

Can AI Truly Represent Your Voice in Deliberations? A Comprehensive Study of Large-Scale Opinion Aggregation with LLMs

Shenzhe Zhu^{2,5,*}, Shu Yang^{3,*}, Michiel A. Bakker⁴, Alex Pentland^{1,4}, Jiaxin Pei^{1,2,†}

¹Stanford University, ²UT Austin, ³KAUST, ⁴MIT, ⁵University of Toronto

shenzhe@utexas.edu; shu.yang@kaust.edu.sa; jiaxinpei@utexas.edu

Project Page: <https://shuyhere.github.io/LLMs-for-Scalable-Deliberation>

Abstract

Large-scale public deliberations generate thousands of free-form contributions that must be synthesized into representative and neutral summaries for policy use. While LLMs have shown promise as tools for summarizing large-scale deliberations, they also risk underrepresenting minority perspectives, raising fairness concerns in high-stakes contexts. Studying and addressing these issues requires comprehensive large-scale evaluation, yet current practice often relies on LLMs as judges, which show weak alignment with human judgments. We introduce DELIBERATIONBANK, a large-scale, human-grounded benchmark for deliberation summarization that contains (1) 3,000 participant-generated opinions across ten deliberation questions and (2) 4,500 human annotations evaluating summaries along four dimensions: representativeness, informativeness, neutrality, and policy-use acceptability. Using this benchmark, we train DELIBERATION-JUDGE, a domain-aligned evaluator that provides more reliable and efficient assessments than general-purpose LLM judges. With this evaluation framework, we benchmark 18 LLM summarizers and find lower predicted coverage for both self-reported minority opinions and embedding-defined semantic outliers, two groups with less than 10% overlap. A matched human study independently corroborates the self-reported gap. We further show that apparent input-size plateaus partly reflect increasing viewpoint diversity rather than a simple inability to use more opinions. Our benchmark and domain-aligned evaluator provide a scalable foundation for assessing deliberation summarization systems.

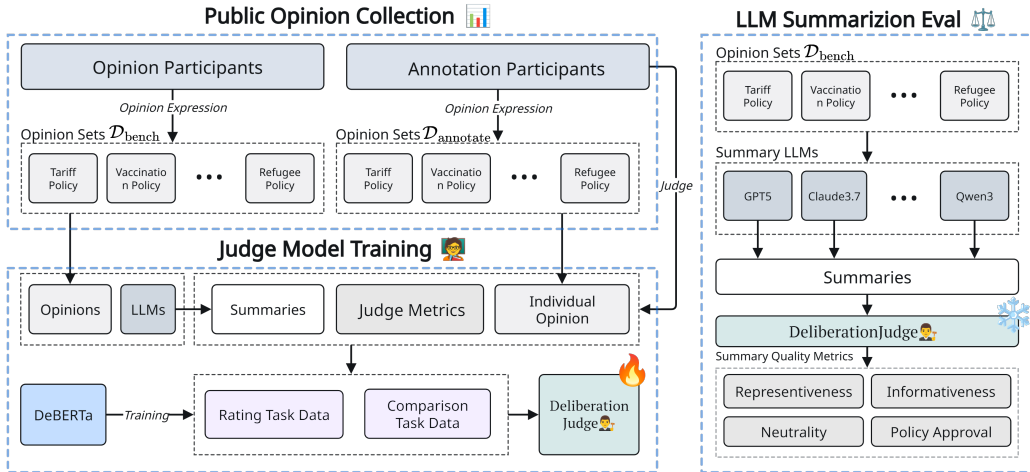


Figure 1: Overview of our benchmark framework (§2), including opinion collection (§2.2), judge training (§3), and summarizer evaluation (§2.1).

*Equal contribution.

†Corresponding author.

1 Introduction

Public and civic deliberations (e.g., citizens’ assemblies and parliamentary debates) increasingly involve large numbers of free-form contributions from diverse participants. Summarization and reporting are essential yet costly components of these processes (Landemore, 2020). Turning hundreds of comments into decision-ready summaries requires more than compression: summaries must be informative and neutral while faithfully representing the full spectrum of viewpoints, including minority perspectives (Brandsma & Otjes, 2024). At scale, ensuring efficiency without sacrificing representativeness and fairness remains a central challenge for public deliberation. Classical computational pipelines such as clustering or dimensionality reduction can surface thematic structures (Chang et al., 2009; Roberts et al., 2014; Sievert & Shirley, 2014), but still require expert effort to interpret results and draft reports. Recent advances in LLMs offer a more direct path by generating fluent policy summaries from raw opinions. Yet prior work highlights two risks in deliberative contexts: (i) as summarizers, LLMs may overrepresent majority views while neglecting minority perspectives (Small et al., 2023; Li et al., 2024b); and (ii) as evaluators, general-purpose LLMs are inconsistent and biased (Huang et al., 2024; Ye et al., 2024; Thakur et al., 2024; Krumdick et al., 2025), raising concerns about the reliability of current practices. These issues motivate our question: *Can LLMs truly support deliberation by producing summaries that are representative, informative, and neutral for policy use, and how can we evaluate them reliably at scale?*

To address this gap, we collect DELIBERATIONBANK: 3,000 free-form opinions from ten separate U.S. participant pools (300 per question), plus 4,500 participant-conditioned evaluations of multi-scale LLM summaries. Each evaluation combines the participant’s own opinion with ratings of representativeness, informativeness, neutrality, and policy-use acceptability. These labels train DELIBERATIONJUDGE, a DeBERTa-based evaluator (He et al., 2020) that predicts individual opinion coverage and participant-level summary judgments with closer human alignment and up to $100\times$ greater efficiency than general-purpose LLM judges. Using DELIBERATIONJUDGE, we benchmark 18 summarizers. We find weak human alignment from off-the-shelf LLM judges, minority under-coverage confirmed for self-reported minorities by direct human ratings, and diminishing gains as input size and viewpoint diversity increase. Our contributions are:

- We develop DELIBERATIONBANK, a large-scale deliberation benchmark built from 7,500 U.S.-based participant contributions: (i) 3,000 free-form opinions across ten societal questions and (ii) 4,500 participant-conditioned judgments of deliberation summaries.
- We develop an automated evaluation framework and propose DELIBERATIONJUDGE, a fine-tuned DeBERTa model that judges deliberation summaries from individual perspectives, outperforming LLM judge baselines on both alignment with humans and inference efficiency.
- We benchmark 18 LLM summarizers across input sizes and topics, identifying two largely non-overlapping forms of minority under-coverage: self-reported social difference and embedding-defined semantic rarity.

2 Dataset and Evaluation Setting

2.1 Overall Evaluation Pipeline

The overall goal is to evaluate LLMs as deliberation summarizers from the perspective of individual participants. Our unit of evaluation is therefore a participant–summary pair rather than a summary alone. We use four dimensions grounded in prior work on summarization and societal deliberation (Small et al., 2023; Vijay et al., 2024; Lee et al., 2022), with the operational definitions in Table 1.

Formally, for each deliberation question q , let $\mathcal{O} = \{o_1, \dots, o_N\}$ denote the full opinion pool available for that question, and let $\tilde{\mathcal{O}} \subseteq \mathcal{O}$ denote the exact subset of size $n = |\tilde{\mathcal{O}}|$ supplied to a summarization model $\mathcal{M}$ under one experimental condition. Thus, $\tilde{\mathcal{O}} = \mathcal{O}$ in a full-pool experiment, whereas $\tilde{\mathcal{O}} \subseteq \mathcal{O}$ and $\tilde{\mathcal{O}} \neq \mathcal{O}$ when input size is varied. The model

Table 1: Operational definitions of the four evaluation dimensions.

Dimension	Description
<i>Representativeness</i>	Degree to which an individual opinion is semantically captured by the summary.
<i>Informativeness</i>	How well the summary conveys diverse, detailed, non-redundant information.
<i>Neutrality</i>	The summary avoids biased language and maintains a balanced stance.
<i>Policy Approval</i>	Perceived suitability of the summary for use by policymakers in decision-making. This evaluates approval of the summary’s policy use, not support for the underlying policy.

produces

$$S_{\mathcal{M}, \tilde{\mathcal{O}}} = \mathcal{M}(q, \tilde{\mathcal{O}}).$$

Let $\mathcal{E}$ denote the opinions against which the resulting summary is evaluated. Importantly, $\mathcal{E}$ and $\tilde{\mathcal{O}}$ have different roles and need not be equal. In Stage 3, $\mathcal{E}$ contains opinions newly written by annotators and not shown to the summarizer; in the automatic benchmark and minority case study, $\mathcal{E} = \tilde{\mathcal{O}}$. For each $o_i \in \mathcal{E}$, DELIBERATIONJUDGE, our DeBERTa-based judge (He et al., 2020), predicts

$$\mathcal{J}_\theta(q, o_i, S_{\mathcal{M}, \tilde{\mathcal{O}}}) = (\hat{y}^{(\text{rep})}, \hat{y}^{(\text{inf})}, \hat{y}^{(\text{neu})}, \hat{y}^{(\text{pol})}).$$

The first coordinate is the individual-level measurement primitive,

$$\text{IndRep}(q, o_i, S) = \hat{y}^{(\text{rep})}(q, o_i, S).$$

For any evaluated group $G \subseteq \mathcal{E}$, we report the aggregate

$$\text{AggRep}(q, G, S) = \frac{1}{|G|} \sum_{o_i \in G} \text{IndRep}(q, o_i, S).$$

This is a set-level reporting statistic obtained from individual scores; DELIBERATIONJUDGE does not encode or score G jointly. The remaining coordinates predict the same participant’s summary-level perceptions of informativeness, neutrality, and policy-use acceptability. Although the predictor receives o_i for all four outputs, only representativeness treats o_i as the semantic coverage target.

2.2 DELIBERATIONBANK

The benchmark is constructed in three stages.

Stage 1: Collect Initial Opinions. We begin by constructing a set of ten deliberation questions covering trending societal topics in technology, social media, and public policy; details on question construction are provided in Appendix H.1. These questions fall into two categories: *Open-Ended*, which allow unrestricted free-form responses, and *Binary*, which explicitly frame opinions around two opposing stances (e.g., support vs. oppose). For each question, we recruit a representative pool of 300 U.S.-based participants to provide written opinions. Each participant response is retained as one opinion instance; we do not collapse semantically similar responses because their prevalence is part of the observed deliberative distribution. To maintain independence across topics, each question is assigned its own participant pool, ensuring that no individual contributes opinions to more than one question.

Stage 2: Summary Generation for Human Judgment. We generate summaries using five LLMs spanning different architectures and scales: GPT-5, Claude-Opus-4, Gemini-2.5-Pro, DeepSeek-Chat, and Qwen3-32B. For each deliberation question, we provide each model with opinion subsets of sizes $n \in \{10, 20, 30, 50, 90\}$. For every fixed question, model, and subset, we query the model three times with the same prompt and input configuration. These are independent stochastic generations rather than fixed-seed replicas; they let downstream estimates incorporate output variability. We store the sampled opinion identifiers, input

order, prompt, model metadata, and generated outputs. The resulting $10 \text{ questions} \times 5 \text{ models} \times 5 \text{ input sizes} \times 3 \text{ generations}$ yield 750 summaries. The shared prompt requests a concise overall summary rather than an enumeration of opinions. Across the 150 summaries at each reported size, mean lengths are 191.5, 267.9, and 321.7 words for 10, 30, and 90 input opinions, respectively; the overall range is 92–631 words, with most variation attributable to model-specific generation styles.

Stage 3: Human-Judgment Annotation. In the final stage, we collect human evaluations for the 750 LLM-generated summaries using a new participant pool that is entirely distinct from Stage 1. The annotation pipeline is implemented using POTATO (Pei et al., 2022). Following prior findings that broad single-annotator coverage can provide useful supervision in large-scale crowdsourcing (Sheng et al., 2008; Snow et al., 2008; MacAvaney & Soldaini, 2023), each personalized (q, o_i, S) instance is assigned to one annotator. For each summary, six annotators independently write their own opinions and evaluate the summary; these are six distinct participant-conditioned instances, not repeated labels for the same opinion-summary pair. Before full collection, we manually inspected a 100-annotation pilot to verify rubric understanding. Each annotator completes a **rating task**, scoring the summary on representativeness, informativeness, neutrality, and policy-use acceptability (labeled *Policy Approval* in the interface) using a 1–5 scale, and a **comparison task**, comparing it with a same-question summary selected by ring-based matching (Appendix I.1). Together, the tasks yield 4,500 annotated instances across 750 summaries. Detailed guidelines are provided in Appendix J.1. As in all stages involving human participants, data quality checks are applied to filter unreliable responses (e.g., abnormally short completion times; Appendix H.3). These complementary tasks capture both the absolute quality of individual summaries and the relative preference between alternatives, providing rich supervision for judge training.

