CAN AI TRULY REPRESENT YOUR VOICE IN DELIBERATIONS? A COMPREHENSIVE STUDY OF LARGE-SCALE OPINION AGGREGATION WITH LLMS

Shenzhe Zhu[1,*], **Shu Yang**[2,*], **Michiel A. Bakker**[3], **Alex Pentland**[3,4], **Jiaxin Pei**[4,✉]
[1]University of Toronto [2]King Abdullah University of Science and Technology
[3]Massachusetts Institute of Technology [4]Stanford University

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 been shown as a promising tool to generate summaries for large-scale deliberations, they also risk underrepresenting minority perspectives and exhibiting bias with respect to the input order, raising fairness concerns in high-stakes contexts. Studying and fixing these issues requires a comprehensive evaluation at a large scale, yet current practice often relies on LLMs as judges, which show weak alignment with human judgments. To address this, we present DELIBERATIONBANK, a large-scale human-grounded dataset with (1) opinion data spanning ten deliberation questions created by 3,000 participants and (2) summary judgment data annotated by 4,500 participants across four dimensions (representativeness, informativeness, neutrality, policy approval). Using these datasets, we train DELIBERATIONJUDGE, a fine-tuned DeBERTa model that can rate deliberation summaries from individual perspectives. DELIBERATIONJUDGE is more efficient and more aligned with human judgements compared to a wide range of LLM judges. With DELIBERATIONJUDGE, we evaluate 18 LLMs and reveal persistent weaknesses in deliberation summarization, especially underrepresentation of minority positions. Our framework provides a scalable and reliable way to evaluate deliberation summarization, helping ensure AI systems are more representative and equitable for policymaking. All the code, data, and model will be released at https://github.com/shuyhere/LLMs-Scalable-Deliberation

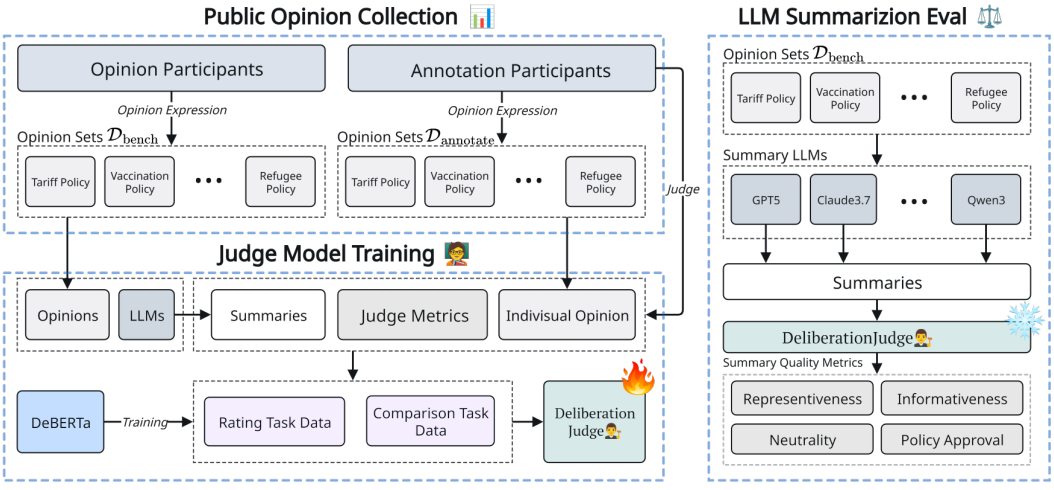


Figure 1: Overview of our benchmark framework(see §2), including opinion collection (§2.2), judge model training (§3), and LLM-based deliberation summarization evaluation (§2.1).

* Equal contribution with randomized author order. Shenzhe contributed writing; Shu contributed experiments. ✉ Correspondence: pedropei@stanford.edu.

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. In citizens' assemblies, facilitators must continuously synthesize perspectives during discussions and later produce comprehensive summaries for participants, stakeholders, policymakers, and the broader public (Landemore, 2020). In parliamentary settings, members and their staff must maintain up-to-date representations of a complex, multi-stakeholder environment (Brandsma & Otjes, 2024). 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. 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 over-represent majority views while neglecting minority perspectives (Small et al., 2023; Li et al., 2024); 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 central 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 question, we create a framework for automating the evaluation of deliberation summarization at a large scale. As shown in Figure 2, at the first step, we design 10 deliberation questions spanning technology, social media, and public policy, and recruit 300 U.S.-based participants per question to provide free-form opinions (3,000 in total). We then generate summaries from these opinions using multiple LLMs under a multi-scale setting that varies the number of inputs. At the second step, to enable reliable large-scale evaluation, we recruit an additional 4,500 annotators (450 per question); each submits a personal opinion and evaluates a corresponding summary along four deliberation-relevant dimensions: *Representativeness*, *Informativeness*, *Neutrality*, and *Policy Approval*. This process results in 3,000 collected opinions and 4,500 evaluation instances, which together constitute our large-scale human-grounded dataset, DELIBERATIONBANK. Based on this dataset, we systematically evaluate LLM-as-judge approaches and develop DELIBERATIONJUDGE, a fine-tuned DeBERTa-based (He et al., 2020) model for personalized summarization evaluation. Compared to general-purpose LLMs, DELIBERATIONJUDGE achieves closer alignment with human judgments and offers up to 100× greater efficiency.

Using DELIBERATIONJUDGE, we benchmark the summarization performance of 18 LLMs using the 3,000 opinions collected in the first stage. We also conduct a systematic study to examine how factors such as model choice, input configuration, and evaluation method influence the effectiveness of deliberation support through summarization. Our analysis reveals three critical insights. First, off-the-shelf LLM judges exhibit only limited correlation with human judgments across dimensions, with agreement varying by model size. Second, LLM-generated summaries tend to under-represent minority stances and show sensitivity to input ordering and verbosity, indicating systematic biases that are particularly problematic in deliberative contexts. Third, a lightweight, preference-aligned DeBERTa judge improves reliability and enables large-scale, cost-effective benchmarking of LLM summarizers for deliberation tasks. Our core contributions are as follows:

- We develop DELIBERATIONBANK, a large-scale deliberation benchmark dataset created by 7,500 participants from a US representative sample with two subsets: (i) a public opinion dataset of 3,000 free-form opinions collected from 10 societal deliberation questions on trending topics, and (ii) a summary judgement dataset of 4,500 annotations that evaluate deliberation summaries from individual perspectives.
- We develop an automated evaluation framework to access LLMs' ability to summarize large-scale deliberations and propose DELIBERATIONJUDGE, a fine-tuned DeBERTa model that can judge deliberation summaries from individual perspectives. DELIBERATIONJUDGE outperforms other LLM judge baselines on both alignment with humans and inference efficiency.

- We conduct a rigorous and comprehensive study of LLM summarizers that surfaces systematic biases (e.g., minority-stance under-coverage, order/verbosity sensitivity) and provides actionable guidance for fairer summarization in public decision-making.

2 Dataset and Evaluation Setting

2.1 Overall evaluation pipeline

The overall goal is to evaluate LLM's capabilities to generate representative and effective summaries for public deliberations. We create four deliberation-relevant metrics, grounded in prior work on summarization and societal deliberation (Small et al., 2023; Vijay et al., 2024; Lee et al., 2022). *Representativeness:* Degree to which an individual opinion is semantically captured by the summary, ranging from *not represented at all* to *fully represented*; *Informativeness:* How well the summary conveys diverse, detailed, non-redundant information; *Neutrality:* The summary avoids bias and subjective language, maintaining a balanced stance. *Policy Approval:* Perceived suitability of the summary for use by policymakers in decision-making.

To assess the ability of LLMs to generate high-quality deliberation summaries, we adopt an automatic evaluation framework. Formally, for each deliberation question q, let $\mathcal{O} = \{o_1, \ldots, o_n\}$ denote the set of collected opinions, and let $\tilde{\mathcal{O}} \subseteq \mathcal{O}$ be the subset provided to a summarization model $\mathcal{M}$. The model then produces a summary

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

To assess quality, each summary $S_{\mathcal{M},\tilde{\mathcal{O}}}$ is paired with an individual opinion $o_i \in \tilde{\mathcal{O}}$ and evaluated by DeliberationJudge, our DeBERTa-based model (He et al., 2020) fine-tuned on human-annotated judgments. The judge produces a four-dimensional continuous score vector in $[0,1]^4$:

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

Higher values indicate stronger performance on each criterion.

2.2 DeliberationBank

The public opinion dataset $\mathcal{D}$ is constructed in three stages.

Stage 1: Collect Initial Opinions. We first construct a deliberation question set $\mathcal{Q} = \{q_1, \ldots, q_{10}\}$ covering trending societal topics in technology, social media, and public policy (Appendix B.1). The questions fall into two categories: *Open-Ended*, which admit free-form responses without restricting scope, and *Binary*, which explicitly frame opinions around two opposing stances (e.g., support vs. oppose). For each $q_i \in \mathcal{Q}$, we recruit 300 U.S.-based participants from a **representative sample** to provide opinions. Let P_i denote the participant set for q_i. To ensure independence across questions, participant pools are disjoint:

$$P_i \cap P_j = \varnothing \quad \text{for } i \neq j.$$

For each q_i, the opinion set is $\mathcal{O}_i = \{o_i^{(1)}, \ldots, o_i^{(300)}\}$, yielding $\mathcal{D}_{\text{bench}} = \bigcup_{i=1}^{10} \mathcal{O}_i, |\mathcal{D}_{\text{bench}}| = 3{,}000.$.

Stage 2: Summary Generation for Human Judgment. We generate summaries using five LLMs of different scales and architectures: `GPT-5`, `Claude-4 Opus`, `Gemini-2.5 Pro`, `DeepSeek-V3.1`, and `Qwen3-32B`. For each $q_i \in \mathcal{Q}$, we apply a multi-scale summarization strategy on its opinion set $\mathcal{O}_i$: the model conditions on subsets $\tilde{\mathcal{O}}_{i,N} \subseteq \mathcal{O}_i$ of varying sizes N (e.g., $N \in \{10, 20, 30, \ldots\}$). Given a fixed subset $\tilde{\mathcal{O}}_{i,N}$, we generate $K = 3$ summaries by independently running the model K times:

$$S^{(k)}_{\mathcal{M},\tilde{\mathcal{O}}_{i,N}} = \mathcal{M}(q_i, \tilde{\mathcal{O}}_{i,N}), \quad k = 1, 2, 3.$$

Overall, this procedure yields 750 summaries across all models, topics, input scales, and resamplings.

