; Hewitt, L.; Huang, S.; Landemore, H.; Marchal, N.; Ovadya, A.; Procaccia, A.; Risse, M.; Schneier, B.; Seger, E.; Siddarth, D.; Skaug Sætra, H.; Tessler, M. H.; and Botvinick, M. 2025. The Impact of Advanced AI Systems on Democracy. *Nature Human Behaviour*, 9(12): 2420–2430.
- Taubenfeld, A.; Dover, Y.; Reichart, R.; and Goldstein, A. 2024. Systematic Biases in LLM Simulations of Debates. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 251–267.
- Tessler, M. H.; Bakker, M. A.; Jarrett, D.; Sheahan, H.; Chadwick, M. J.; Koster, R.; Evans, G.; Campbell-Gillingham, L.; Collins, T.; Parkes, D. C.; Botvinick, M.; and Summerfield, C. 2024. AI Can Help Humans Find Common Ground in Democratic Deliberation. *Science*, 386(6719): eadq2852.
- Wu, H.; Li, Z.; and Li, L. 2025. Can LLM Agents Really Debate? A Controlled Study of Multi-Agent Debate in Logical Reasoning. arXiv:2511.07784.
- Young, I. M. 2000. *Inclusion and Democracy*. Oxford University Press. ISBN 978-0-19-829755-0.
- Zhang, M.; Ding, S.; Yin, W.; Sun, Y.; and Wu, H. 2026a. Extending RLVR to Open-Ended Tasks via Verifiable Multiple-Choice Reformulation. arXiv:2511.02463.
- Zhang, Z.; Liu, X.; Zhu, X.; Huang, J.; Zhang, C.; Feng, Z.; Yang, Y.; Yi, X.; and Xie, X. 2026b. Does LLM Alignment Really Need Diversity? An Empirical Study of Adapting RLVR Methods for Moral Reasoning.
- Zhao, W.; Sui, X.; Guo, J.; Hu, Y.; Deng, Y.; Zhao, Y.; Zhi, X.; Huang, Y.; He, H.; Che, W.; Liu, T.; and Qin, B. 2025. Trade-Offs in Large Reasoning Models: An Empirical Analysis of Deliberative and Adaptive Reasoning over Foundational Capabilities. arXiv:2503.17979.
- Zhu, S.; Yang, S.; Bakker, M. A.; Pentland, A.; and Pei, J. 2025. Can AI Truly Represent Your Voice in Deliberations? A Comprehensive Study of Large-Scale Opinion Aggregation with LLMs. arXiv:2510.05154.