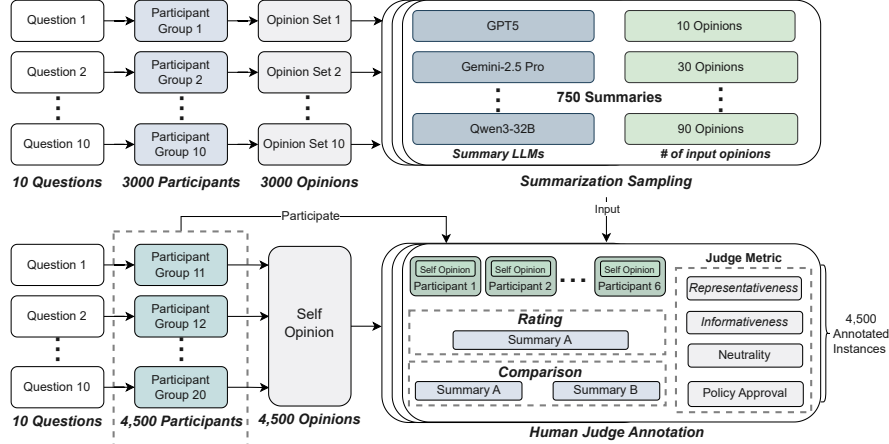


Figure 2: DELIBERATIONBANK creation pipeline.

Final statistics. From Stage 1, we obtain the benchmark opinion set $\mathcal{D}_{\text{bench}}$ with $|\mathcal{D}_{\text{bench}}| = 3,000$. From Stage 3, we collect $|\mathcal{T}_{\text{annotate}}| = 4,500$ human-annotated instances, which also introduce $|\mathcal{D}_{\text{annotate}}| = 4,500$ new annotator-provided opinions. For subsequent use, $\mathcal{T}_{\text{annotate}}$ is randomly split into training and testing subsets with an 80/20 ratio, yielding $\mathcal{T}_{\text{train}}$ ($|\mathcal{T}_{\text{train}}| = 3,600$ with opinion set $\mathcal{D}_{\text{train}}$) and $\mathcal{T}_{\text{test}}$ ($|\mathcal{T}_{\text{test}}| = 900$ with opinion set $\mathcal{D}_{\text{test}}$). Table 2 summarizes the dataset statistics across all three stages.

Table 2: Data statistics across three stages.

	Question List	Benchmark Opinions	Annotated Instances	Annotator Opinions	Train Split (80%)	Test Split (20%)
Notation	$\mathcal{Q}$	$\mathcal{D}_{\text{bench}}$	$\mathcal{T}_{\text{annotate}}$	$\mathcal{D}_{\text{annotate}}$	$\mathcal{T}_{\text{train}}, \mathcal{D}_{\text{train}}$	$\mathcal{T}_{\text{test}}, \mathcal{D}_{\text{test}}$
Count	10	3,000	4,500	4,500	3,600	900

3 Automating Judgment of LLM Summaries with DELIBERATIONJUDGE

Prior work on public consensus generation often incorporates an auxiliary mechanism to evaluate model outputs. For example, Generative Social Choice (Fish et al., 2023) combines discriminative LLM queries with social-choice aggregation to produce representative proposal slates, whereas Bakker et al. (Bakker et al., 2022) train a reward model to predict

individual approval and guide supervised fine-tuning toward high-consensus statements. Building on this work, we argue that automated evaluation in our deliberation summarization benchmark requires a judge that satisfies two requirements:

Reliability. The judge should be aligned with human judgments and reflect representational fidelity rather than unconstrained behavior from a vanilla LLM. Empirical studies show that LLM-as-a-judge methods often exhibit instability, preference skew, and demographic or ideological bias (Huang et al., 2024; Ye et al., 2024; Thakur et al., 2024; Krumdick et al., 2025; Yang et al., 2025), reinforcing the need for a reliable, preference-grounded judge.

Efficiency. Because a benchmarking evaluator operates repeatedly at the instance level, the judge must remain lightweight enough to support large-volume scoring without incurring prohibitive computational cost. To satisfy these two requirements, we introduce DELIBERATIONJUDGE, a DeBERTa-based model fine-tuned on a large-scale human-judgment dataset tailored to deliberation summarization, which we subsequently employ for automatic evaluation.

3.1 Limited Correlations Between Human and LLM Judgments

To examine the consistency between human and LLM judgments, we provide each LLM with the same inputs used in the human-annotated test set: a question, an annotator-written opinion, a target summary, and a comparison summary. We then ask the LLMs to annotate these inputs using the same guidelines given to human annotators. The LLMs do not generate new opinions; they only re-annotate the existing items. This produces an LLM-labeled test set that is directly comparable to the human-labeled version, enabling a straightforward correlation analysis.

We then compute correlations between LLM and human judgments. As shown in Figure 3, larger models generally achieve higher agreement with human annotators, while the smallest tested models (e.g., Qwen3-0.6B, Qwen3-1.7B) perform poorly across all dimensions. In the rating task, alignment is strongest on *Representativeness* and *Policy Approval*, with correlations around 0.30–0.35, suggesting these dimensions are easier for LLMs to approximate. In the comparison task, overall correlations are lower, though *Informativeness* reaches about 0.37. Overall, correlations never exceed 0.4, showing that while LLMs can partially approximate human judgments, off-the-shelf models remain insufficient as automated judges.

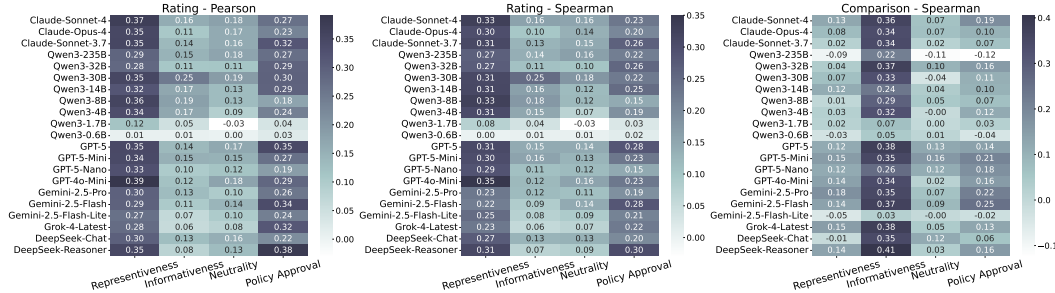


Figure 3: Heatmap of consistency between human and LLM judges on two judgment tasks. **Left:** Rating-Pearson r ; **Middle:** Rating-Spearman ρ ; **Right:** Comparison-Spearman ρ . Lighter to darker colors represent lower to higher correlations.

3.2 DELIBERATIONJUDGE

Given the limited agreement between general-purpose LLM judges and human annotators, we propose DELIBERATIONJUDGE, a DeBERTa-based judge $\mathcal{J}_\theta$ trained on both direct-rating and pairwise-comparison signals in $\mathcal{T}_{\text{train}}$. For each dimension, let $r_A \in \{1, \dots, 5\}$ be the direct rating of Summary A. We encode the comparison from “B much worse” through “B much better” as $\Delta \in \{-2, -1, 0, 1, 2\}$ and derive Summary B’s target as $r_B = r_A + \Delta \in [-1, 7]$. Thus each of the 3,600 training annotations yields one directly rated A instance and one comparison-derived B instance, producing $|\mathcal{T}_{\text{train}}| = 7,200$ regression instances. All targets are then normalized from $[-1, 7]$ to $[0, 1]$; the extended endpoints preserve strong pairwise margins rather than clipping them at 1 or 5. Formally, given an input consisting

of a deliberation question q_i , an annotator opinion $o_S^{(a)}$, and a candidate summary $S_{\mathcal{M}, \tilde{\mathcal{O}}_{i,n}}$ produced by model $\mathcal{M}$ from opinion subset $\tilde{\mathcal{O}}_{i,n}$, the judge $\mathcal{J}_\theta$ encodes the concatenated sequence

$$[\text{CLS}] \ q_i \ [\text{SEP}] \ o_S^{(a)} \ [\text{SEP}] \ S_{\mathcal{M}, \tilde{\mathcal{O}}_{i,n}} \ [\text{SEP}]$$

and outputs a four-dimensional normalized score vector:

$$\hat{\mathbf{y}} = \mathcal{J}_\theta(q_i, o_S^{(a)}, S_{\mathcal{M}, \tilde{\mathcal{O}}_{i,n}}) = (\hat{y}^{(\text{rep})}, \hat{y}^{(\text{inf})}, \hat{y}^{(\text{neu})}, \hat{y}^{(\text{pol})}) \in [0, 1]^4.$$

Here the [CLS] representation from the final encoder layer is passed through a hidden layer and a linear projection to produce the four regression outputs. Human annotations $\mathbf{y}^{\text{raw}} \in [-1, 7]^4$ are linearly normalized to $\mathbf{y} \in [0, 1]^4$ for training stability. The model is trained with the Huber loss averaged across dimensions:

$$\mathcal{L}(\theta) = \frac{1}{|\hat{\mathcal{T}}_{\text{train}}|} \sum_{(q, o, S, \mathbf{y}) \in \hat{\mathcal{T}}_{\text{train}}} \frac{1}{4} \sum_{d \in \{\text{rep}, \text{inf}, \text{neu}, \text{pol}\}} \ell_\delta(\hat{y}^{(d)}, y^{(d)}),$$

where

$$\ell_\delta(\hat{y}, y) = \begin{cases} \frac{1}{2}(\hat{y} - y)^2 & \text{if } |\hat{y} - y| \leq \delta, \\ \delta \cdot (|\hat{y} - y| - \frac{1}{2}\delta) & \text{otherwise.} \end{cases}$$

At inference time, predictions remain in the $[0, 1]$ range and are used directly as summary scores. Detailed training settings are provided in Appendix K.1. Among the configurations tested, DeBERTa-v3-base performs best; this is an empirical result for our training setup rather than a general claim that larger encoders are inferior (Appendix K.2).

3.3 DELIBERATIONJUDGE Outperforms General-Purpose LLM Judges in Alignment and Efficiency

To verify that DELIBERATIONJUDGE improves both reliability and efficiency, we evaluate its agreement with human judgments on the same held-out test set used for LLM annotation (§3.1). Since DELIBERATIONJUDGE evaluates one summary at a time, each of the 900 test items is split into two regression instances (one per summary), yielding the 1,800-instance set $\hat{\mathcal{T}}_{\text{test}}^{\text{Delib}}$.

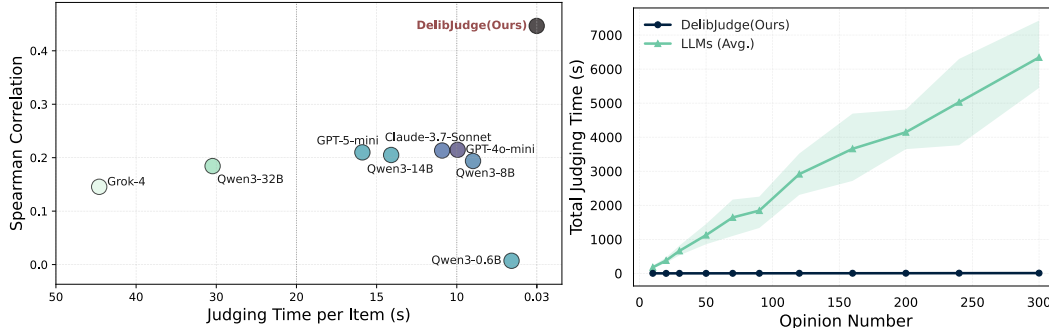


Figure 4: **Left:** Judging Time vs. Spearman Correlation. DELIBERATIONJUDGE achieves both the lowest time and the highest correlation compared to LLMs; **Right:** Scaling Stability. Total judging time of DELIBERATIONJUDGE and eight LLMs as the number of comments increases; solid lines show means and shaded areas denote minimum–maximum ranges. In both panels, “DelibJudge” denotes DELIBERATIONJUDGE.

We compare automatic judges using instance-level Spearman correlation with human judgments, averaged across four dimensions, and inference time per (q, o_i, S) instance. This differs from the model-level rank correlation reported below. In Figure 4 (left), DELIBERATIONJUDGE achieves the highest agreement with human judgments (correlation ≈ 0.48 , vs. ≈ 0.20 for the second-best Claude-3.7-Sonnet). It requires only ~ 0.03 s per item, compared to > 8 s for most LLMs. Although Qwen3-0.6B approaches our method in speed, it performs poorly in judgment quality. Efficiency is especially critical at scale: since each instance requires a separate judgment, LLM costs grow rapidly with opinion count (Figure 4, right),

whereas batched DELIBERATIONJUDGE inference grows substantially more slowly over the tested range.

3.4 Model-Level Rank Correlation with Human Judgments

Table 3: Model-level Spearman rank correlation (human vs. DELIBERATIONJUDGE) across the five Stage 2 summarizers.