Stage 3: Human Judge Annotation. In Stage 3, we collect human judgment for LLM-generated summaries, using a new pool of participants distinct from those in Stage 1. We use POTATO (Pei et al., 2022) to set up the annotation pipeline.

For each generated summary $S_{\mathcal{M},\tilde{\mathcal{O}}_i^{(k)}}$, we assign six independent annotators $\mathcal{A}_S = \{a_1, \ldots, a_6\}$. Each annotator $a \in \mathcal{A}_S$ first provides their own opinion $o^{(a)}$ on the underlying question q_i. Next,

they evaluate S in the **rating task**, producing a score vector $r^{(a)}(q_i, S) \in \{1, \ldots, 5\}^4$, where dimensions correspond to *Representativeness*, *Informativeness*, *Neutrality*, and *Policy Approval*. They also participate in the **comparison task**, where S is paired with another summary $S' \in \mathcal{S}_{q_i} \setminus \{S\}$ (via ring-based matching; Appendix C.1), yielding $c^{(a)}(q_i, S, S') \in \{1, \ldots, 5\}^4$. Together, these steps define the annotation mapping at the level of a single annotator:

$$\Phi : (q_i, o^{(a)}, S, S') \mapsto \big(r(q_i, S),\, c(q_i, S, S')\big) \tag{1}$$

For a fixed summary S, the annotated instances are

$$\mathcal{T}_S = \{\Phi(q_i, o^{(a)}, S, S') : a \in \mathcal{A}_S,\ S' \in \mathcal{S}_{q_i} \setminus \{S\}\}.$$

Aggregating over all 750 summaries, we obtain the full dataset $\mathcal{T}_{\text{annotate}} = \bigcup_S \mathcal{T}_S, |\mathcal{T}_{\text{annotate}}| = 4{,}500$.

As in all stages involving human participants, data quality checks are applied to filter unreliable responses (e.g., abnormally short completion times; Appendix B.3). These complementary tasks capture both the absolute quality of individual summaries and the relative preference between alternatives, providing rich supervision for judge training.

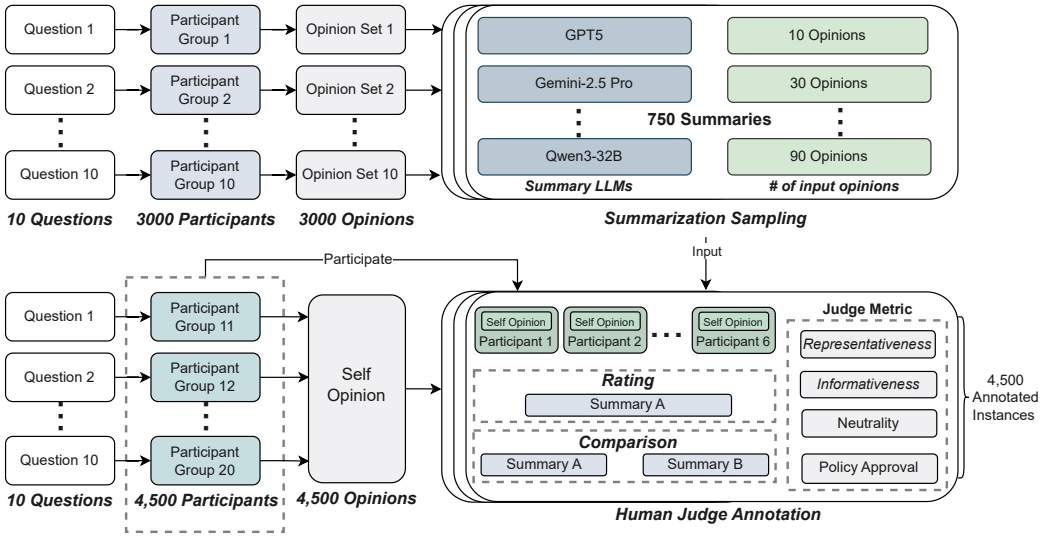


Figure 2: Human Judge Annotation pipeline

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 opinions set $\mathcal{D}_{\text{train}}$) and $\mathcal{T}_{\text{test}}$ ($|\mathcal{T}_{\text{test}}| = 900$ with opinions set $\mathcal{D}_{\text{test}}$). Table 1 summarizes the dataset statistics across all three stages.

Table 1: 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 Judgement of LLM Summaries with DeliberationJudge

While recent work has increasingly adopted pretrained LLMs as automated judges, prior studies (Huang et al., 2024; Ye et al., 2024; Thakur et al., 2024; Krumdick et al., 2025; Yang et al., 2025) reveal their limitations, including systematic biases and instability. These shortcomings largely stem from their black-box nature, with billions of internal parameters, and from the constraints of current alignment paradigms. To balance efficiency and reliability, 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 consistency between human and LLM judgments, we take the raw inputs of $\mathcal{T}_{\text{test}}$, each quadruple (q_i, o, S, S'), and ask LLMs to annotate them under the same guidelines as humans (see Equation 1 and instruction in Table 6). This process yielding the set $\hat{\mathcal{T}}_{\text{test}}$ comparable to $\mathcal{T}_{\text{test}}$ for correlation analysis.

After annotation, we compute correlations between LLM and human judgments. As shown in Figure 3, larger models achieve higher agreement with human judges, while very small 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 and motivate the need for more reliable, dedicated judge models.

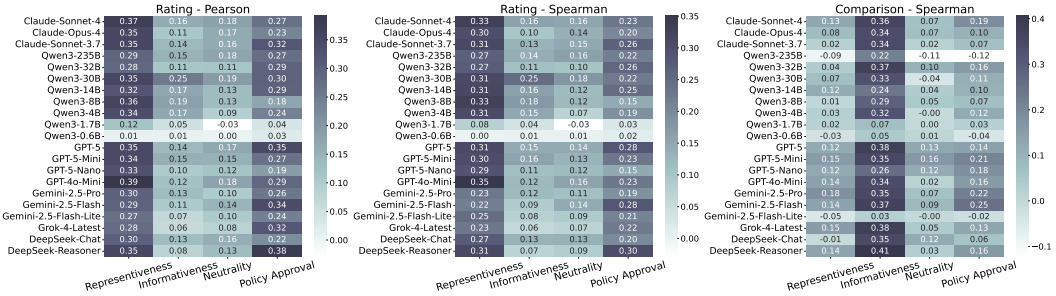


Figure 3: Heatmap of consistency between human and LLM judges on two judgment tasks. **Left**: Rating–Pearson r; Center: Rating–Spearman ρ; Right: Comparison–Spearman ρ. Lighter to darker colors represent lower to higher correlations.

3.2 DeliberationJudge

Considering the lack of consistency between LLM as Judge and human annotation, we propose DeliberationJudge, a DeBERTa-based judge $\mathcal{J}_\theta$, trained with supervised fine-tuning (SFT) on the two types of human annotations introduced in § 2.2. To place both rating and comparison tasks on a unified scale, we normalize the labels for each evaluation dimension. Specifically, five-point Likert scores from the rating task (Summary A alone) are retained on $[1, 5]$, while five-point relative judgments from the comparison task (Summary A vs. Summary B) are linearly mapped into the extended interval $[-1, 7]$. This normalization allows both annotation sources to be trained jointly as graded supervision on the same scale. Formally, given an input consisting of a deliberation question q_i, an annotator opinion $o_i^{(j)}$, and a candidate summary $S_{\mathcal{M},\tilde{\mathcal{O}}_i}$ produced by model $\mathcal{M}$ from opinion subset $\tilde{\mathcal{O}}_i$, the judge $\mathcal{J}_\theta$ encodes the concatenated sequence

$$\texttt{[CLS]}\ q_i\ \texttt{[SEP]}\ o_i^{(j)}\ \texttt{[SEP]}\ S_{\mathcal{M},\tilde{\mathcal{O}}_i}\ \texttt{[SEP]}$$

and outputs a four-dimensional normalized score vector:

$$\hat{\mathbf{y}} = \mathcal{J}_\theta(q_i, o_i^{(j)}, S_{\mathcal{M},\tilde{\mathcal{O}}_i}) = \left(\hat{y}^{(\text{rep})}, \hat{y}^{(\text{inf})}, \hat{y}^{(\text{neu})}, \hat{y}^{(\text{pol})}\right) \in [0,1]^4. \tag{2}$$

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:

$$\mathcal{L}(\theta) = \frac{1}{|\mathcal{T}_{\text{train}}|} \sum_{(q,o,S)\in\mathcal{T}_{\text{train}}} \frac{1}{4} \sum_{d\in\{\text{rep},\text{inf},\text{neu},\text{pol}\}} \ell_\delta\left(\hat{y}^{(d)}, y^{(d)}\right), \tag{3}$$

where

$$\ell_\delta(\hat{y}, y) = \begin{cases} \frac{1}{2}(\hat{y} - y)^2 & \text{if } |\hat{y} - y| \le \delta, \\ \delta \cdot \left(|\hat{y} - y| - \frac{1}{2}\delta\right) & \text{otherwise.} \end{cases} \tag{4}$$

At inference time, predictions remain in the $[0, 1]$ range and are used directly as summary scores. Detailed training settings are provided in Appendix E.1. We also compared multiple base models and found DeBERTa to perform best, full comparison results are reported in Appendix E.2.

3.3 DELIBERATIONJUDGE OUTPERFORMS SOTA LLMS IN CONSISTENCY AND EFFICIENCY

We evaluate the agreement between human judges and DELIBERATIONJUDGE, and compare it against Human–LLM judgment correlations as defined in § 3.1. Specifically, we provide DELIBERATIONJUDGE with the same held-out test set $\mathcal{T}_{\text{test}}$ that was used for LLM-based annotation. While $\mathcal{T}_{\text{test}}$ is defined over quadruples $(q_i, o_i^{(j)}, S, S')$, DELIBERATIONJUDGE takes as input a pair $(q_i, o_i^{(j)}, S)$. Therefore, each quadruple is converted into two regression instances, $(q_i, o_i^{(j)}, S)$ and $(q_i, o_i^{(j)}, S')$, with outputs normalized to $[0, 1]$. Since $\mathcal{T}_{\text{test}}$ contains 900 quadruples, this conversion yields 1,800 (question, opinion, summary) triples for DELIBERATIONJUDGE, forming a new annotated set $\hat{\mathcal{T}}_{\text{test}}^{\text{Delib}}$ for subsequent correlation analysis.

We then examine the average Pearson correlation of automatic judges with human annotations, as well as their efficiency in producing judgments for each (opinion, summary) pair. From Figure 4 (left), we observe that DELIBERATIONJUDGE achieves the highest scores in both agreement with human preferences and efficiency. Its overall correlation reaches approximately 0.48, outperforming the second-best model, `Claude-3.7-Sonnet`, which achieves only around 0.20. In terms of efficiency, DELIBERATIONJUDGE requires only about 0.03 seconds per item, while most LLM judges take more than 8 seconds on average. Although `Qwen3-0.6B` approaches our method in inference speed, it performs poorly in judgment quality. From another perspective, efficiency

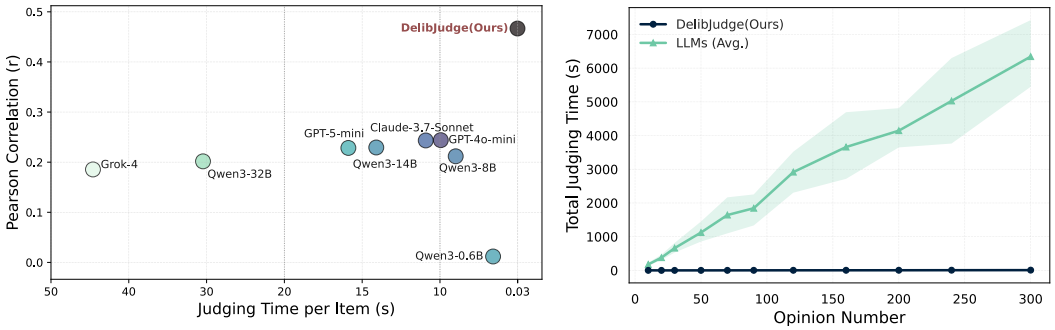


Figure 4: **Left**: Judging Time vs. Pearson Correlation. DELIBERATIONJUDGE achieves both the lowest time and the highest correlation compared to LLMs. **Right**: Scaling Stability. Total judging time of DELIBERATIONJUDGE and eight LLMs as the number of comments increases; solid lines show means and shaded areas denote min–max ranges. In the figures, DelibJudge=DELIBERATIONJUDGE.

becomes even more critical when scaling to large deliberation settings. For example, if we need to evaluate summaries conditioned on thousands of deliberation opinions, the inference time cost quickly becomes prohibitive. As shown in Figure 4 (right), we compute the total time required by DELIBERATIONJUDGE and the average of eight LLM judges as the number of input opinions increases. Since judgments must be executed once for each (opinion, summary) pair, the total cost for LLMs grows rapidly with opinion number. In contrast, the total time for DELIBERATIONJUDGE remains largely unaffected by input size, owing to its fast per-pair processing speed.

3.4 STRONG RANK CORRELATION BETWEEN DELIBERATIONJUDGE AND HUMAN JUDGMENTS

Table 2: Spearman rank correlation (human vs. DELIBERATIONJUDGE) across four dimensions.

	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 models used in Stage 2 summarization for building DELIBERATIONBANK (2.2). Concretely, for each evaluation dimension and overall average score, we compute the Spearman rank correlation across models between human scores and DELIBERATIONJUDGE predictions on $\hat{\mathcal{T}}_{\text{test}}^{\text{Delib}}$ and the original $\hat{\mathcal{T}}_{\text{test}}$. As shown in Table 2, DELIBERATIONJUDGE exhibits strong rank alignment overall (ρ = 0.70), with near-perfect agreement on Informativeness (ρ = 1.00) and high alignment on Neutrality and Policy

Approval ($\rho = 0.90$ each). Representativeness shows moderate alignment ($\rho = 0.60$), reflecting minor ordering differences in that dimension.

4 BENCHMARKING LLMS FOR DELIBERATION SUMMARIZATION

We evaluate a broad set of language models, covering both proprietary frontier systems and open-weight models across diverse parameter scales (see Table 4 in Appendix A). For each deliberation question $q_i \in \mathcal{Q}$, we sample an opinion set $\mathcal{O}_i \subseteq \mathcal{D}_{\text{bench}}$ of size $|\mathcal{O}_i| = 300$. From this pool, opinion subsets $\tilde{\mathcal{O}}_{i,N} \subseteq \mathcal{O}_i$ of size $|\tilde{\mathcal{O}}_{i,N}| = N$ are drawn, with $N \in \{10, 20, 30, 50, 70, 90, 120, 160, 200, 240, 300\}$. For each pair $(q_i, \tilde{\mathcal{O}}_{i,N})$, we perform $K = 3$ independent resamplings to obtain $\tilde{\mathcal{O}}_{i,N}^{(k)}$, and generate summaries via

$$S^{(k)}_{\mathcal{M},\,\tilde{\mathcal{O}}_{i,N}} = \mathcal{M}(q_i,\, \tilde{\mathcal{O}}_{i,N}), \quad k = 1, 2, 3.$$

Each summary $S^{(k)}_{\mathcal{M},\,\tilde{\mathcal{O}}_{i,N}}$ is then evaluated by DELIBERATIONJUDGE against every opinion in the corresponding subset. Formally, for each $o_i^{(j)} \in \tilde{\mathcal{O}}_{i,N}$, the judge maps

$$\mathcal{J}_\theta\big(q_i,\, o_i^{(j)},\, S^{(k)}_{\mathcal{M},\,\tilde{\mathcal{O}}_{i,N}}\big) = \big(\hat{y}^{(\text{rep})},\, \hat{y}^{(\text{inf})},\, \hat{y}^{(\text{neu})},\, \hat{y}^{(\text{pol})}\big) \in [0,1]^4.$$

Finally, results from the three resampled summaries are aggregated to compute 95% confidence interval (CI) for each evaluation metric. Finally, results from the three resampled summaries are aggregated to compute 95% confidence intervals (CI) for each evaluation metric. To report a global-wise average score on four dimensions for model $\mathcal{M}$, we first take the mean across the four evaluation dimensions for each summary:

$$y_{i,N}^{(k)} = \tfrac{1}{4}\Big(\hat{y}^{(\text{rep})} + \hat{y}^{(\text{inf})} + \hat{y}^{(\text{neu})} + \hat{y}^{(\text{pol})}\Big).$$

The model's global average score (GAS) is then defined as

$$GAS(\mathcal{M}) = \frac{1}{|\mathcal{Q}|}\sum_{i=1}^{|\mathcal{Q}|}\frac{1}{|N|}\sum_{N}\frac{1}{K}\sum_{k=1}^{K} y_{i,N}^{(k)}, \tag{5}$$

where $|\mathcal{Q}| = 10$ denotes the number of questions, $N \in \{10, 20, \ldots, 300\}$ the set of subset sizes, and $K = 3$ the number of resamplings. Beyond this global score, we later conduct fine-grained analyses by computing subset-wise and question-wise average scores.

Table 3: Model's global average score (GAS), expressed as $GAS \pm$ 95% CI half-width, computed across four evaluation dimensions (Representativeness, Informativeness, Neutrality, and Policy Approval) over all data samples. Models are arranged from left to right and top to bottom in descending order of mean performance.

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

4.1 MAIN RESULTS

As shown in Table 3, the `GPT-5` and `Claude-4` families lead the leaderboard, averaging 0.61–0.62 across four dimensions. The `Qwen3-32B` outperforms several closed-source models (e.g., `Gemini-2.5`, `Grok-4`), 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 gap between best and worst is < 0.1, indicating that most models capture deliberation summaries at a broadly similar quality level. 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).

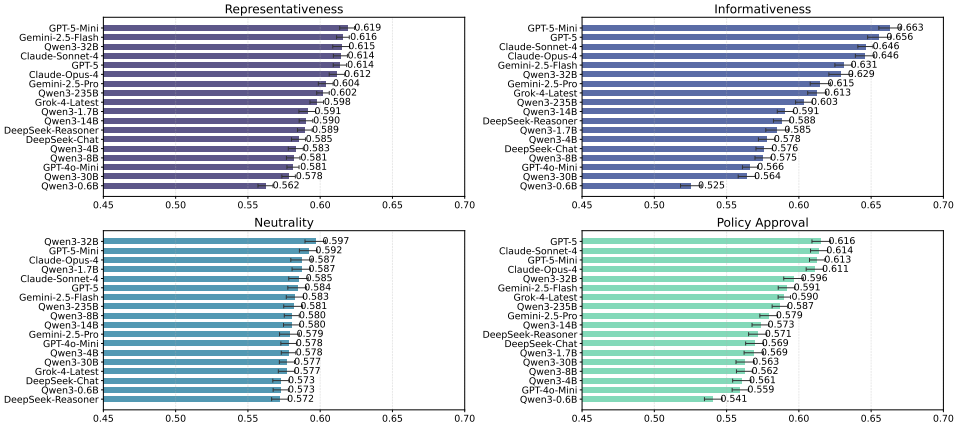


Figure 5: Comparative performance on Representativeness, Informativeness, Neutrality, and Policy Approval (mean ± 95% CI); higher is better.

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. As shown in Figure 6 (upper left), the model's average score at each opinion subset size N improves consistently as N increases from 10 to about 100. beyond which it plateaus. Thus, while LLMs benefit from moderate input sizes, they fail to leverage substantially larger sets, revealing limited scalability in handling large deliberation contexts and underscoring the need for more effective aggregation mechanisms.