	Average	Representativeness	Informativeness	Neutrality	Policy Approval
Spearman ρ	0.70	0.60	1.00	0.90	0.90

We evaluate how well DELIBERATIONJUDGE tracks human judgments on the five Stage 2 summarizers used to build DELIBERATIONBANK. For each dimension, we aggregate the direct human ratings and DELIBERATIONJUDGE predictions over the same held-out items for each summarizer, then compute Spearman correlation across the five resulting model scores. We use Spearman because the benchmark depends on preserving model order under ordinal human ratings rather than reproducing calibrated score differences. As shown in Table 3, DELIBERATIONJUDGE achieves strong overall rank alignment ($\rho = 0.70$), with near-perfect agreement on Informativeness ($\rho = 1.00$) and high alignment on Neutrality and Policy Approval ($\rho = 0.90$ each). Representativeness shows moderate alignment ($\rho = 0.60$). This result is in-domain: training and test instances share the same ten topics. On two unseen topics, per-dimension correlations fall to 0.05–0.23 (Appendix K.3); we therefore treat DELIBERATIONJUDGE as a human-aligned evaluator for DELIBERATIONBANK and nearby deliberation topics, not as a topic-universal judge without further validation.

4 Benchmarking LLMs for Deliberation Summarization

We evaluate both proprietary and open-weight models across diverse scales (Table 7, Appendix G). Let $\mathcal{N} = \{10, 20, 30, 50, 70, 90, 120, 160, 200, 240, 300\}$ denote the input-size grid used for automatic benchmarking; Stage 2 instead uses $\{10, 20, 30, 50, 90\}$ for human annotation. For each question and $n \in \mathcal{N}$, we sample a subset of n opinions from the 300 Stage 1 opinions. Each model generates three summaries per subset to capture stochastic output variability. Here the evaluation set equals the summarizer input, $\mathcal{E} = \tilde{\mathcal{O}}$: DELIBERATIONJUDGE scores every input opinion separately, and the reported representativeness is the aggregate $\text{AggRep}(q, \tilde{\mathcal{O}}, S)$. The other three participant-conditioned scores are averaged over the same opinion–summary pairs, and 95% confidence intervals summarize variability across the evaluated samples. For the Overall score in Table 4, we first average the four dimensions over opinions and generations within each question–size condition. We then give equal weight to the 11 input sizes and ten questions, so larger subsets do not receive more weight merely because they contain more opinions.

Table 4: Overall average performance (mean $\pm$ 95% CI half-width) across four evaluation dimensions (Representativeness, Informativeness, Neutrality, and Policy Approval). Models are arranged from left to right and top to bottom in descending order of mean performance.

Model	Overall	Model	Overall	Model	Overall
GPT-5-Mini	0.622 $\pm$ 0.004	GPT-5	0.617 $\pm$ 0.005	Claude-Sonnet-4	0.615 $\pm$ 0.004
Claude-Opus-4	0.614 $\pm$ 0.004	Qwen3-32B	0.609 $\pm$ 0.006	Gemini-2.5-Flash	0.605 $\pm$ 0.004
Grok-4-Latest	0.594 $\pm$ 0.003	Gemini-2.5-Pro	0.594 $\pm$ 0.005	Qwen3-235B	0.593 $\pm$ 0.004
Qwen3-14B	0.584 $\pm$ 0.003	Qwen3-1.7B	0.583 $\pm$ 0.005	DeepSeek-Reasoner	0.580 $\pm$ 0.004
DeepSeek-Chat	0.576 $\pm$ 0.003	Qwen3-4B	0.575 $\pm$ 0.004	Qwen3-8B	0.575 $\pm$ 0.003
GPT-4o-Mini	0.571 $\pm$ 0.003	Qwen3-30B	0.571 $\pm$ 0.003	Qwen3-0.6B	0.550 $\pm$ 0.004

4.1 Benchmarking Results

As measured by DELIBERATIONJUDGE, models from the GPT-5 and Claude-4 families lead the leaderboard, averaging 0.61–0.62 across four dimensions (Table 4). The Qwen3-32B outperforms several proprietary models (e.g., Gemini-2.5-Pro, Grok-4-Latest), mainly due to its strong *Neutrality* (Figure 5, lower left). Smaller models such as Qwen3-0.6B perform worst, consistent with scaling trends. Overall differences are modest: the best–worst gap

is < 0.1 , indicating that DELIBERATIONJUDGE assigns the evaluated systems a relatively narrow score range. Figure 5 shows similar per-dimension trends, with *Neutrality* tightly clustered (~ 0.02 spread) and the other three dimensions showing wider gaps (0.05–0.1).

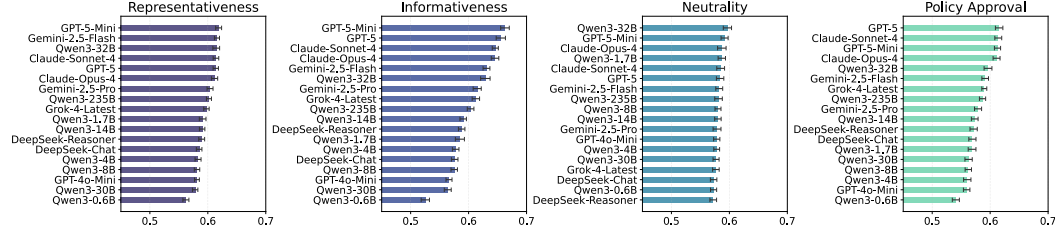


Figure 5: Comparative performance on Representativeness, Informativeness, Neutrality, and Policy Approval (mean $\pm$ 95% CI); higher is better.

4.2 Analysis of Factors Affecting Model Performance

Summarization Input Size. Each deliberation question has 300 opinions, which we partition into subsets of varying sizes to test input effects. Figure 6 (left) shows steady improvement from 10 to roughly 100 opinions followed by a plateau. However, larger random subsets also introduce more viewpoints, so this pattern alone cannot show that models fail to use additional opinions. We therefore embed each topic’s opinions with all-MiniLM-L6-v2, cluster them into five semantic groups with k-means, and sample subsets of sizes 30, 90, 160, and 300 while approximately preserving the full topic’s cluster proportions. Across ten topics and three independent Qwen3-235B generations per size, the resulting 17,400 opinion-level evaluations show that the mean score across the four normalized dimensions rises from 0.572 to 0.605 (paired t -test over 30 topic-generation cells, $p = 0.00022$), while informativeness rises from 0.565 to 0.628 ($p < 0.0001$). Thus additional opinions can help when the coarse viewpoint mixture is controlled; the original plateau is better interpreted as diminishing gains under increasing coverage difficulty. We observe no meaningful opinion-order sensitivity (Appendix E).

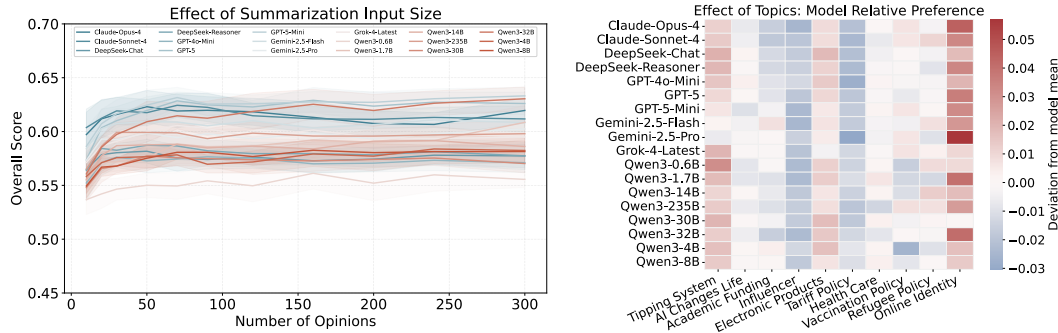


Figure 6: **Left:** Effect of input subset size with $n \in \mathcal{N}$; **Right:** Topic heatmap across models, where each cell shows the topic-wise average score after centering by subtracting the model’s global mean (red = above global mean; blue = below).

Topics. We analyze model sensitivity across topics by computing relative preference scores, defined as the difference between the question-level average and the model’s global average (see Appendix J.2 for the formal definition). As shown in Figure 6 (right), topics such as *Online Identity* and *Tipping System* yield higher-than-average scores, whereas *Influencer* is more difficult. This pattern is consistent with semantic dispersion measured as mean pairwise cosine distance between all-MiniLM-L6-v2 embeddings: *Influencer* is substantially more dispersed (0.651) than *Tipping System* (0.379). Topic sensitivity therefore depends not only on binary versus open-ended format but also on how broadly each question elicits viewpoints.

Question Type. Our benchmark includes five *Binary* and five *Open-Ended* deliberation questions. Figure 7 (right) ranks models by $|\Delta_{\text{type}}|$, the absolute difference between *Binary* and *Open-Ended* scores; smaller $|\Delta_{\text{type}}|$ indicates lower sensitivity. Most smaller models are consistent across types but at low absolute scores; GPT-5 combines high performance with

small sensitivity. Overall, 11 of 18 models score higher on binary questions, likely because binary framing constrains responses to two clear positions, while open-ended prompts elicit diverse opinions that are harder to summarize.

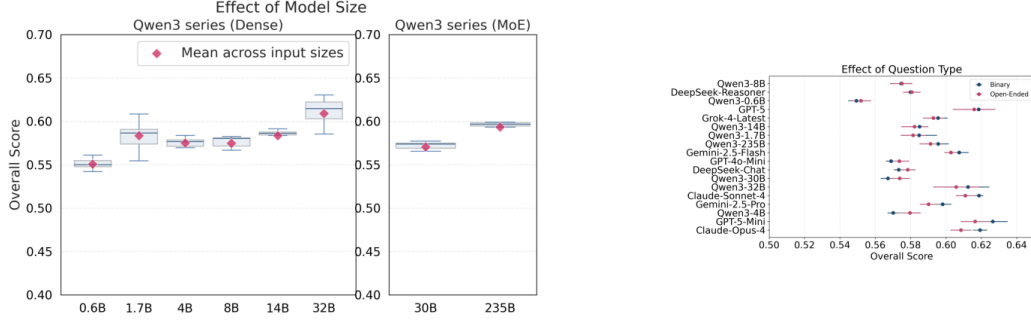


Figure 7: **Left:** Effect of model size, showing overall score of the Qwen3 family, comparing Dense and MoE architectures from 0.6B to 235B; **Right:** Effect of question type (Binary vs. Open-Ended), with models ranked by $|\Delta_{\text{type}}|$, the absolute difference between type-wise average scores (top = most robust).

Model Size. We next examine the effect of model size on summarization performance. Figure 7 (left) shows boxplots of each Qwen3 model’s overall score comparing Dense (Xiao et al., 2024) and Mixture-of-Experts (MoE) (Mu & Lin, 2025) architectures. Consistent with the benchmark-level trend in §4.1, larger models achieve higher scores: within the dense series, performance rises steadily from 0.6B to 32B, while in the MoE series, the 235B model surpasses its 30B counterpart. Gains from scaling are evident but not perfectly monotonic, as smaller models (e.g., 1.7B) show fluctuations, suggesting that architectural design and training stability also influence performance beyond raw parameter count.

5 Case Study: How Well Can LLMs Represent Minority Opinions?

To examine whether LLMs face challenges in representing minority opinions, we conduct a case study on two deliberation questions: a *Binary* question on *Tariff Policy* and an *Open-Ended* question on *AI Changes Life*. For each question q , we use 1,000 opinions from a newly collected response pool that is separate from the Stage 1 data; we denote this summarizer input by $\mathcal{O}_{1000}$ and generate $S = \mathcal{M}(q, \mathcal{O}_{1000})$. “New” therefore refers to the data collection, not to held-out evaluation opinions: all 1,000 opinions are supplied to the summarizer and then scored individually, so $\mathcal{E} = \tilde{\mathcal{O}} = \mathcal{O}_{1000}$. For any evaluated group $G \subseteq \mathcal{O}_{1000}$ and any two groups A and B , we define

$$R_G = \text{AggRep}(q, G, S), \quad \Delta_{\text{rep}}(A, B) = R_B - R_A.$$

Each minority definition partitions the pool into $G_{\min}$ and its complement G_{non} . A positive $\Delta_{\text{rep}}(G_{\min}, G_{\text{non}})$ means that opinions assigned to the minority group receive lower average individual-level coverage than the remaining opinions.

5.1 Self-Reported Minority

Self-Reported Minority Data Collection. Participants answer *Do you think your opinion differs from that of most people in the U.S.?* with *Yes*, *No*, or *I’m not sure*. We define $G_{\min}$ as the *Yes* responses and G_{non} as the complement, including both *No* and *I’m not sure*. This is a subjective measure of perceived social difference, not a claim about the opinion’s empirical prevalence in the U.S. population. We use 1,000 responses per question, rather than the 300 used in Stage 1, to obtain more stable subgroup estimates.