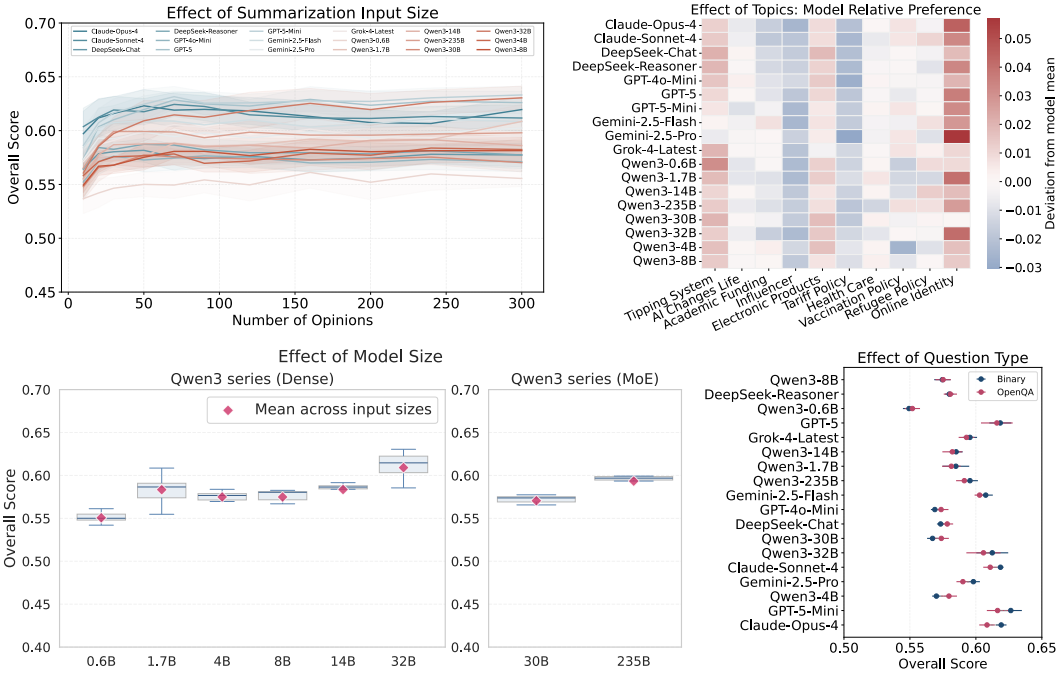


Figure 6: Overview of factors affecting model performance. **Upper Left**: Effect of input subset size with $N \in \{10, 20, \ldots, 300\}$. **Upper Right**: Topic heatmap across models, where each cell shows the topic-wise average score after centering by subtracting the model's global mean (red = above global mean; blue = below). **Lower Left**: Effect of model size, showing GAS of the `Qwen3` family, comparing Dense and MoE architectures from 0.6B to 235B. **Lower Right**: Effect of question type (Binary vs. Open-Ended), with models ranked by $|\Delta|$, the absolute difference between type-wise average scores (top = most robust).

Topics. We further analyze model sensitivity across topics by computing relative preference scores, defined as, over the full dataset, the difference between the question-level average score (for question q) and the model's global average score (pooled across all questions, subset sizes, and runs); (see Appendix D.2 for the formal definition). As shown in Figure 6 (upper right), topics such as *Online Identity* and *Tipping System* yield higher-than-average scores due to more concentrated opinions, whereas open-ended topics like *Influencers* produce dispersed views, making it harder for models to capture all perspectives. These results suggest that sensitivity depends not only on question type but also on framing, with narrower questions yielding more consistent summaries.

Question Type. Our benchmark includes ten deliberation questions, evenly split between binary and open-ended types. Figure 6 (lower right) presents a type-wise comparison of model average performance, ranking systems by $|\Delta|$, the absolute difference between average scores on Binary and Open-Ended questions; smaller $|\Delta|$ indicates lower sensitivity. Most smaller models (except `Qwen3-4B`) are consistent across types, though at low absolute scores. `GPT-5` combines high average performance with relatively small sensitivity. Overall, 11 of 18 models perform better 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.

Model Size. We next examine the effect of model size on summarization performance. Figure 6 (lower left) shows boxplots of each `Qwen3` model's GAS comparing Dense (Xiao et al., 2024) and Mixture-of-Experts (MoE) (Mu & Lin, 2025) architectures. In line with scaling law observations 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 conducted a focused case study on two representative deliberation questions: a Binary question on *Tariff Policy* and an Open-Ended question on *AI Change Life*.

Minority Data Collection. For each question, we collected 1,000 opinions from a new pool of U.S. participants. In addition to providing their stance on the question, participants were explicitly asked to self-identify whether they believed their opinion belonged to a minority group (i.e., *Do you think your opinion differs from that of most people in the U.S.?*). Three response options were provided: *Yes*, *No*, and *I'm not sure*. This self-reported annotation enables a relative ground-truth partition of the data into minority and non-minority subsets. Compared to earlier collection settings with 300 responses, we increased the sample size to 1,000 to ensure more reliable minority/non-minority estimates.

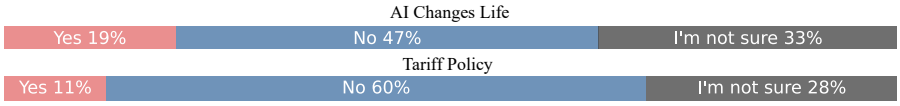


Figure 7: Distribution of minority opinions in the newly data for two representative questions.

Key Findings. We focus on the *representativeness* dimension, as it best reflects whether an individual opinion is covered by the generated summary. Based on participant self-reports, we partitioned the dataset into minority and non-minority subsets, treating only responses marked as *Yes* as minority and all others as non-minority, and then computed the average representativeness score for each model on the two groups using 1,000 opinions as summarization input. As shown in Figure 8, across all models, representativeness scores for non-minority opinions are higher than those for minority opinions, revealing a systematic bias. The effect is strongest in the binary *Tariff Policy* case (gap up to 0.08) and smaller in the open-ended question (about 0.02). We attribute this disparity to the relative scarcity of minority responses in the *Tariff Policy* setting (see the minority distribution in Figure 7), which makes models more likely to overlook them.

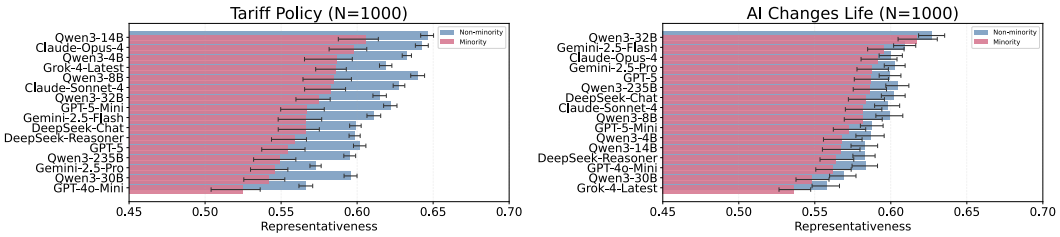


Figure 8: Comparison of model performance on representativeness for minority and non-minority opinions across two representative deliberation questions. Results are reported under summarization input sizes of 1000.

6 Conclusion

Summarizing public opinions is a key step to enabling large-scale deliberations. While existing studies have demonstrated the potential of LLMs for deliberation summarization, it remains unclear which LLM and what setting lead to the optimal summarization experience. One of the key bottlenecks of deliberation summarization evaluation is the scalability of human evaluation, as the perception of LLM summarization is highly subjective. In this paper, we present DeliberationBank, a large-scale deliberation and summarization evaluation dataset, and DeliberationJudge, a fine-tuned model that can accurately and efficiently scale the judgment of deliberation summarizations. Leveraging the dataset and model, we provide a systematic evaluation of 18 LLMs and our study reveals key insights and limitations of LLM for deliberation summarization.

Ethics Statement

This work relies on human annotation to evaluate deliberation summaries. All annotators were recruited through the Prolific platform and managed with the Potato annotation system, which ensured complete anonymization of responses and secure task assignment. Participants provided informed consent prior to annotation, and no personally identifiable information (PII) was collected or stored at any stage. Sensitive questions were included solely for research purposes and were framed in a neutral manner. Annotators were compensated at fair market rates.

Reproducibility Statement

We provide full details of training hyperparameters, model settings, and evaluation protocols in Appendix E. Upon publication, we will release the detailed dataset information, annotation guidelines, preprocessing scripts, training code, and evaluation scripts to enable independent replication and further study. Together, these resources will allow researchers to reproduce our experiments and extend our framework.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, 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.

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.

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 Krumdick, 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.

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*, 2024.

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.

Siyuan Mu and Sen Lin. A comprehensive survey of mixture-of-experts: Algorithms, theory, and applications. *arXiv preprint arXiv:2503.07137*, 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.

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, 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 arXiv:2306.11932*, 2023.

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.

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.

A Details of Model Choice

As shown below, we include 18 widely used LLMs, covering both proprietary (closed-source) and open-weight models across several major families: OpenAI GPT, Anthropic Claude, Google Gemini, Alibaba Qwen3, DeepSeek, and xAI Grok. Within the Qwen3 series, `Qwen3-30B-a3b` and `Qwen3-235B-a22b` adopt a Mixture-of-Experts (MoE) architecture, with 3B and 22B active parameters during inference, respectively.

Table 4: List of evaluated summary models, including both API-based frontier systems and open-weight models of varying scales.

Model	#Size	Form	Creator	Model	#Size	Form	Creator
`GPT-4o-Mini` (Achiam et al., 2023)	N/A	api	OpenAI	`Qwen3-0.6B` (Team, 2025)	0.6B	open	Alibaba
`GPT-5-Mini` (OpenAI, 2025)	N/A	api	OpenAI	`Qwen3-1.7B` (Team, 2025)	1.7B	open	Alibaba
`GPT-5` (OpenAI, 2025)	N/A	api	OpenAI	`Qwen3-4B` (Team, 2025)	4B	open	Alibaba
`Claude-4-Sonnet` (Anthropic, 2025)	N/A	api	Anthropic	`Qwen3-8B` (Team, 2025)	8B	open	Alibaba
`Claude-4-Opus` (Anthropic, 2025)	N/A	api	Anthropic	`Qwen3-14B` (Team, 2025)	14B	open	Alibaba
`Gemini-2.5-Flash` (Comanici et al., 2025)	N/A	api	Google	`Qwen3-30B-a3b` (Team, 2025)	30B(a3B)	open	Alibaba
`Gemini-2.5-Pro` (Comanici et al., 2025)	N/A	api	Google	`Qwen3-32B` (Team, 2025)	32B(a22B)	open	Alibaba
`DeepSeek-V3.1 (Thinking)` (Guo et al., 2025)	671B	api	DeepSeek	`Qwen3-235B-a22b` (Team, 2025)	235B	open	Alibaba
`DeepSeek-V3.1 (No-Thinking)` (Liu et al., 2024)	671B	api	DeepSeek	`Grok-4` (xAI, 2025)	N/A	api	xAI

B Details of Public Opinion Datasets

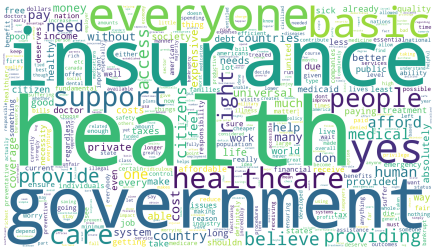
B.1 Deliberation questions.

As shown in Table 5, we design ten deliberation questions spanning two categories: five binary-choice policy questions (e.g., health care, refugee policy, tariff policy) and five open-ended societal questions (e.g., AI and human life, tipping systems, electronic products). This design ensures coverage of both structured policy debates and broader social issues, capturing diverse perspectives for summarization and evaluation.

Table 5: Deliberation questions.

Questions	Type	Description
"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?

B.2 WORLDCLOUD.



(a) Health Care

(b) Online Identity

(c) Refugee Policy

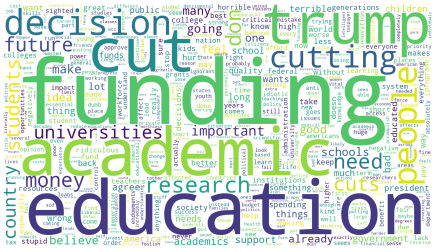
(d) Tariff Policy

(e) Vaccination Policy

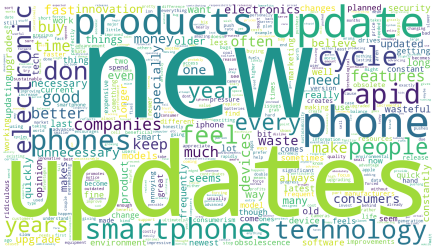
(f) AI Changes Life

(g) Influencer

(h) Tipping System



(i) Academic Funding



(j) Electronic Products

Figure 9: Wordclouds for the annotation dataset across all topics.

The wordclouds (see Figure 9) visualize participant opinions, highlighting salient terms and thematic differences across binary and open-ended deliberation topics.

B.3 Data Quality Checking

For all tasks involving human participation (i.e., the collection of the Public Opinion Dataset and the Human Judge Annotations), we conducted systematic quality control. Specifically, we monitored completion times for each participant to filter out responses that were submitted unrealistically quickly, which are indicative of inattentive or low-effort behavior. In addition, we randomized the order of questions for each participant to mitigate potential position bias and ensure fairness in evaluation. All retained data thus reflect responses that passed both attention and fairness checks.

C Details of Human Annotation Process

C.1 Ring-Based Summary Matching Algorithm

For the comparison task (see §2.2), we adopt a ring-based matching algorithm (see Alg. 1) rather than random sampling. This ensures balanced and repeated pairing of summaries within each deliberation question.

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

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

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

Algorithm 1 Ring-Based Summary Matching

Require: Summaries $S = [s_0, \ldots, s_{n-1}]$, seed, either k or M
$S \leftarrow \textsc{Permute}(S, \text{seed})$; $n \leftarrow |S|$
if M is specified **then**
  $k \leftarrow \lfloor M/n \rfloor$; $r \leftarrow M \bmod n$
else
  $k \leftarrow$ `min_comparisons_per_summary`
end if
$Pairs \leftarrow \varnothing$
for $i = 0$ **to** $n - 1$ **do**
  for $o = 1$ **to** k **do**
    $j \leftarrow (i + o) \bmod n$
    append (s_i, s_j) to $Pairs$
  end for
end for
if M is specified **then**
  for $i = 0$ **to** $r - 1$ **do**
    $j \leftarrow (i + (k + 1)) \bmod n$
    append (s_i, s_j) to $Pairs$
  end for
end if
return $Pairs$

D Details of Metrics

D.1 Summarization Quality Metric

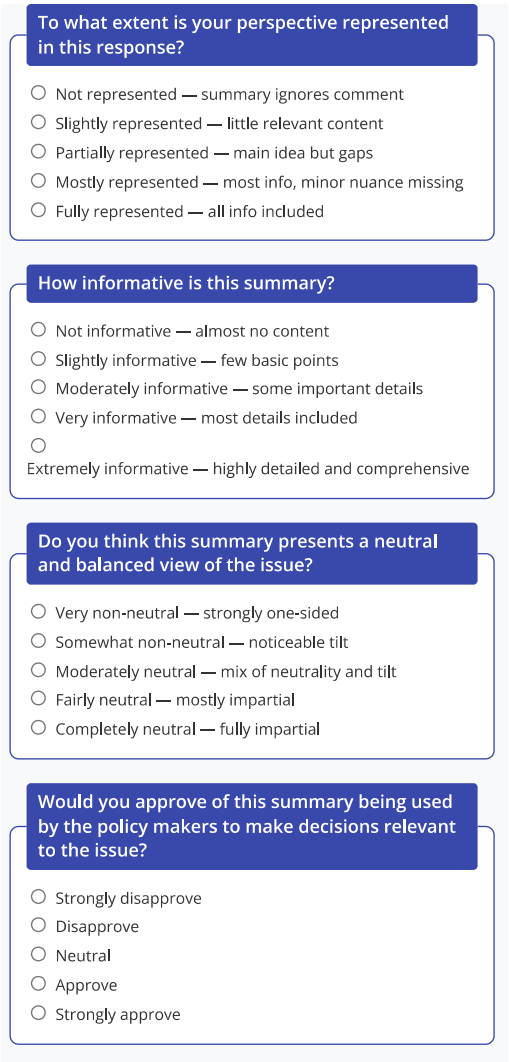
We evaluate summary quality along four deliberation-relevant dimensions:

- **Representativeness:** To what extent does the summary reflect the annotator's perspective?

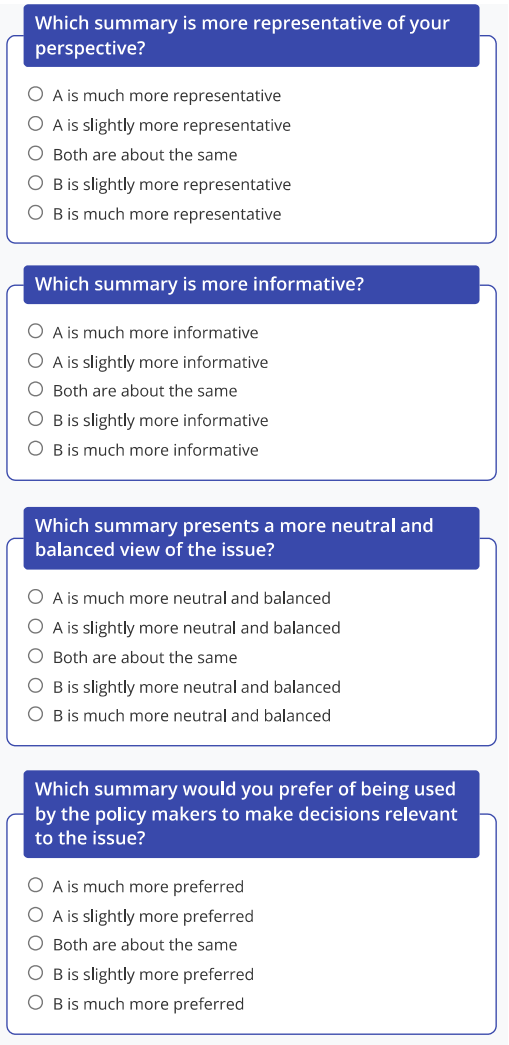
- **Informativeness:** How much useful information does the summary provide?
- **Neutrality:** Does the summary present a balanced and unbiased view of the issue?
- **Policy Approval:** Would the annotator approve of this summary being used by policymakers to make decisions?

Human annotation. These four metrics were operationalized in two complementary annotation tasks. As shown in Figure 10, 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, we benchmark LLMs by comparing their predicted judgments against human annotations. Our model, DELIBERATIONJUDGE, is a DeBERTa-based judge that outputs a four-dimensional score vector in the normalized range $[0, 1]$, where each value corresponds to one of the four metrics. These continuous predictions serve as summary-level scores for downstream comparison with human ratings and judgments.



(a) Rating task with Likert-scale options.



(b) Comparison task with five ordinal outcomes.

Figure 10: Screenshots of the annotation interfaces: (a) rating task and (b) comparison task.

D.2 Analytical Metric

Relative Preference Score For each model $\mathcal{M}$ and question $q \in \mathcal{Q}$, we define

$$\text{Diff}(\mathcal{M}, q) = \frac{1}{4} \sum_{d \in \{\text{rep,inf,neu,pol}\}} \hat{y}_q^{(d)} - \frac{1}{|\mathcal{Q}|} \sum_{q' \in \mathcal{Q}} \frac{1}{4} \sum_{d \in \{\text{rep,inf,neu,pol}\}} \hat{y}_{q'}^{(d)},$$

where $\hat{y}_q^{(d)}$ denotes the score on dimension d for question q produced by judge $\mathcal{J}_\theta$. Positive values indicate above-average performance, while negative values indicate below-average performance.

E Details of Judge Model Training

E.1 Multi-Output Regression Model Training

We implement a multi-output regression model to predict quality ratings across four key dimensions: perspective representation, informativeness, neutrality balance, and policy approval. Our approach employs several optimization strategies to enhance correlation performance between predicted and ground truth scores.

E.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: $S_{\mathcal{M},\tilde{\mathcal{O}}_i}$ [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 for enhanced feature learning. The regression head consists of a linear layer that maps the transformed features to four-dimensional outputs corresponding to the evaluation criteria (perspective representation, informativeness, neutrality balance, and policy approval). 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.

E.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 $\delta = 1.0$ instead of standard MSE to enhance robustness against outliers in human annotations. 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 L2 norm = 1.0 for training stability.

Optimization and Learning Rate Scheduling: We utilize 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.

E.2 Comparison of different base models

We have trained additional encoder-based models and large language models with different scales for comparison. For encoder-based models, we trained DeBERTa-v3-large and Longformer-base-4096 (Beltagy et al., 2020) using the same dataset with hyperparameter sweeping to identify optimal training configurations. These models were specifically chosen for their ability to process longer

sequences than other BERT-like models. For LLMs, we implemented two distinct training approaches. The first approach follows the same regression framework as the encoder-based models, where we adapt Qwen3-0.6B and Qwen3-4B for multi-output regression by adding a custom regression head on top of the pre-trained causal language model. This regression head employs a multi-layer architecture that progressively reduces dimensionality using LayerNorm, GELU activation, and dropout for regularization. The second approach leverages the instruction-following capabilities of LLMs by constructing an alpaca-format SFT dataset (Ding et al., 2023) based on the annotation data. In this approach, we directly fine-tune Qwen3-4B to generate structured JSON-formatted scores as text output, treating the regression task as a text generation problem rather than a traditional regression objective.

An example of the SFT data can be found below:

Alpaca Format SFT data

Instruction: We have made a deliberation with many annotators on the issue: Do you support the government accepting more refugees fleeing war or persecution?
One annotator's opinion on this question is: Only if they are properly vetted before being set free in our country, and have real proof that they are refugees.
Below is a summary of all people's opinions on the issue... `[summary content]`
Please evaluate this summary on the following 4 criteria using a 1-5 scale:

1. **To what extent is the annotator's opinion represented?** (1: Not represented — summary ignores comment, 2: Slightly represented — little relevant content, 3: Partially represented — main idea but gaps, 4: Mostly represented — most info, minor nuance missing, 5: Fully represented — all info included)
2. **How informative is this summary?** (1: Not informative ... 5: Extremely informative ...)
3. **Neutral and balanced view?** (1: Very non-neutral ... 5: Completely neutral ...)
4. **Policy maker approval?** (1: Strongly disapprove ... 5: Strongly approve ...)