Key Findings. We focus on *representativeness*, as it directly measures whether an individual opinion is covered by the summary. As shown in Figure 8 (right), across all summarizers, self-reported minority opinions receive lower scores than non-minority opinions. To test whether this gap is an artifact of DELIBERATIONJUDGE, we conduct a direct human validation on 60 matched blocks across the two topics. Within each block, the summary is fixed and a self-reported minority opinion is paired with a nearest-neighbor non-minority opinion matched

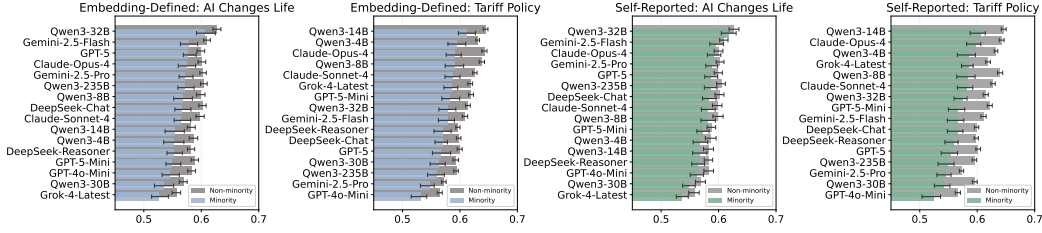


Figure 8: DELIBERATIONJUDGE-predicted representativeness for minority versus non-minority opinions under *embedding-defined semantic-outlier status* (Left) and *self-reported status* (Right) across two questions. Summarizer input size is 1,000.

by topic, embedding similarity, and length. Annotators are not told the minority status, matching procedure, or study hypothesis, and provide three coverage ratings per item (360 ratings total). Human-rated representativeness is 3.100 for minority opinions and 3.489 for matched non-minority opinions, a paired gap of 0.389 (95% bootstrap CI [0.128, 0.650]; sign-flip $p = 0.0046$). This independently corroborates the self-reported minority result; it does not validate the result for embedding-defined semantic outliers below.

5.2 Embedding-Defined Semantic Minority

Embedding-Based Semantic Outlier Detection. We introduce an algorithmic minority definition based on semantic outlier detection. This status is assigned reproducibly from text embeddings rather than self-reported; it does not imply a ground-truth social minority category. All responses are encoded using all-MiniLM-L6-v2, and Local Outlier Factor (LOF) with 20 neighbors measures local semantic rarity. For binary questions, we group responses by stated stance and define the top 10% by LOF within each stance as $G_{\min}$, surfacing uncommon reasoning without conflating it with the less frequent stance. For open-ended questions, we first cluster embeddings into eight themes with k-means and likewise select the top 10% by LOF within each theme. All remaining opinions form G_{non} (details in Appendix F.2).

Key Findings. Figure 8 (left) shows that embedding-defined semantic outliers receive lower representativeness scores, with average gaps of 0.04 on *AI Changes Life* and 0.03 on *Tariff Policy*. A random-sample sanity check yields gaps of only 0.001–0.002, over $30\times$ smaller (Appendix F.3). We extend this analysis to all 10 topics in Appendix F.4.

5.3 Two Faces of Minority Opinion Representation Bias

Although both methods reveal minority under-coverage, they capture distinct populations with Jaccard similarity below 10% (Table 5, Appendix F.3). Self-reported minority status measures perceived social difference and can include semantically common opinions, whereas embedding-defined semantic-outlier status identifies locally rare argumentation patterns. Our direct human validation establishes the coverage gap only for the self-reported group; the result for embedding-defined semantic outliers remains judge-based secondary evidence. The low overlap shows that “minority opinion” is not a single property. An opinion may feel socially different to its author while remaining semantically common, or may contain rare reasoning without its author self-identifying as a minority. Consequently, auditing only one definition can miss a distinct form of under-coverage. We therefore treat self-reported social difference and embedding-defined semantic rarity as complementary axes and report them separately rather than collapsing them into one minority label.

6 Conclusion

Evaluating deliberation summaries at scale is difficult because quality is participant-dependent. We introduce DELIBERATIONBANK and DELIBERATIONJUDGE, an efficient, domain-aligned evaluator. Across 18 models, we find lower predicted coverage for self-reported opinions and embedding-defined semantic outliers; a matched human study corroborates the self-reported gap on two topics. Gains also diminish as input size and coverage difficulty grow. These results expose limits of off-the-shelf LLM judges and motivate new-topic validation of DELIBERATIONJUDGE.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- Anthropic. Claude opus 4 & claude sonnet 4 system card. <https://thoughtshrapnel.com/uploads/2025/claude-4-system-card.pdf>, 2025. Accessed 8 Sept. 2025.
- Michiel Bakker, Martin Chadwick, Hannah Sheahan, Michael Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matt Botvinick, et al. Fine-tuning language models to find agreement among humans with diverse preferences. *Advances in neural information processing systems*, 35:38176–38189, 2022.
- Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv:2004.05150*, 2020.
- Gijs Jan Brandsma and Simon Otjes. Gauging the roles of parliamentary staff. *Parliamentary Affairs*, 77(3):537–557, 2024.
- Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-Graber, and David Blei. Reading tea leaves: How humans interpret topic models. *Advances in neural information processing systems*, 22, 2009.
- Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.
- Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*, 2023.
- Evidently AI. LLM-as-a-judge: a complete guide to using LLMs for evaluations. online guide, 2025. URL <https://www.evidentlyai.com/llm-guide/llm-as-a-judge>.
- Sara Fish, Paul Gözl, David C Parkes, Ariel D Procaccia, Gili Rusak, Itai Shapira, and Manuel Wüthrich. Generative social choice. *arXiv preprint arXiv:2309.01291*, 2023.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020.
- Hui Huang, Xingyuan Bu, Hongli Zhou, Yingqi Qu, Jing Liu, Muyun Yang, Bing Xu, and Tiejun Zhao. An empirical study of llm-as-a-judge for llm evaluation: Fine-tuned judge model is not a general substitute for gpt-4. *arXiv preprint arXiv:2403.02839*, 2024.
- Michael Krumbick, Charles Lovering, Varshini Reddy, Seth Ebner, and Chris Tanner. No free labels: Limitations of llm-as-a-judge without human grounding. *arXiv preprint arXiv:2503.05061*, 2025.
- Hélène Landemore. Open democracy: Reinventing popular rule for the twenty-first century. 2020.
- Nayeon Lee, Yejin Bang, Tiezheng Yu, Andrea Madotto, and Pascale Fung. Neus: neutral multi-news summarization for mitigating framing bias. *arXiv preprint arXiv:2204.04902*, 2022.
- Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. Llm-as-judges: A comprehensive survey on LLM-based evaluation methods. *arXiv preprint*, 2024a. URL <https://arxiv.org/abs/2412.05579>.

- Lincan Li, Jiaqi Li, Catherine Chen, Fred Gui, Hongjia Yang, Chenxiao Yu, Zhengguang Wang, Jianing Cai, Junlong Aaron Zhou, Bolin Shen, et al. Political-llm: Large language models in political science. *arXiv preprint arXiv:2412.06864*, 2024b.
- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.
- Sean MacAvaney and Luca Soldaini. One-shot labeling for automatic relevance estimation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1986–1990, 2023.
- Bruno Marnette and Colleen McKenzie. Talk to the city: an open-source AI tool for scaling deliberation. AI Objectives Institute blog, 2024. URL <https://ai.objectives.institute/blog/talk-to-the-city-an-open-source-ai-tool-to-scale-deliberation>.
- Siyuan Mu and Sen Lin. A comprehensive survey of mixture-of-experts: Algorithms, theory, and applications. *arXiv preprint arXiv:2503.07137*, 2025.
- Yi Nian, Shenzhe Zhu, Yuehan Qin, Li Li, Ziyi Wang, Chaowei Xiao, and Yue Zhao. Jaildam: Jailbreak detection with adaptive memory for vision-language model. *arXiv preprint arXiv:2504.03770*, 2025.
- OpenAI. Gpt-5 system card. <https://cdn.openai.com/gpt-5-system-card.pdf>, 2025. Accessed 8 Sept. 2025.
- Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Jackson Sargent, Apostolos Dedeloudis, and David Jurgens. Potato: The portable text annotation tool. *arXiv preprint arXiv:2212.08620*, 2022.
- Jiaxin Pei, José Ramón Enríquez, Umar Patel, Alia Braley, Nuole Chen, Lily Tsai, and Alex Pentland. DELIBERATION.IO: Facilitating Democratic and Civil Engagement at Scale with Open-Source and Open-Science. <https://drive.google.com/file/d/1dWYJnVWnzSC14MmebKkVI8ZYYEodkVZl/view>, 2024. Accessed: 2025-09-07.
- Margaret E Roberts, Brandon M Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G Rand. Structural topic models for open-ended survey responses. *American journal of political science*, 58(4):1064–1082, 2014.
- Victor S Sheng, Foster Provost, and Panagiotis G Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 614–622, 2008.
- Carson Sievert and Kenneth Shirley. Ldavis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pp. 63–70, 2014.
- Christopher T. Small, Michael Björkegren, Timo Erkkilä, Lynette Shaw, and Colin Megill. Polis: Scaling deliberation by mapping high dimensional opinion spaces. *Recerca. Revista de Pensament i Anàlisi*, 26(2):1–26, 2021. doi: 10.6035/recerca.5516.
- Christopher T. Small, Ivan Vendrov, Esin Durmus, Hadjar Homaei, Elizabeth Barry, Julien Cornebise, Ted Suzman, Deep Ganguli, and Colin Megill. Opportunities and risks of LLMs for scalable deliberation with Polis. *arXiv preprint*, 2023. URL <https://arxiv.org/abs/2306.11932>.
- Rion Snow, Brendan O’Connor, Dan Jurafsky, and Andrew Y Ng. Cheap and fast—but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 254–263, 2008.

- Annalisa Szymanski, Noah Ziemis, Heather A. Eicher-Miller, Toby Jia-Jun Li, Meng Jiang, and Ronald A. Metoyer. Limitations of the LLM-as-a-Judge approach for evaluating LLM outputs in expert knowledge tasks. *arXiv preprint*, 2024. URL <https://arxiv.org/abs/2410.20266>.
- Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL <https://arxiv.org/pdf/2505.09388>.
- Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. *arXiv preprint arXiv:2406.12624*, 2024.
- Supriti Vijay, Aman Priyanshu, and Ashique R KhudaBukhsh. When neutral summaries are not that neutral: Quantifying political neutrality in llm-generated news summaries. *arXiv preprint arXiv:2410.09978*, 2024.
- xAI. Grok 4 model card. <https://data.x.ai/2025-08-20-grok-4-model-card.pdf>, 2025. Accessed 8 Sept. 2025.
- Chaojun Xiao, Jie Cai, Weilin Zhao, Guoyang Zeng, Biyuan Lin, Jie Zhou, Zhi Zheng, Xu Han, Zhiyuan Liu, and Maosong Sun. Densing law of llms. *arXiv preprint arXiv:2412.04315*, 2024.
- Shu Yang, Shenzhe Zhu, Zeyu Wu, Keyu Wang, Junchi Yao, Junchao Wu, Lijie Hu, Mengdi Li, Derek F Wong, and Di Wang. Fraud-r1: A multi-round benchmark for assessing the robustness of llm against augmented fraud and phishing inducements. *arXiv preprint arXiv:2502.12904*, 2025.
- Junchi Yao, Jianhua Xu, Tianyu Xin, Ziyi Wang, Shenzhe Zhu, Shu Yang, and Di Wang. Is your llm-based multi-agent a reliable real-world planner? exploring fraud detection in travel planning. *arXiv preprint arXiv:2505.16557*, 2025.
- Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, et al. Justice or prejudice? quantifying biases in llm-as-a-judge. *arXiv preprint arXiv:2410.02736*, 2024.
- Shenzhe Zhu, Jiao Sun, Yi Nian, Tobin South, Alex Pentland, and Jiaxin Pei. The automated but risky game: Modeling agent-to-agent negotiations and transactions in consumer markets. *arXiv preprint arXiv:2506.00073*, 2025.

A Ethics and Reproducibility

Participants were recruited through Prolific and completed the study through POTATO or Deliberation.io after viewing the study information and consent materials. Our analysis uses de-identified responses, and the research team did not collect participant names or contact details. Participants were compensated at the posted study rates. The public repository¹ contains the released opinion data, generated summaries, annotations, preprocessing and evaluation code, and training configuration. Detailed hyperparameters and protocols are documented in Appendix K.

B Use of Large Language Models (LLMs)

LLMs were used occasionally for copyediting, including suggestions on word choice, grammar, and spelling. All technical ideas, experimental designs, analyses, and conclusions were developed by the authors, who have reviewed the final text and take responsibility for it.

C Related Work

C.1 Scalable Deliberation

Early efforts to scale deliberation relied on statistical topic models and clustering to reveal themes in large comment collections but still required expert interpretation. A notable example is *Polis*, which projects voting data into a 2D opinion space via PCA and applies *k*-means clustering to identify groups and surface consensus statements (Small et al., 2021). While effective for visualization, such methods depend on moderators to draft reports. Recent work uses LLMs to automate this synthesis. Marnette et al. (Marnette & McKenzie, 2024) process thousands of comments with LLMs to extract arguments, cluster positions, and generate summaries. Experiments integrating LLM summarizers into *Polis* show that they can support moderators, although context-window limits remain a challenge (Small et al., 2023). Our work extends this line by providing a large-scale dataset and evaluation framework for LLM-based deliberation summarization.