Please provide your evaluation in the following JSON format: { "Representiveness": <1-5>, "Informativeness": <1-5>, "Neutrality": <1-5>, "Policy Approval": <1-5> }
Important: Select exactly one option for each criterion.
Output:

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

Table 6 presents a comparison of results across different models and training strategies. Several key insights emerge from this comparison. DeBERTa-v3-base with regression training consistently outperforms other models across all evaluation dimensions, achieving the highest correlations. This supports our core approach of employing encoder-based models with task-specific regression heads for multi-dimensional summary evaluation. Interestingly, increasing model size does not necessarily lead to better performance: DeBERTa-v3-large underperforms the base model across all dimensions, indicating that the base model achieves an optimal balance between capacity and generalization for this task.

Table 6: 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

F Details of Experiment

F.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}`

G Details of Human Data Collection

All human-involved procedures were implemented using two annotation platforms, `potato` (Pei et al., 2022) and `deliberation.io` (Pei et al., 2024), with participant recruitment conducted via Prolific[1]. The following screenshots illustrate the user interface and the design of annotation questions for each stage of human comment collection and evaluation.

G.1 Public Opinion Data Collection

For the ten questions introduced in Table 5, we launched ten separate studies on Prolific, each corresponding to one question. In each study, we recruit participants from the United States, aiming for a representative population across demographic groups (see §I for details). Moreover, Figures 11-14 illustrate the full opinion-submission workflow for the *tipping system* study, which we present as a pipeline example. In addition, we show in Figure 15 the Deliberation.io management interface that we used to coordinate the collection of all 10 studies.

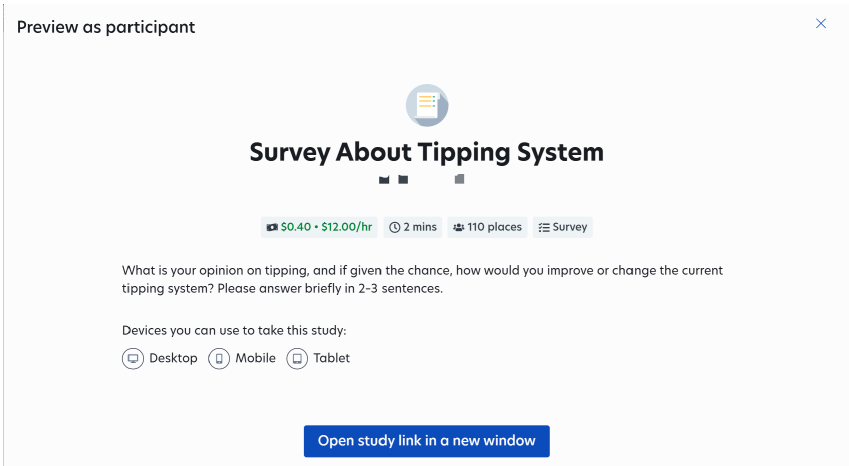


Figure 11: Example of Public opinion collection recuitment page powered by Prolific

G.2 Human Judge Data Collection

To collect human judgment data, we designed a structured annotation pipeline implemented on the `potato` platform (Pei et al., 2022). Recruitment of annotators was conducted via Prolific, ensuring a pool of participants distinct from those who contributed to the public opinion dataset. Each annotator was guided through a standardized workflow: recruitment and consent, task instructions, writing their own opinion, rating a model-generated summary along four evaluation dimensions, and final submission. Figures 16–21 provide interaction interface screenshots of this process.

[1] https://www.prolific.com/

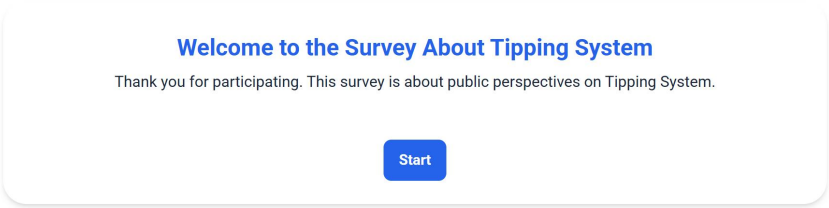


Figure 12: Example of public opinion collection start page powered by Deliberation.io

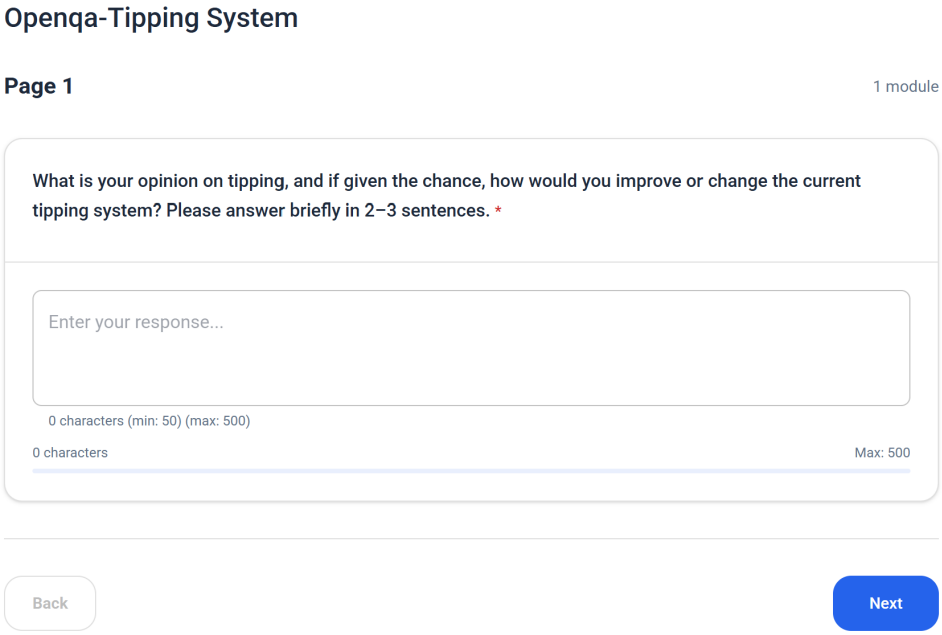


Figure 13: Example of public opinion collection question page powered by Deliberation.io

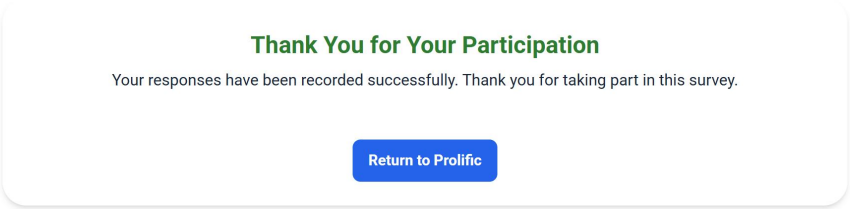


Figure 14: Example of public opinion collection ending page powered by Deliberation.io

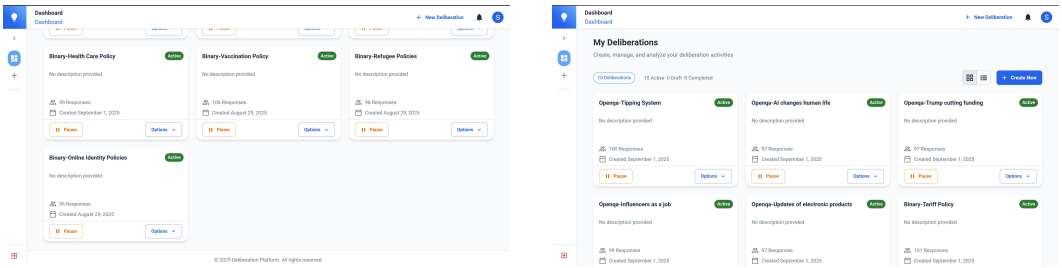


Figure 15: Example of the public opinion collection management interface on Deliberation.io.

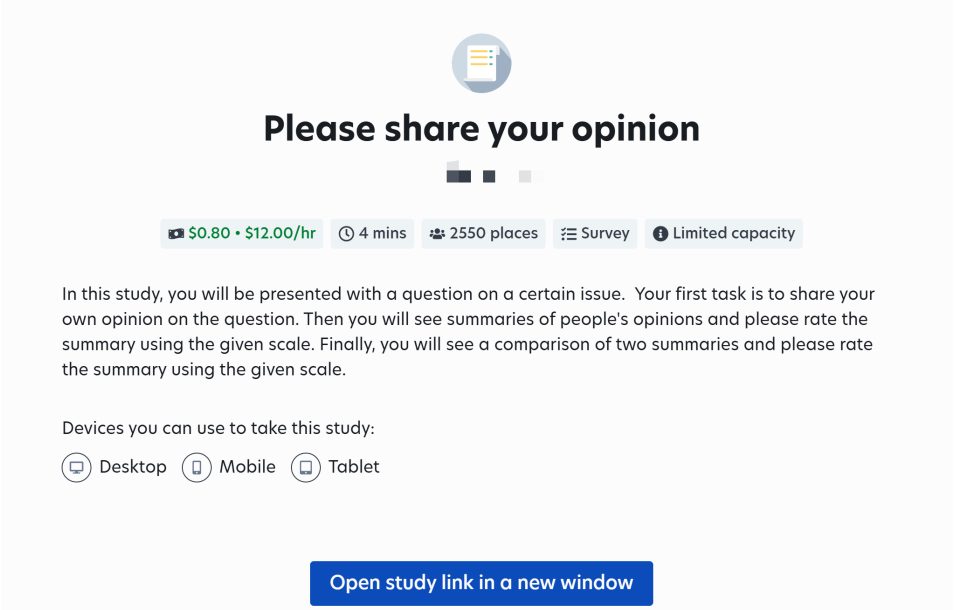


Figure 16: Recruitment page for human judge annotation (via Prolific).

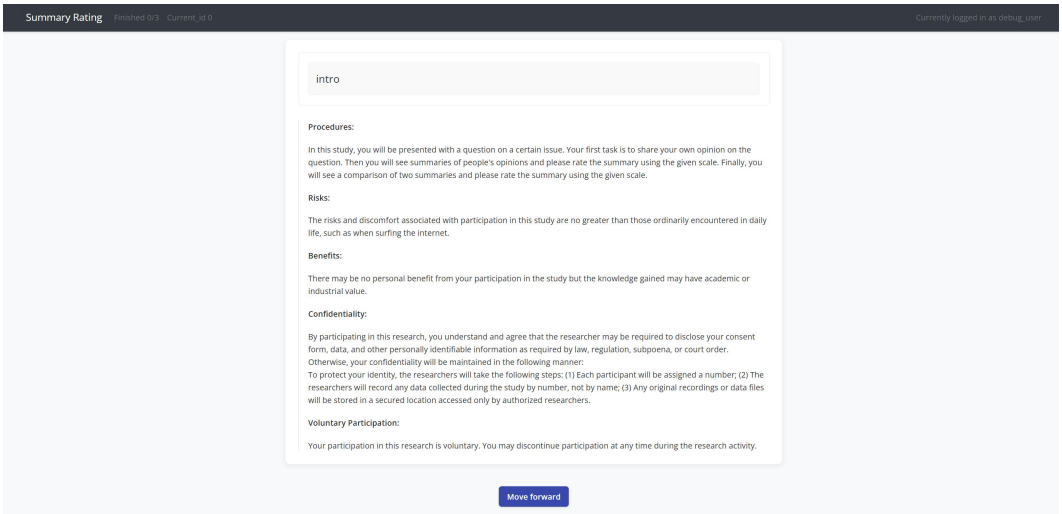


Figure 17: Task introduction page in the potato annotation system.

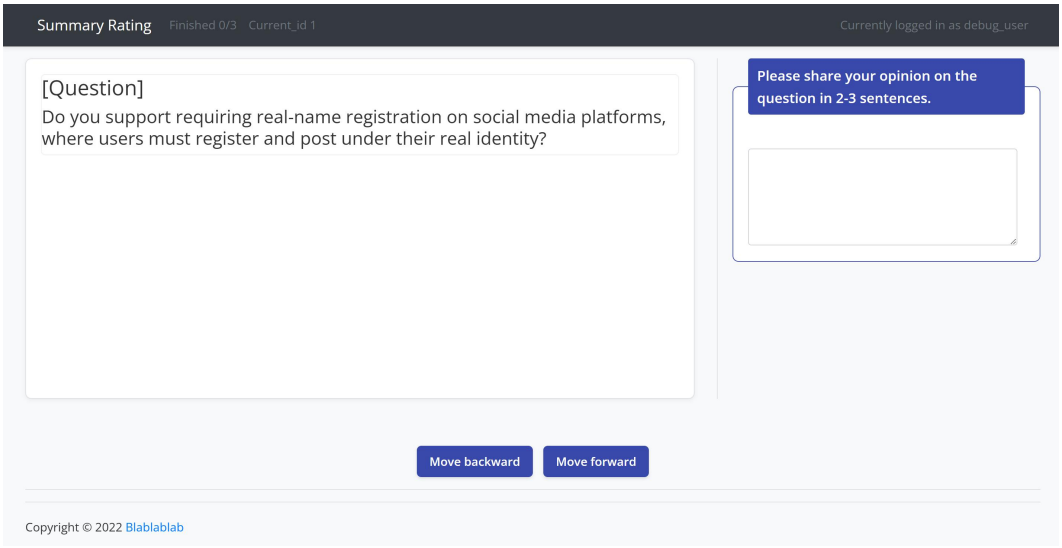


Figure 18: Annotator writing their own opinion for the given deliberation question.

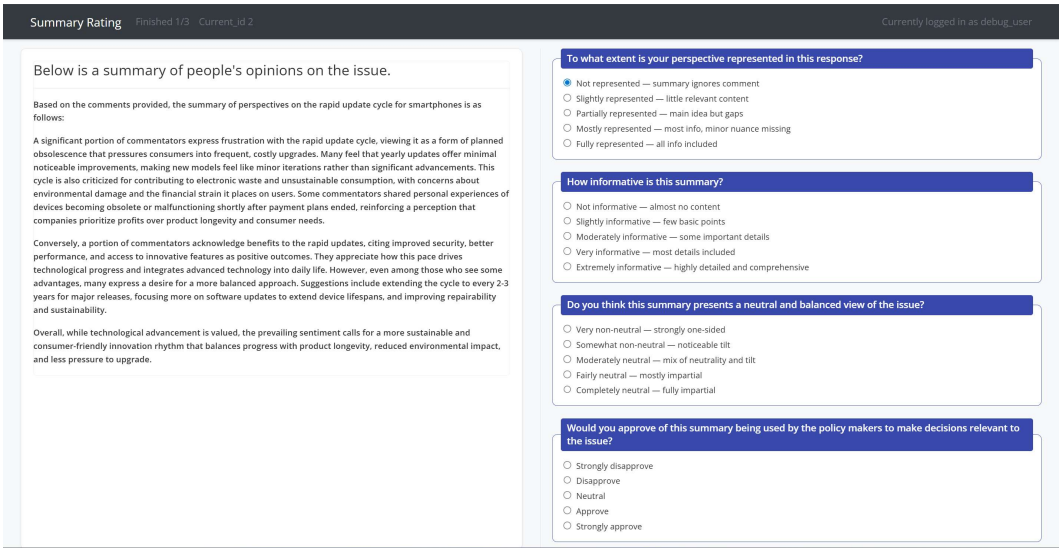


Figure 19: Annotation interface for rating judgment task.

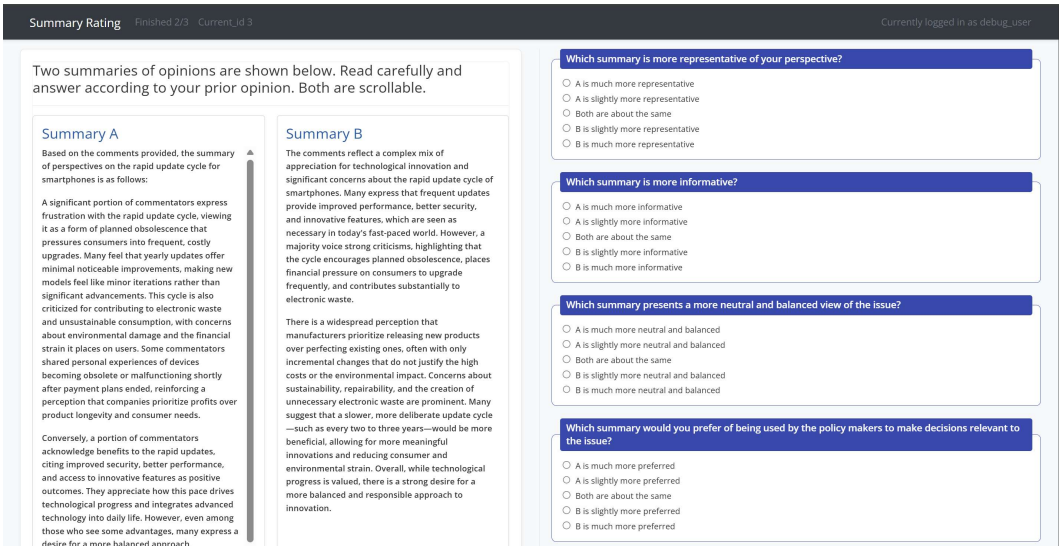


Figure 20: Annotation interface for compare judgment task.

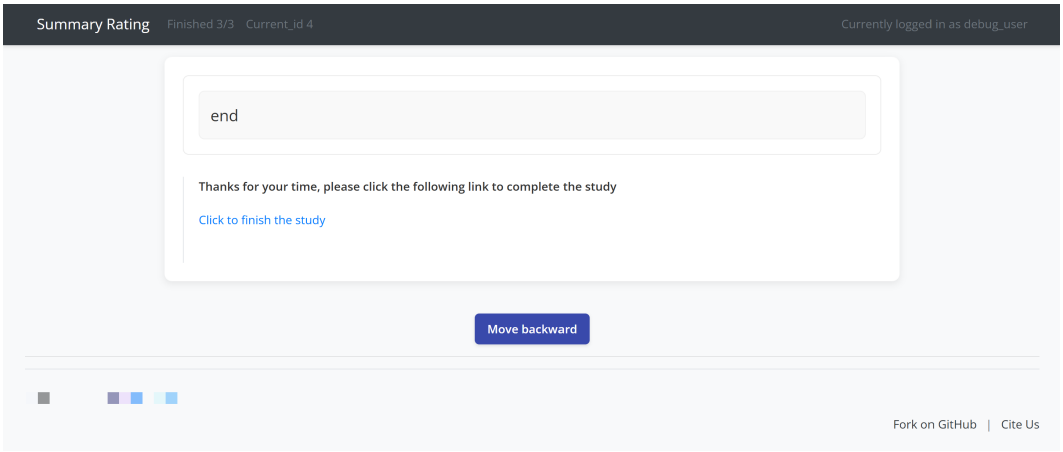


Figure 21: Final submission and completion page in the potato system.

H Additional Results

H.1 Human annotations

In this subsection, we present the detailed analysis of the human annotation results, as summarized in Figures 22–26. Figure 22 shows the distribution of overall ratings on a 5-point scale, while Figure 23 reports the corresponding distribution in the binary comparison setting. To examine the alignment between rating schemes and the binary comparison setting, Figure 24 illustrates the relationship between rating label distributions and sample-level win rates. Figure 25 further analyzes the impact of the number involved in summary on win rates, suggesting potential position or length biases in human judgments. Finally, Figure 26 presents the correlations among the four evaluated dimensions

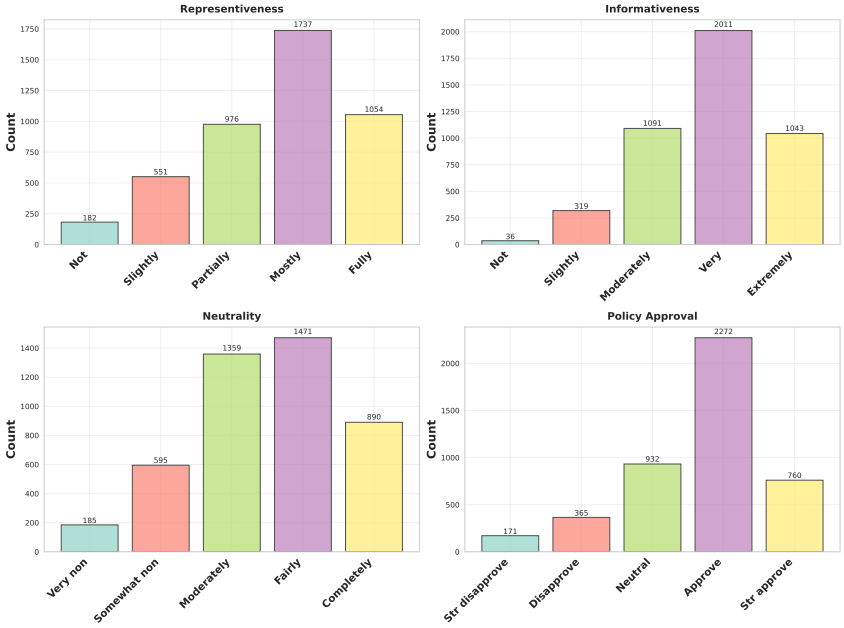


Figure 22: Distribution of overall human ratings on a 5-point scale.