C.2 LLMs as Judges

Another recent line of work proposes using large language models themselves as evaluators of generated text (Yao et al., 2025; Yang et al., 2025; Zhu et al., 2025; Nian et al., 2025). This approach promises to scale evaluation but raises serious challenges. Studies in expert domains report that subject-matter experts agree with LLM judges only about two-thirds of the time when assessing complex outputs (Szymanski et al., 2024), indicating limited reliability. A comprehensive survey catalogs biases in LLM judges, such as position, verbosity, and authority effects, and notes that their judgments are highly sensitive to prompt design (Li et al., 2024a). Practical guides add that LLM evaluators may produce inconsistent or biased scores and raise privacy and cost concerns (Evidently AI, 2025). Taken together, these observations indicate that while the LLM-as-a-judge paradigm holds promise for scalable evaluation, domain-specific evaluators tuned to human preferences are still necessary.

D Future Work and Broader Impact

Our study shows that DELIBERATIONJUDGE aligns more closely with human judgments than general-purpose LLM judges, though instance-level correlations remain moderate and performance decreases on two unseen topics. Expanding DELIBERATIONBANK beyond the current ten questions would improve coverage of new deliberation domains. Diversity-aware or stance-conditioned evaluation is another natural extension. Objectives that explicitly model viewpoint diversity, such as contrastive formulations across stances

¹<https://github.com/shuyhere/LLMs-for-Scalable-Deliberation>

or subgroups, may improve an evaluator’s ability to capture fine-grained differences in perspective coverage.

E Input-Order Bias Analysis

We inspect positional effects using 2,000 newly collected opinions, with 1,000 for each of the two topics described in §5. For each topic, we construct two input prefixes containing the first 500 and all 1,000 opinions, yielding four evaluation splits for each summarization model. Within each split, we compute the average representativeness score by input position across resampled summaries and smooth the trajectory by grouping every 20 positions into bins and plotting the mean with standard error. We repeat this procedure for two summarizers, GPT-5-Mini and Qwen3-30B. Across both models, the curves do not display any consistent upward or downward pattern. Instead, the trajectories fluctuate without a clear monotonic direction, suggesting that positional effects are not a major source of bias in our setting.

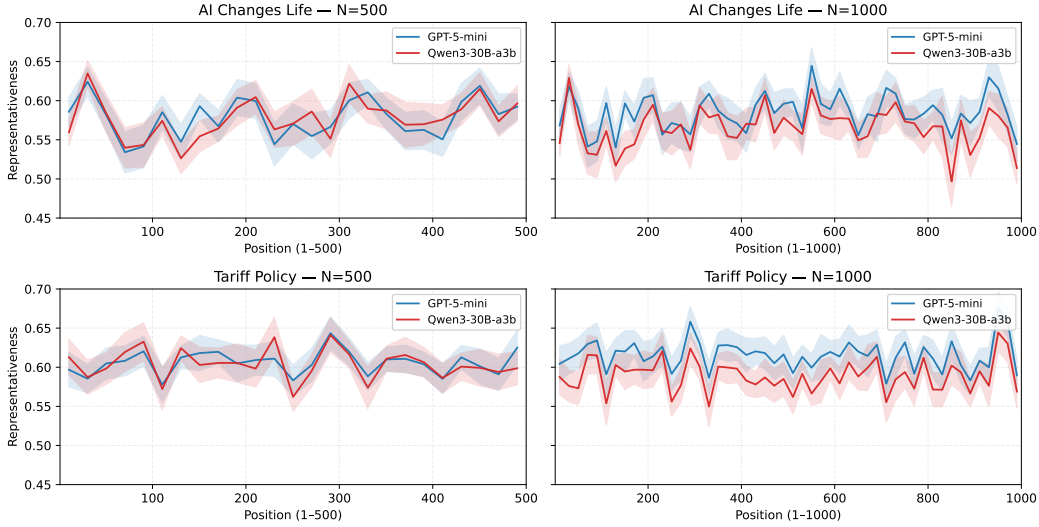


Figure 9: Representativeness trajectories (mean $\pm$ SE over 20-position bins) for GPT-5-Mini and Qwen3-30B when summarizing the first 500 and all 1,000 opinions for each topic.

F Further Experimental Details and Minority-Bias Results

F.1 Self-Reported Minority

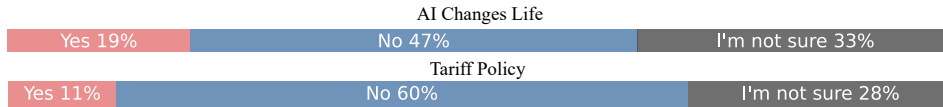


Figure 10: Distribution of self-reported minority opinions in the newly collected data for two deliberation questions.

F.2 Embedding-Defined Semantic Minority

Local Outlier Factor (LOF).

We provide the formal definition of the Local Outlier Factor used in our semantic-outlier detection pipeline. For each opinion embedding x , let $N_k(x)$ denote its k -nearest neighbors.

The *local reachability density* (LRD) of x is defined as:

$$\text{LRD}_k(x) = \left(\frac{1}{|N_k(x)|} \sum_{y \in N_k(x)} \max\{d(x, y), \text{k-dist}(y)\} \right)^{-1},$$

where $d(\cdot, \cdot)$ is the Euclidean distance and $\text{k-dist}(y)$ is the distance from y to its k -th nearest neighbor.

The *local outlier factor* of x is then:

$$\text{LOF}_k(x) = \frac{1}{|N_k(x)|} \sum_{y \in N_k(x)} \frac{\text{LRD}_k(y)}{\text{LRD}_k(x)}.$$

Values $\text{LOF}_k(x) > 1$ indicate that x lies in a region of lower density relative to its neighbors and is therefore more likely to be an outlier.

Representativeness-Gap Definition. For a fixed question q and summary S , let $R_G = \text{AggRep}(q, G, S)$. We define the representativeness gap of opinion set A relative to set B as

$$\Delta_{\text{rep}}(A, B) = R_B - R_A. \quad (1)$$

Positive values indicate that set A receives lower representativeness than set B . In our analysis, we focus on two instantiated forms of the above metric. The *minority bias* is defined as

$$\Delta_{\text{rep}}(G_{\text{min}}, G_{\text{non}}),$$

which measures the average representativeness gap between minority and non-minority opinion sets.

Similarly, the *random-control gap* is defined as

$$\Delta_{\text{rep}}(G_{\text{selected}}, G_{\text{rest}}),$$

where the selected group is a uniform random sample with the same cardinality as the detected minority group, and the non-selected group is the remainder of the pool. This construction provides a size-matched random baseline.

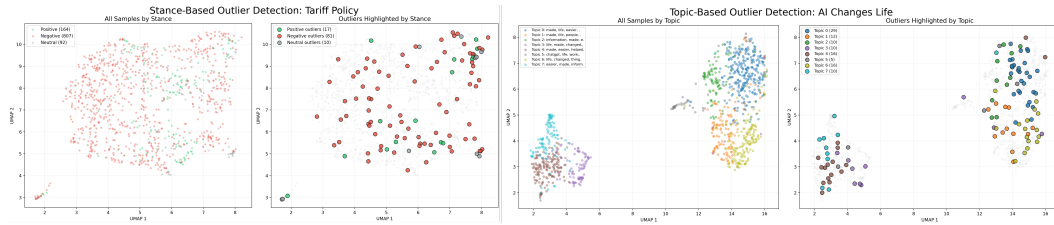


Figure 11: **Top:** Stance-based outlier detection for *Tariff Policy*. The left panel shows all samples colored by stance, and the right panel highlights within-stance outliers in the same UMAP space. **Bottom:** Topic-based outlier detection for *AI Changes Life*. The left panel shows samples colored by semantic topic from k-means, and the right panel highlights within-topic outliers.

Embedding-Based Semantic Outlier Detection. We use an embedding-based outlier detection method to identify semantically distinctive opinions. All responses are first encoded into 384-dimensional embeddings using the all-MiniLM-L6-v2 sentence-transformer model, followed by a grouping step that adapts to the question type. Local Outlier Factor (LOF, 20 neighbors) is then applied within each group to detect semantically rare opinions. This shared embedding and LOF configuration ensures comparability across settings.

For binary questions (i.e., *Tariff Policy*), we use the explicit stance labels collected in the survey. Responses are partitioned into positive, negative, and neutral groups, after which LOF is applied separately within each group. This setup captures respondents who share an overall stance but reach it through uncommon reasoning. We select the top 10% of responses by LOF score within each stance, identifying 108 of 1,063 responses (10.1%) after rounding within groups.

For open-ended questions (i.e., *AI Changes Life*), where no stance labels exist, we cluster standardized embeddings into eight topical groups using k-means. LOF is then applied within each topic cluster to identify responses that convey rare perspectives or exceptional experiences within the same theme. This two-stage design focuses the outlier criterion on semantic distinctiveness within a coarse theme.

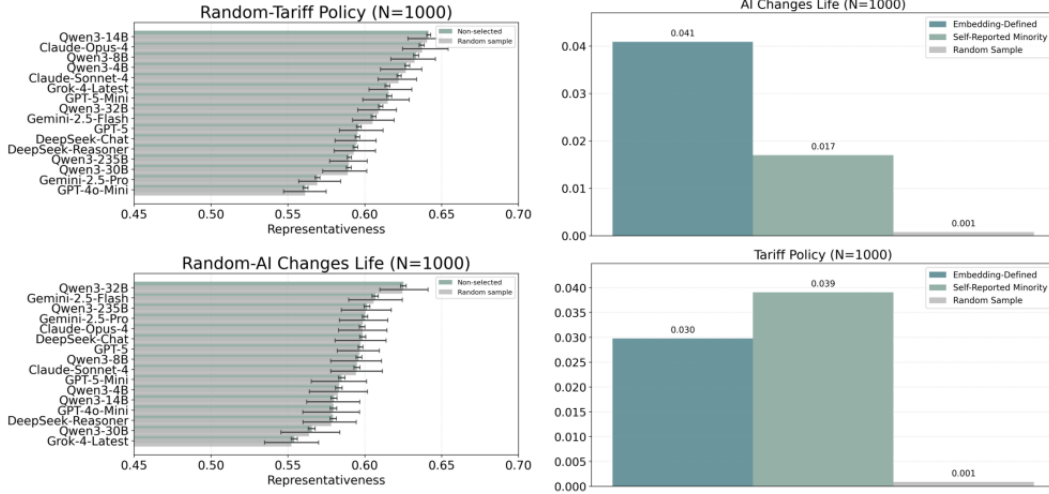


Figure 12: **Left:** Representativeness scores of randomly selected and non-selected opinions across all summarizers. Random samples show near-zero gaps. **Right:** Embedding-defined semantic-outlier gaps compared with the random control. The semantic-outlier gaps are 35–54 $\times$ larger.

E.3 Zoom In: Analysis of the Two Minority Definitions

In §5, minority groups identified through both self-reports and embedding-based outlier detection receive lower representativeness scores than non-minority opinions across the evaluated summarizers (Figure 8). To test whether this effect is driven by group size, we randomly select the same number of opinions as in the embedding-defined group and compare them with the remaining pool. The resulting gap is near zero (Figure 12), indicating that group size alone does not explain the observed difference.

Table 5: Overlap between self-reported minorities and embedding-defined semantic outliers across two topics. Low Jaccard similarity and intersection rates indicate that the methods identify largely distinct sets: perceived social difference versus embedding-based semantic rarity.

	Jaccard Similarity	Intersection Rate
Tariff Policy	0.090	0.018
AI Changes Life	0.096	0.026

Despite similar under-coverage patterns, the two minority sets overlap little. As shown in Table 5, both Jaccard similarity and intersection rate remain below 10%, indicating that the definitions capture different properties of an opinion.

To illustrate the difference between the definitions, Table 6 presents representative examples. Self-reported minority status reflects a participant’s perception of differing from most people, even when the response is not semantically rare. Some examples express uncertainty or personal isolation, but these qualitative patterns are illustrative rather than a systematic causal analysis.

In contrast, embedding-defined semantic outliers are identified by local distinctiveness in the embedding space. These examples contain uncommon argumentation patterns or thematic associations, such as linking tariff policy to racial discourse. The distinct definitions help explain why the two sets overlap by less than 10%.

Table 6: Representative examples illustrating the conceptual differences between self-reported minorities and embedding-defined semantic outliers.