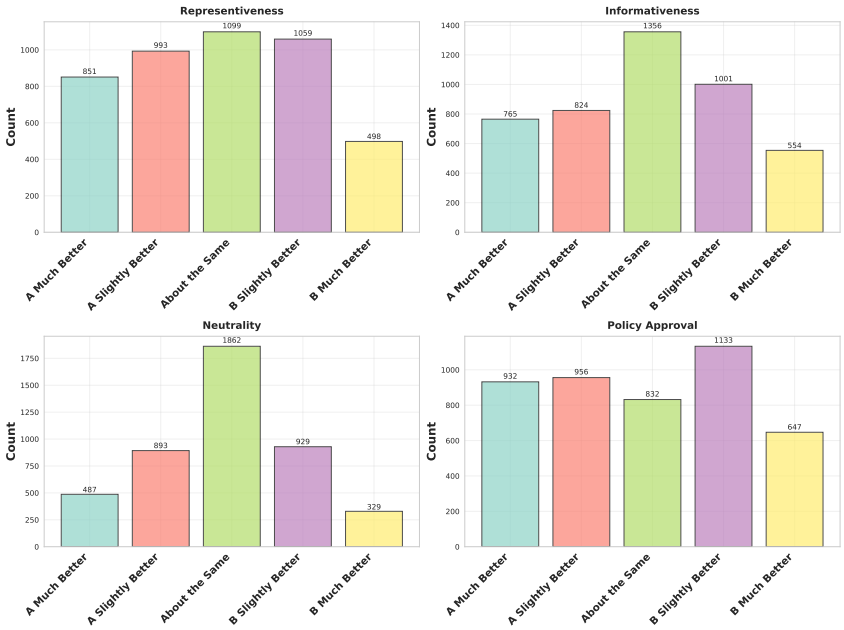


Figure 23: Distribution of human annotations in the binary comparison setting.

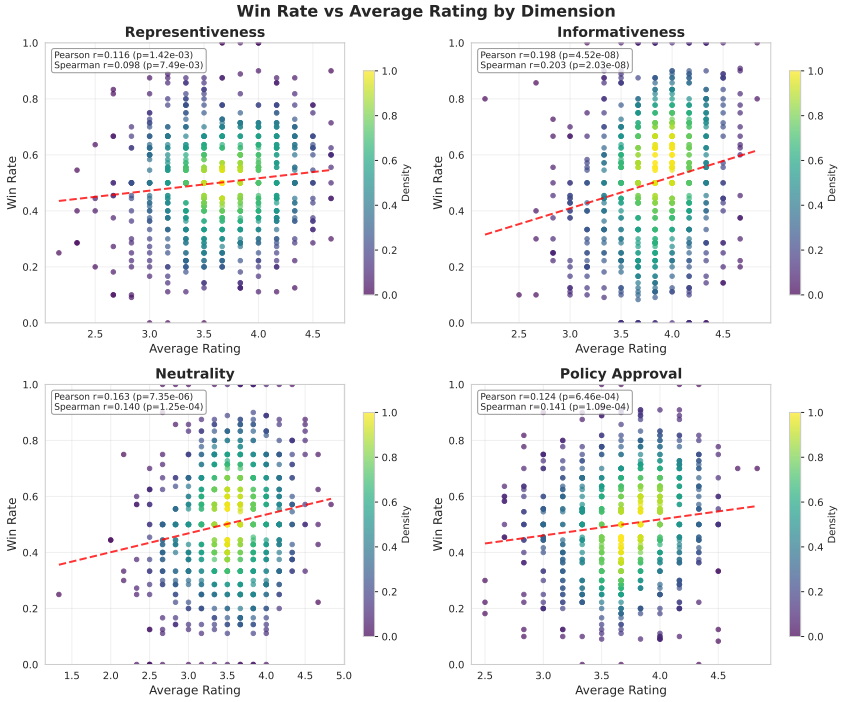


Figure 24: Relationship between rating label distribution and win rate at the sample level.

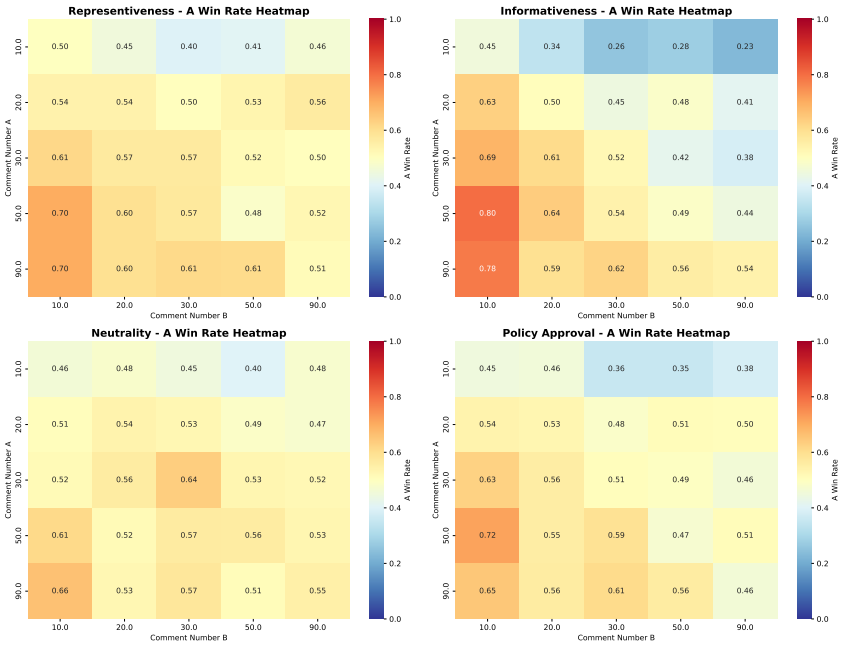


Figure 25: Heatmap of the relationship between the number of comments in a summary and its win rate in the human annotation dataset.

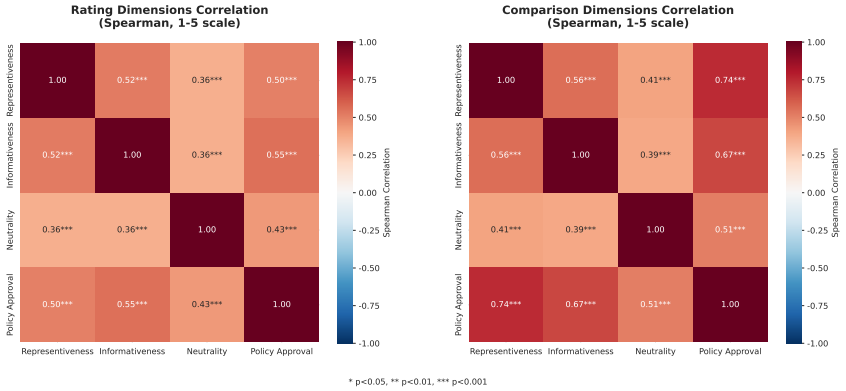


Figure 26: Correlations among the four evaluated dimensions.

I Human Participants Demographic

I.1 Public Opinion Dataset

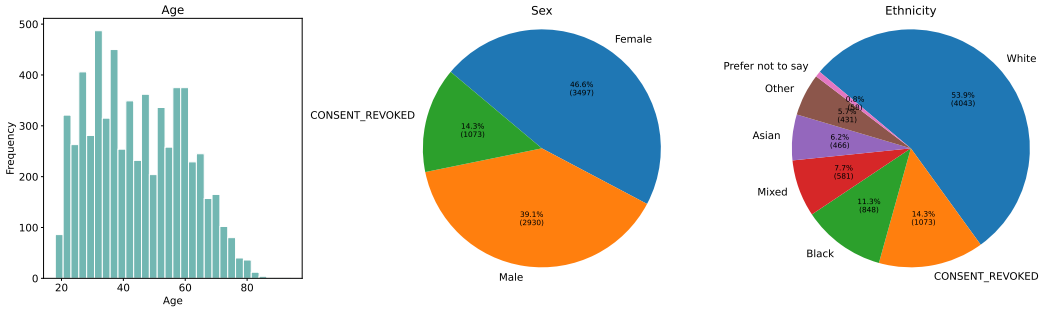


Figure 27: participants demographics for giving public opinion sample (n=7500), including 3000 sample from §2.2 and 4500 sample from §3: age distribution, sex and ethnicity breakdown; pie labels show % and counts.

I.2 Human Judge Annotation

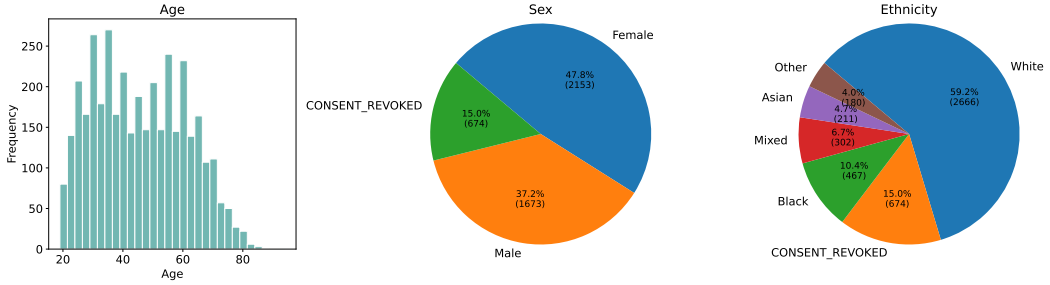


Figure 28: Human annotation participant demographics (n=4500): age histogram; sex and ethnicity pies; labels show % and counts.

J The Usage of Large Language Models (LLMs)

LLMs were used only occasionally to help polish the writing (propose new words, grammar, and spelling correction). All technical ideas, experimental designs, analyses, conclusions, and writing

were developed and carried out entirely by the authors. The authors have full responsibility for the final text.