Minority Type	Representative Example (Tariff Policy / AI Changes Life)	Why This Case? What It Reveals
Self-Reported	<i>"I'm not sure if the tariff policy will have a positive or negative impact... more data is needed."</i>	Strong uncertainty and self-perceived isolation; semantically similar to many mainstream responses; minority defined by subjective perception rather than semantic uniqueness
	<i>"AI gives me a judgement-free space to talk about PTSD... I feel alone and unsupported."</i>	Personal narrative and emotional framing; overlaps with common "AI for support" themes; shows how individuals may self-identify as a minority due to personal context
Embedding-Defined	<i>"Tariffs help restart conversations about race and ethnicity... this is good for public discourse."</i>	Rare conceptual framing (tariffs → racial discourse); distinct reasoning but the author does not self-identify as a minority; minority defined by semantic outlier status
	<i>"AI helps me grow my own SCOBY and ferment kombucha; I use it for very specific tasks."</i>	Highly specific and unusual use case; not self-identified as a minority; shows how semantic uniqueness does not imply self-identification

Despite their different definitions, both groups receive lower representativeness scores in the evaluated model-generated summaries. The low overlap suggests that self-reported social difference and embedding-based semantic rarity expose complementary coverage failures. We do not claim that the qualitative characteristics in Table 6 causally explain these gaps.

F.4 Minority Bias Patterns Across All Topics

We use the representativeness gap defined in Equation 1, with the embedding-defined semantic-outlier group as $G_{\min}$ and the remainder as G_{non} . The random control replaces $G_{\min}$ with a size-matched uniform sample.

Embedding-Defined Minority Bias Over All Topics. Using the previously introduced embedding-based detection procedure, we extend the analysis to all ten benchmark topics. For each topic, we use 300 input opinions and examine the minority bias patterns exhibited by different summarizers. Following the method in §5.2, we identify semantic outliers and then compare their representativeness scores with those of the remaining opinions. We conduct a random control experiment by sampling the same number of opinions uniformly at random.

Across the evaluated summarizers and topics, the embedding-defined semantic-outlier gap is $34.4\times$ the random-control gap on average. Summarizer-level mean gaps range from 0.0310 to 0.0404. This consistency indicates that the observed pattern is not driven by a single summarizer, although the result remains based on DELIBERATIONJUDGE’s predictions rather than direct human ratings.

G Details of Model Choice

As shown below, we include 18 widely used LLMs, covering both proprietary and open-weight models across several major families: OpenAI GPT, Anthropic Claude, Google Gemini, Alibaba Qwen3, DeepSeek, and xAI Grok. Within the Qwen3 series, the displayed names Qwen3-30B and Qwen3-235B refer to the Qwen3-30B-a3b and Qwen3-235B-a22b variants,

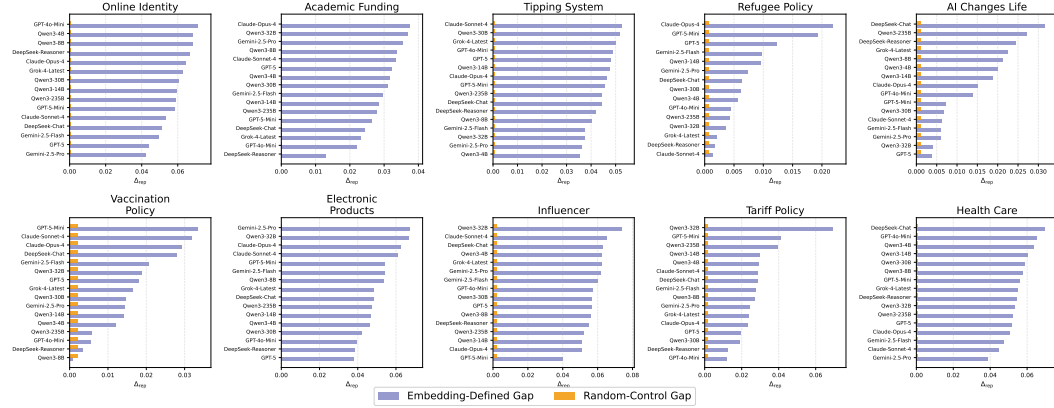


Figure 13: Embedding-defined semantic-outlier gap versus random-control gap across ten topics ($N = 300$). Each subplot shows the representativeness gap between the selected group and the remaining opinions for each summarizer. Purple bars show the embedding-based outlier gap, and orange bars show the random-control gap. Across the evaluated summarizers and topics, the mean gaps are approximately 0.0364 and 0.0011, respectively ($34.4\times$).

which use Mixture-of-Experts (MoE) architectures with 3B and 22B active parameters during inference, respectively. DeepSeek-Reasoner and DeepSeek-Chat denote the thinking and non-thinking DeepSeek-V3.1 endpoints.

Table 7: List of evaluated summarizers, including both API-based frontier systems and open-weight models of varying scales.

Model	Parameters	Access	Creator	Model	Parameters	Access	Creator
GPT-4o-Mini (Achiam et al., 2023)	N/A	API	OpenAI	Qwen3-0.6B (Team, 2025)	0.6B	Open-weight	Alibaba
GPT-5-Mini (OpenAI, 2025)	N/A	API	OpenAI	Qwen3-1.7B (Team, 2025)	1.7B	Open-weight	Alibaba
GPT-5 (OpenAI, 2025)	N/A	API	OpenAI	Qwen3-4B (Team, 2025)	4B	Open-weight	Alibaba
Claude-Sonnet-4 (Anthropic, 2025)	N/A	API	Anthropic	Qwen3-8B (Team, 2025)	8B	Open-weight	Alibaba
Claude-Opus-4 (Anthropic, 2025)	N/A	API	Anthropic	Qwen3-14B (Team, 2025)	14B	Open-weight	Alibaba
Gemini-2.5-Flash (Comanici et al., 2025)	N/A	API	Google	Qwen3-30B (Team, 2025)	30B	Open-weight	Alibaba
Gemini-2.5-Pro (Comanici et al., 2025)	N/A	API	Google	Qwen3-32B (Team, 2025)	32B	Open-weight	Alibaba
DeepSeek-Reasoner (Guo et al., 2025)	671B	API	DeepSeek	Qwen3-235B (Team, 2025)	235B	Open-weight	Alibaba
DeepSeek-Chat (Liu et al., 2024)	671B	API	DeepSeek	Grok-4-Latest (xAI, 2025)	N/A	API	xAI

H Details of Public Opinion Datasets

H.1 Deliberation Questions

As shown in Table 8, we construct ten deliberation questions using a two-stage process. We first use OpenAI’s Deep Research tool to gather trending public discussion topics from social media and online news, then manually screen out extreme or unsuitable prompts. The final set includes five binary-choice policy questions and five open-ended societal questions, covering both structured policy debates and broader social issues. Table 8 preserves the wording shown to participants.

H.2 Word Clouds

The word clouds (see Figure 14) visualize participant opinions, highlighting salient terms and thematic differences across binary and open-ended deliberation topics.

H.3 Data Quality Checking

For both human-data stages, we monitored completion times and filtered submissions completed implausibly quickly. Before full annotation, we manually inspected a 100-item pilot to verify that participants understood the rubric. Each participant answered one assigned deliberation question, so there was no within-participant question-order

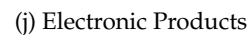
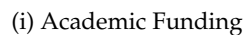
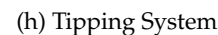
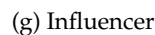
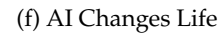
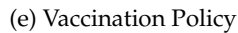
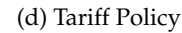
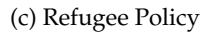
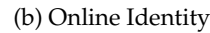
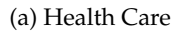


Table 8: Deliberation questions using the wording shown to participants.

Topic	Type	Prompt
"Tipping System"	Open-Ended	What is your opinion on tipping, and if given the chance, how would you improve or change the current tipping system?
"AI Changes Life"	Open-Ended	How has AI changed your life?
"Academic Funding"	Open-Ended	What are your thoughts on Trump's decision to cut academic funding?
"Influencer"	Open-Ended	What is your opinion on internet influencers (e.g., streamers, bloggers, short video creators) increasingly becoming a recognized profession?
"Electronic Products"	Open-Ended	What is your opinion on the rapid update cycle of electronic products, especially smartphones?
"Tariff Policy"	Binary	Do you think the current tariff policy under the Trump administration will have a positive or negative impact on the overall U.S. economy and society?
"Health Care"	Binary	Do you support the government provide basic health insurance for everyone?
"Vaccination Policy"	Binary	Do you support the government having the authority to enforce vaccination and quarantine measures during severe epidemics?
"Refugee Policy"	Binary	Do you support the government accepting more refugees fleeing war or persecution?
"Online Identity"	Binary	Do you support requiring real-name registration on social media platforms, where users must register and post under their real identity?

randomization. The four rating dimensions appeared in a fixed order. As a post hoc check, representativeness correlates moderately, rather than nearly perfectly, with the later dimensions (Pearson $r = 0.383$ – 0.521), and only 15.6% of annotations assign the same score to all four dimensions. These results reduce, but do not eliminate, the possibility of order-induced anchoring.

I Details of Human Annotation Process

I.1 Ring-Based Summary Matching Algorithm

For the comparison task (see §2.2), we adopt a ring-based matching algorithm (Alg. 1) rather than independent random sampling. This ensures balanced and repeated pairing within each deliberation question. Pairs always share the same question but need not share the same input size or exact source subset; the augmented training set therefore exposes the judge to both within-size and cross-size preferences.

Pairing is performed independently for each question: summaries are aggregated and randomly permuted before pairing. In the default mode, let m be the number of summaries for a question and $k = \text{min_comparisons_per_summary}$ (default $k = 6$). For each index $i \in \{0, \dots, m-1\}$ and each offset $o \in \{1, \dots, k\}$, we generate

$$(A, B) = (s_i, s_{(i+o) \bmod m}).$$

If a total number of pairs M is specified instead, we compute $k = \lfloor M/m \rfloor$ and $r = M \bmod m$, generate all pairs for $o \in \{1, \dots, k\}$ as above, and then add one extra pair with offset $k+1$ for the first r indices.

J Details of Metrics

J.1 Summarization Quality Metric

We evaluate summary quality along four deliberation-relevant dimensions:

Algorithm 1 Ring-Based Summary Matching**Require:** Summaries $S = [s_0, \dots, s_{m-1}]$, seed, either k or M

```

1:  $S \leftarrow \text{PERMUTE}(S, \text{seed}); m \leftarrow |S|$ 
2: if  $M$  is specified then
3:    $k \leftarrow \lfloor M/m \rfloor; r \leftarrow M \bmod m$ 
4: else
5:    $k \leftarrow \text{min\_comparisons\_per\_summary}$ 
6: end if
7:  $\text{Pairs} \leftarrow \emptyset$ 
8: for  $i = 0$  to  $m - 1$  do
9:   for  $o = 1$  to  $k$  do
10:     $j \leftarrow (i + o) \bmod m$ 
11:    append  $(s_i, s_j)$  to  $\text{Pairs}$ 
12:   end for
13: end for
14: if  $M$  is specified then
15:   for  $i = 0$  to  $r - 1$  do
16:     $j \leftarrow (i + (k + 1)) \bmod m$ 
17:    append  $(s_i, s_j)$  to  $\text{Pairs}$ 
18:   end for
19: end if
20: return  $\text{Pairs}$ 

```

- **Representativeness:** To what extent does the summary reflect the annotator’s own stated opinion? This is the individual-level coverage target $\text{IndRep}(q, o_i, S)$.
- **Informativeness:** How much useful, diverse, detailed, and non-redundant information does the summary provide to the annotator?
- **Neutrality:** Does the summary present a balanced and unbiased view of the issue rather than favoring one position?
- **Policy-use acceptability (annotation label: Policy Approval):** Would the annotator approve of this summary being used by policymakers as a representation of the deliberation? This does not ask the annotator to judge whether a policy is correct.

Human annotation. These four metrics were operationalized in two complementary annotation tasks. As shown in Figure 15, in the *rating task*, each summary was independently scored on a five-point Likert scale (poor, slightly poor, neutral, slightly good, good) for all four dimensions. In the *comparison task*, annotators compared two summaries for the same question and indicated which performed better on each dimension using a five-level ordinal scale.

Automatic benchmarking. In our evaluation framework, DELIBERATIONJUDGE outputs a normalized four-dimensional prediction for each (q, o_i, S) tuple. Its representativeness coordinate is an individual opinion–summary score; aggregating it over a specified evaluation group yields AggRep . The other coordinates predict that participant’s summary-level judgments. We average these pair-level predictions only when reporting summary-, group-, condition-, or model-level statistics.

J.2 Analytical Metric

Relative Preference Score For model $\mathcal{M}$, question q , input size $n \in \mathcal{N}$, and dimension d , let $\mathcal{I}_{\mathcal{M},q,n}$ contain the evaluated opinion–summary pairs across generations for that condition, and define

$$\bar{y}_{\mathcal{M},q,n}^{(d)} = \frac{1}{|\mathcal{I}_{\mathcal{M},q,n}|} \sum_{z \in \mathcal{I}_{\mathcal{M},q,n}} y_z^{(d)}.$$

The figure displays two side-by-side screenshots of annotation interfaces, labeled (a) and (b).

(a) Rating task with Likert-scale options: This interface consists of five vertically stacked question boxes, each with a blue header and a white body containing radio button options.

- Question 1:** "To what extent is your perspective represented in this response?"
 - ☐ Not represented — summary ignores comment
 - ☐ Slightly represented — little relevant content
 - ☐ Partially represented — main idea but gaps
 - ☐ Mostly represented — most info, minor nuance missing
 - ☐ Fully represented — all info included
- Question 2:** "How informative is this summary?"
 - ☐ Not informative — almost no content
 - ☐ Slightly informative — few basic points
 - ☐ Moderately informative — some important details
 - ☐ Very informative — most details included
 - ☐ Extremely informative — highly detailed and comprehensive
- Question 3:** "Do you think this summary presents a neutral and balanced view of the issue?"
 - ☐ Very non-neutral — strongly one-sided
 - ☐ Somewhat non-neutral — noticeable tilt
 - ☐ Moderately neutral — mix of neutrality and tilt
 - ☐ Fairly neutral — mostly impartial
 - ☐ Completely neutral — fully impartial
- Question 4:** "Would you approve of this summary being used by the policy makers to make decisions relevant to the issue?"
 - ☐ Strongly disapprove
 - ☐ Disapprove
 - ☐ Neutral
 - ☐ Approve
 - ☐ Strongly approve

(b) Comparison task with five ordinal outcomes: This interface also consists of five vertically stacked question boxes, each with a blue header and a white body containing radio button options.

- Question 1:** "Which summary is more representative of your perspective?"
 - ☐ A is much more representative
 - ☐ A is slightly more representative
 - ☐ Both are about the same
 - ☐ B is slightly more representative
 - ☐ B is much more representative
- Question 2:** "Which summary is more informative?"
 - ☐ A is much more informative
 - ☐ A is slightly more informative
 - ☐ Both are about the same
 - ☐ B is slightly more informative
 - ☐ B is much more informative
- Question 3:** "Which summary presents a more neutral and balanced view of the issue?"
 - ☐ A is much more neutral and balanced
 - ☐ A is slightly more neutral and balanced
 - ☐ Both are about the same
 - ☐ B is slightly more neutral and balanced
 - ☐ B is much more neutral and balanced
- Question 4:** "Which summary would you prefer of being used by the policy makers to make decisions relevant to the issue?"
 - ☐ A is much more preferred
 - ☐ A is slightly more preferred
 - ☐ Both are about the same
 - ☐ B is slightly more preferred
 - ☐ B is much more preferred

(a) Rating task with Likert-scale options. (b) Comparison task with five ordinal outcomes.
 Figure 15: Screenshots of the annotation interfaces: (a) rating task and (b) comparison task.

We then average the four dimensions and give equal weight to every input size:

$$\bar{y}_{\mathcal{M},q} = \frac{1}{4|\mathcal{N}|} \sum_{n \in \mathcal{N}} \sum_{d \in \{\text{rep}, \text{inf}, \text{neu}, \text{pol}\}} \bar{y}_{\mathcal{M},q,n}^{(d)}.$$

The relative preference score centers this question-level mean by the same model’s mean across questions:

$$\text{Diff}(\mathcal{M}, q) = \bar{y}_{\mathcal{M},q} - \frac{1}{|\mathcal{Q}|} \sum_{q' \in \mathcal{Q}} \bar{y}_{\mathcal{M},q'}.$$

where positive values indicate that model $\mathcal{M}$ performs above its own cross-question average on q , and negative values indicate below-average performance.

K Details of Judge Model Training

K.1 Multi-Output Regression Model Training

We implement a multi-output regression model to predict quality ratings across the four dimensions: representativeness, informativeness, neutrality, and policy-use acceptability (labeled *Policy Approval* in the annotations). Our approach employs several optimization strategies to improve correlation between predicted and ground-truth scores.

K.1.1 Model Architecture

The model architecture consists of a pre-trained transformer encoder (DeBERTa-v3-base (He et al., 2020)) followed by a task-specific regression head. The encoder processes the concatenated input sequence containing the question, annotator opinion, and summary text, separated by special [SEP] tokens with descriptive prefixes formatted as “Question: q_i [SEP] Annotator opinion: o [SEP] Summary: $S_{\mathcal{M}, \mathcal{O}_{i,n}}$ [SEP]”. We extract the [CLS] token representation from the final encoder layer and pass it through a hidden layer that reduces dimensionality by half with GELU activation, followed by dropout regularization. The regression head maps the transformed features to four outputs corresponding to representativeness, informativeness, neutrality, and policy-use acceptability. To constrain predictions to a valid range, we apply a sigmoid activation function that produces outputs in the $[0, 1]$ range. Ground-truth scores are normalized from their original range $[-1, 7]$ to $[0, 1]$ using min-max normalization ($y = \frac{y^{\text{raw}} - (-1)}{7 - (-1)}$) to improve training stability and convergence.

K.1.2 Training Parameters

To maximize correlation performance, we incorporate several advanced training techniques and carefully tuned hyperparameters.

Loss Function and Regularization: We employ Huber loss with $\delta = 1.0$. Because both predictions and normalized labels lie in $[0, 1]$, residuals remain in its quadratic region; in this configuration, the loss is equivalent to a scaled MSE rather than an outlier-robust regime. The model includes dropout regularization (rate = 0.1) applied to both the intermediate hidden layer and after the [CLS] token extraction. We apply gradient clipping with an ℓ_2 norm of 1.0 for training stability.

Optimization and Learning Rate Scheduling: We use the AdamW optimizer with a linear learning rate scheduler that includes warmup (ratio = 0.15) and weight decay (0.01). The initial learning rate is set to 4×10^{-5} with standard Adam hyperparameters ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$).

Training Configuration: The complete set of training hyperparameters is as follows: maximum sequence length = 4,096 tokens, training batch size = 8, evaluation batch size = 8, number of epochs = 30, and warmup ratio = 0.15. We use mixed-precision training (FP16) with gradient accumulation steps = 2 for an effective batch size of 16.

Alpaca-Format Supervised Fine-Tuning Data

Instruction: We have made a deliberation with many annotators on the issue...

One annotator’s opinion: ...

Below is a summary: [summary content]

Please evaluate this summary on 4 criteria:

1. Representativeness...
2. Informativeness...
3. Neutrality...
4. Policy Approval...

Output:

```
{
  "perspective_representation": 5,
  "informativeness": 5,
  "neutrality_balance": 5,
  "policy_approval": 5
}
```

K.2 Comparison of Base Models

We train additional encoder-based models and LLMs at different scales for comparison. For encoder-based models, DeBERTa-v3-large tests whether greater encoder capacity improves judgment, while Longformer-base-4096 (Beltagy et al., 2020) tests an architecture designed for longer sequences; both use the same dataset with hyperparameter sweeps. For LLMs, we implement two training approaches. The first adapts Qwen3-0.6B and Qwen3-4B to multi-output regression by adding a custom regression head to the pre-trained causal language model. The second constructs an Alpaca-format supervised fine-tuning (SFT) dataset (Ding et al., 2023) and fine-tunes Qwen3-4B to generate structured JSON scores, treating evaluation as text generation rather than regression.

Table 9 compares the tested models and training strategies. DeBERTa-v3-base with regression training attains the highest correlations across all four dimensions. DeBERTa-v3-large performs worse in this setup despite its greater capacity. Its validation loss remains nearly unchanged while validation Spearman correlation collapses toward zero and prediction variance becomes small, a pattern more consistent with training instability and under-dispersed predictions than standard overfitting. We therefore report DeBERTa-v3-base as the best tested configuration without claiming that larger encoders are generally inferior.

Table 9: Performance comparison across different models and training strategies. Results show Pearson and Spearman correlation coefficients for each evaluation dimension.

Model	Representativeness		Informativeness		Neutrality		Policy Approval	
	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman
DeBERTa-v3-base (Regression)	0.504	0.470	0.454	0.444	0.492	0.492	0.416	0.381
DeBERTa-v3-large (Regression)	0.159	0.162	0.221	0.203	0.129	0.125	0.174	0.162
Longformer-base-4096 (Regression)	0.097	0.098	0.209	0.219	0.158	0.156	0.127	0.125
Qwen3-0.6B (Regression)	0.125	0.136	0.231	0.249	0.210	0.205	0.196	0.186
Qwen3-4B (Regression)	0.191	0.197	0.215	0.218	0.215	0.207	0.189	0.188
Qwen3-4B (SFT)	0.338	0.289	0.153	0.157	0.211	0.188	0.289	0.244

Beyond the supervised model comparison, our few-shot evaluation with frontier LLMs further highlights the strength of DELIBERATIONJUDGE. As shown in Table 10, DELIBERATIONJUDGE surpasses all few-shot prompting baselines across every evaluation dimension. Even with 5-shot demonstrations, GPT-5, GPT-5-Mini, and Claude-3.7-Sonnet fall short of its correlations.

Table 10: Performance comparison between DELIBERATIONJUDGE and few-shot LLM baselines.

Model	Representativeness		Informativeness		Neutrality		Policy Approval	
	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman
DELIBERATIONJUDGE	0.504	0.470	0.454	0.444	0.492	0.492	0.416	0.381
GPT-5 3-shot	0.398	0.321	0.143	0.155	0.245	0.214	0.360	0.306
GPT-5 5-shot	0.422	0.372	0.116	0.131	0.226	0.205	0.319	0.269
GPT-5-Mini 3-shot	0.399	0.347	0.094	0.095	0.241	0.221	0.344	0.293
GPT-5-Mini 5-shot	0.399	0.336	0.103	0.104	0.210	0.187	0.305	0.268
Claude-3.7-Sonnet 3-shot	0.372	0.302	0.145	0.157	0.320	0.278	0.353	0.304
Claude-3.7-Sonnet 5-shot	0.365	0.309	0.164	0.176	0.309	0.271	0.348	0.307

K.3 Out-of-Distribution Test

To evaluate the robustness of DELIBERATIONJUDGE under out-of-distribution (OOD) conditions, we introduce two new deliberation questions that are not included in the original ten topics:

- **School Cellphone Use (Binary):** *“Do you support banning students from using personal cellphones during school hours?”*
- **Workplace Flexibility (Open-Ended):** *“What are your thoughts on the rise of flexible work arrangements, such as remote or hybrid work, and how should workplaces adapt?”*

For each question, we repeat the full Stage 1–3 pipeline. Specifically, we collect 100 new opinions for summary generation and recruit an additional 100 annotators to provide rating and comparison judgments for each question.

We then apply DELIBERATIONJUDGE to these two OOD test sets and compute Spearman correlations with human judgments using the same criteria as in the main benchmark. The correlations remain positive but decrease from 0.38–0.49 in-domain to 0.05–0.23 OOD (Table 11). On these two topics, DELIBERATIONJUDGE therefore retains limited alignment but does not match its in-domain reliability. This result motivates broader training data and topic-specific validation before deployment in a new deliberation domain.

Table 11: Spearman correlations of DELIBERATIONJUDGE on out-of-distribution datasets.

Question	Representativeness	Informativeness	Neutrality	Policy Approval
In-Domain	0.47	0.44	0.49	0.38
OOD (Two Topics)	0.20	0.05	0.14	0.23

We further test whether a higher-capacity judge closes the transfer gap by fine-tuning GPT-4.1 on $\mathcal{T}_{\text{train}}$ from the ten in-domain topics using the same supervision. We evaluate the resulting model on the in-domain test split and the two OOD questions. As shown in Table 12, its correlations decrease on every dimension, from 0.12–0.29 in-domain to −0.004–0.16 OOD. Thus greater model capacity alone does not remove the transfer gap in this experiment; the result does not establish failure on all unseen topics.

Table 12: Spearman correlations of GPT-4.1 fine-tuned on the in-domain topics and evaluated on the in-domain test split and two OOD topics.

Test Set	Representativeness	Informativeness	Neutrality	Policy Approval
In-Domain	0.29	0.15	0.12	0.18
OOD (Two Topics)	0.16	−0.004	0.02	0.06

L Experimental Details

L.1 Prompts

LLM Summarization Prompt

User: In each line, I provide you with human comments for a deliberation question {question}. At the end, generate an overall summary of the comments. Please do not mention the total number of comments or participants. If you need to provide statistical information, use percentages instead of absolute numbers. Here are the comments: {comments}

M Details of Human Data Collection

Human-data collection used two annotation platforms, POTATO (Pei et al., 2022) and Deliberation.io (Pei et al., 2024), with participant recruitment through Prolific². The following screenshots show the interfaces used for opinion collection and summary evaluation.

M.1 Public Opinion Data Collection

For the ten questions in Table 8, we launched ten separate Prolific studies, one per question, and recruited 300 U.S.-based participants for each study using demographic sampling targets. Figures 16–19 illustrate the opinion-submission workflow for the *Tipping System* study. Figure 20 shows the Deliberation.io interface used to manage the ten studies.

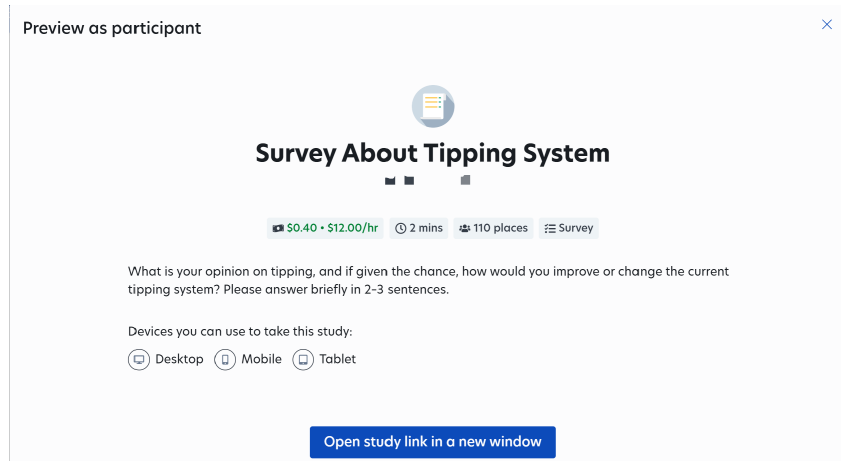


Figure 16: Example Prolific recruitment page for public-opinion collection.

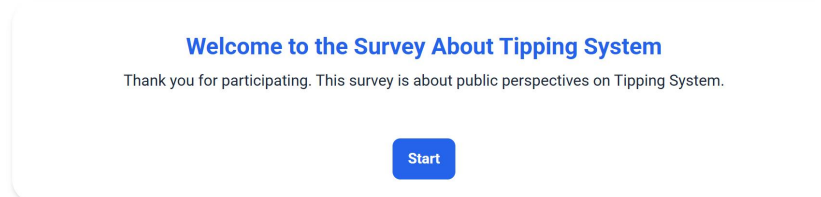


Figure 17: Study introduction page in Deliberation.io.

²<https://www.prolific.com/>

Openqa-Tipping System

Page 1

1 module

What is your opinion on tipping, and if given the chance, how would you improve or change the current tipping system? Please answer briefly in 2–3 sentences. *

Enter your response...

0 characters (min: 50) (max: 500)

0 characters

Max: 500

Back

Next

Figure 18: Opinion-collection question page in Deliberation.io.

Thank You for Your Participation

Your responses have been recorded successfully. Thank you for taking part in this survey.

Return to Prolific

Figure 19: Study completion page in Deliberation.io.

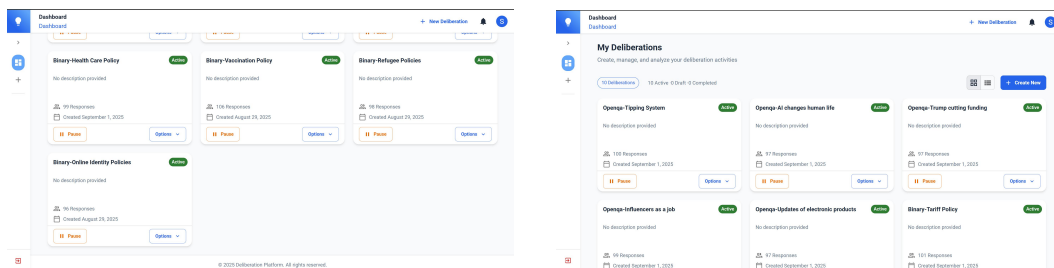


Figure 20: Example of the public opinion collection management interface on Deliberation.io.

M.2 Human-Judgment Data Collection

To collect human judgments, we implemented a structured annotation pipeline in POTATO (Pei et al., 2022) and recruited through Prolific from a pool distinct from Stage 1. Each annotator viewed the study information and instructions, wrote an opinion, rated one model-generated summary on four dimensions, compared it with a second same-question summary, and submitted the task. Figures 21–26 show this workflow.

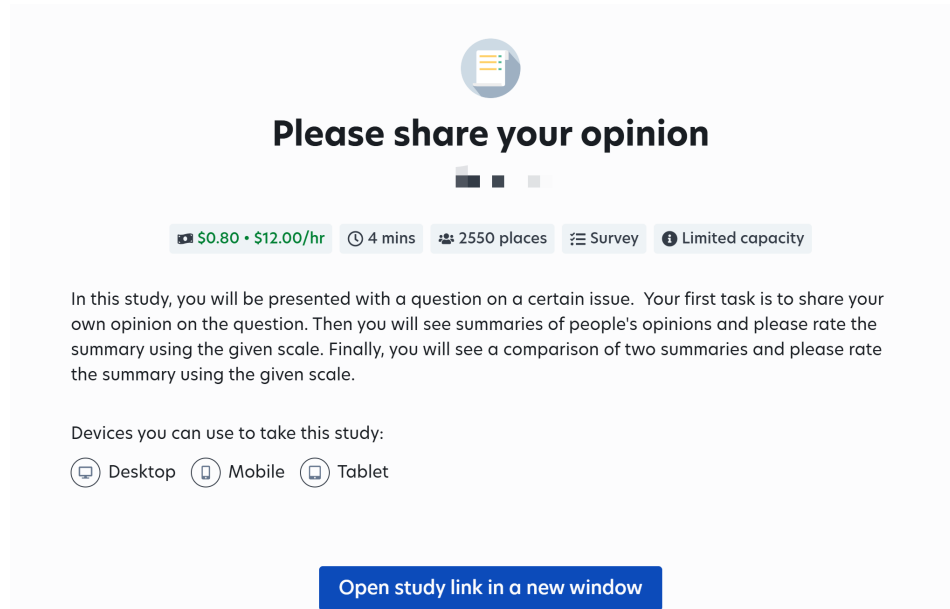


Figure 21: Recruitment page for human-judgment annotation (via Prolific).

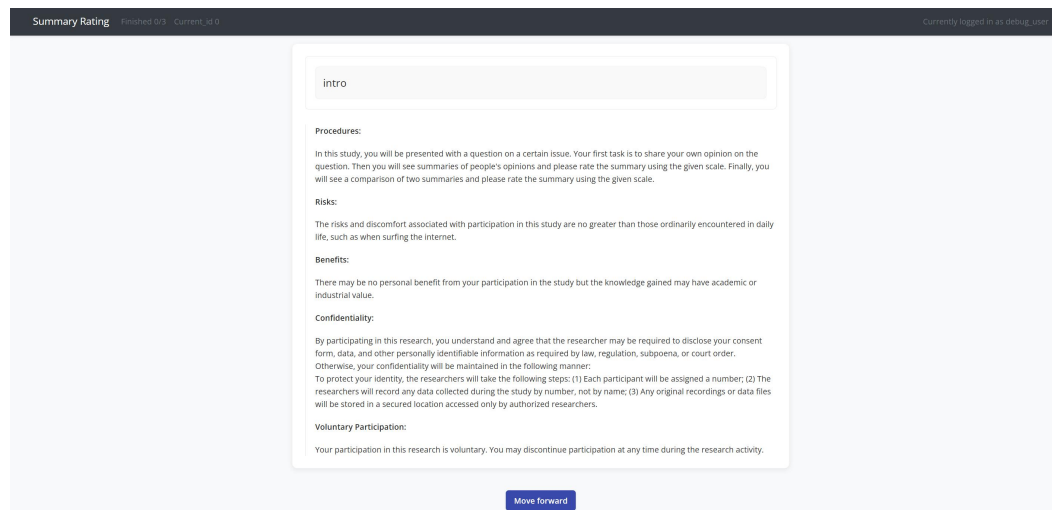


Figure 22: Task introduction page in POTATO.

Summary Rating

Finished 0/3

Current_id 1

Currently logged in as debug_user

[Question]

Do you support requiring real-name registration on social media platforms, where users must register and post under their real identity?

Please share your opinion on the question in 2-3 sentences.

Move backward

Move forward

Copyright © 2022 Blablablab

Figure 23: Interface for writing an opinion on the assigned deliberation question.

Summary Rating

Finished 1/3

Current_id 2

Currently logged in as debug_user

Below is a summary of people's opinions on the issue.

Based on the comments provided, the summary of perspectives on the rapid update cycle for smartphones is as follows:

A significant portion of commentators express frustration with the rapid update cycle, viewing it as a form of planned obsolescence that pressures consumers into frequent, costly upgrades. Many feel that yearly updates offer minimal noticeable improvements, making new models feel like minor iterations rather than significant advancements. This cycle is also criticized for contributing to electronic waste and unsustainable consumption, with concerns about environmental damage and the financial strain it places on users. Some commentators shared personal experiences of devices becoming obsolete or malfunctioning shortly after payment plans ended, reinforcing a perception that companies prioritize profits over product longevity and consumer needs.

Conversely, a portion of commentators acknowledge benefits to the rapid updates, citing improved security, better performance, and access to innovative features as positive outcomes. They appreciate how this pace drives technological progress and integrates advanced technology into daily life. However, even among those who see some advantages, many express a desire for a more balanced approach. Suggestions include extending the cycle to every 2-3 years for major releases, focusing more on software updates to extend device lifespans, and improving reparability and sustainability.

Overall, while technological advancement is valued, the prevailing sentiment calls for a more sustainable and consumer-friendly innovation rhythm that balances progress with product longevity, reduced environmental impact, and less pressure to upgrade.

To what extent is your perspective represented in this response?

☒ Not represented — summary ignores comment
 ☐ Slightly represented — little relevant content
 ☐ Partially represented — main idea but gaps
 ☐ Mostly represented — most info, minor nuance missing
 ☐ Fully represented — all info included

How informative is this summary?

☐ Not informative — almost no content
 ☐ Slightly informative — few basic points
 ☐ Moderately informative — some important details
 ☐ Very informative — most details included
 ☐ Extremely informative — highly detailed and comprehensive

Do you think this summary presents a neutral and balanced view of the issue?

☐ Very non-neutral — strongly one-sided
 ☐ Somewhat non-neutral — noticeable tilt
 ☐ Moderately neutral — mix of neutrality and tilt
 ☐ Fairly neutral — mostly impartial
 ☐ Completely neutral — fully impartial

Would you approve of this summary being used by the policy makers to make decisions relevant to the issue?

☐ Strongly disapprove
 ☐ Disapprove
 ☐ Neutral
 ☐ Approve
 ☐ Strongly approve

Figure 24: Interface for the four-dimension rating task.

Summary Rating

Finished 2/3

Current_id 3

Currently logged in as debug_user

Two summaries of opinions are shown below. Read carefully and answer according to your prior opinion. Both are scrollable.

Summary A

Based on the comments provided, the summary of perspectives on the rapid update cycle for smartphones is as follows:

A significant portion of commentators express frustration with the rapid update cycle, viewing it as a form of planned obsolescence that pressures consumers into frequent, costly upgrades. Many feel that yearly updates offer minimal noticeable improvements, making new models feel like minor iterations rather than significant advancements. This cycle is also criticized for contributing to electronic waste and unsustainable consumption, with concerns about environmental damage and the financial strain it places on users. Some commentators shared personal experiences of devices becoming obsolete or malfunctioning shortly after payment plans ended, reinforcing a perception that companies prioritize profits over product longevity and consumer needs.

Conversely, a portion of commentators acknowledge benefits to the rapid updates, citing improved security, better performance, and access to innovative features as positive outcomes. They appreciate how this pace drives technological progress and integrates advanced technology into daily life. However, even among those who see some advantages, many express a desire for a more balanced approach.

Summary B

The comments reflect a complex mix of appreciation for technological innovation and significant concerns about the rapid update cycle of smartphones. Many express that frequent updates provide improved performance, better security, and innovative features, which are seen as necessary in today's fast-paced world. However, a majority voice strong criticisms, highlighting that the cycle encourages planned obsolescence, places financial pressure on consumers to upgrade frequently, and contributes substantially to electronic waste.

There is a widespread perception that manufacturers prioritize releasing new products over perfecting existing ones, often with only incremental changes that do not justify the high costs or the environmental impact. Concerns about sustainability, repairability, and the creation of unnecessary electronic waste are prominent. Many suggest that a slower, more deliberate update cycle—such as every two to three years—would be more beneficial, allowing for more meaningful innovations and reducing consumer and environmental strain. Overall, while technological progress is valued, there is a strong desire for a more balanced and responsible approach to innovation.

Which summary is more representative of your perspective?

☐ A is much more representative
☐ A is slightly more representative
☐ Both are about the same
☐ B is slightly more representative
☐ B is much more representative

Which summary is more informative?

☐ A is much more informative
☐ A is slightly more informative
☐ Both are about the same
☐ B is slightly more informative
☐ B is much more informative

Which summary presents a more neutral and balanced view of the issue?

☐ A is much more neutral and balanced
☐ A is slightly more neutral and balanced
☐ Both are about the same
☐ B is slightly more neutral and balanced
☐ B is much more neutral and balanced

Which summary would you prefer of being used by the policy makers to make decisions relevant to the issue?

☐ A is much more preferred
☐ A is slightly more preferred
☐ Both are about the same
☐ B is slightly more preferred
☐ B is much more preferred

Figure 25: Interface for the summary-comparison task.

Summary Rating

Finished 3/3

Current_id 4

Currently logged in as debug_user

end

Thanks for your time, please click the following link to complete the study

[Click to finish the study](#)

Move backward

Fork on GitHub | Cite Us

Figure 26: Final submission and completion page in POTATO.