AI for helping people to find consensus

Zhaowei Zhang

Insitute for AI, Peking University

25/11/2025

Motivations

Howard Raiffa 1982

p. 358,

“The Art and Science of Negotiation”

Often, disputants fail to reach an agreement when, in fact, a compromise does exist that could be to the advantage of all concerned.

And the agreements they do make are frequently inefficient: they could have made others that they all would have preferred.

Motivations

- Empirical evidence suggests that in practice it is **difficult to reach consensus** in deliberations.

- It is **nontrivial to provide utility function and clarify the optimizing direction** in practice.

- LLMs' impressive text generation capability may have potential to overcome the problem.

Motivations

Empirical evidence suggests that in practice it is **difficult to reach consensus** in deliberations.

- **Limited Bandwidth:** We cannot process thousands of free-text opinions.
- **Anchor Thinking:** We often get stuck on “A vs. B”, hard to think the hidden “Option C”.
- **Inefficiency:** Hard to achieve and always leaving value on the table.

Motivations

LLMs' impressive text generation capability may have potential to overcome the problem.

- **Inferring Utility from Language**: LLMs can process unstructured natural language preference directly.

- **Powerful text generation**: LLMs can generate solutions jumping out of the thinking anchor.

- **High efficiency**: LLMs almost have no time cost compared with human.

OUTLINE

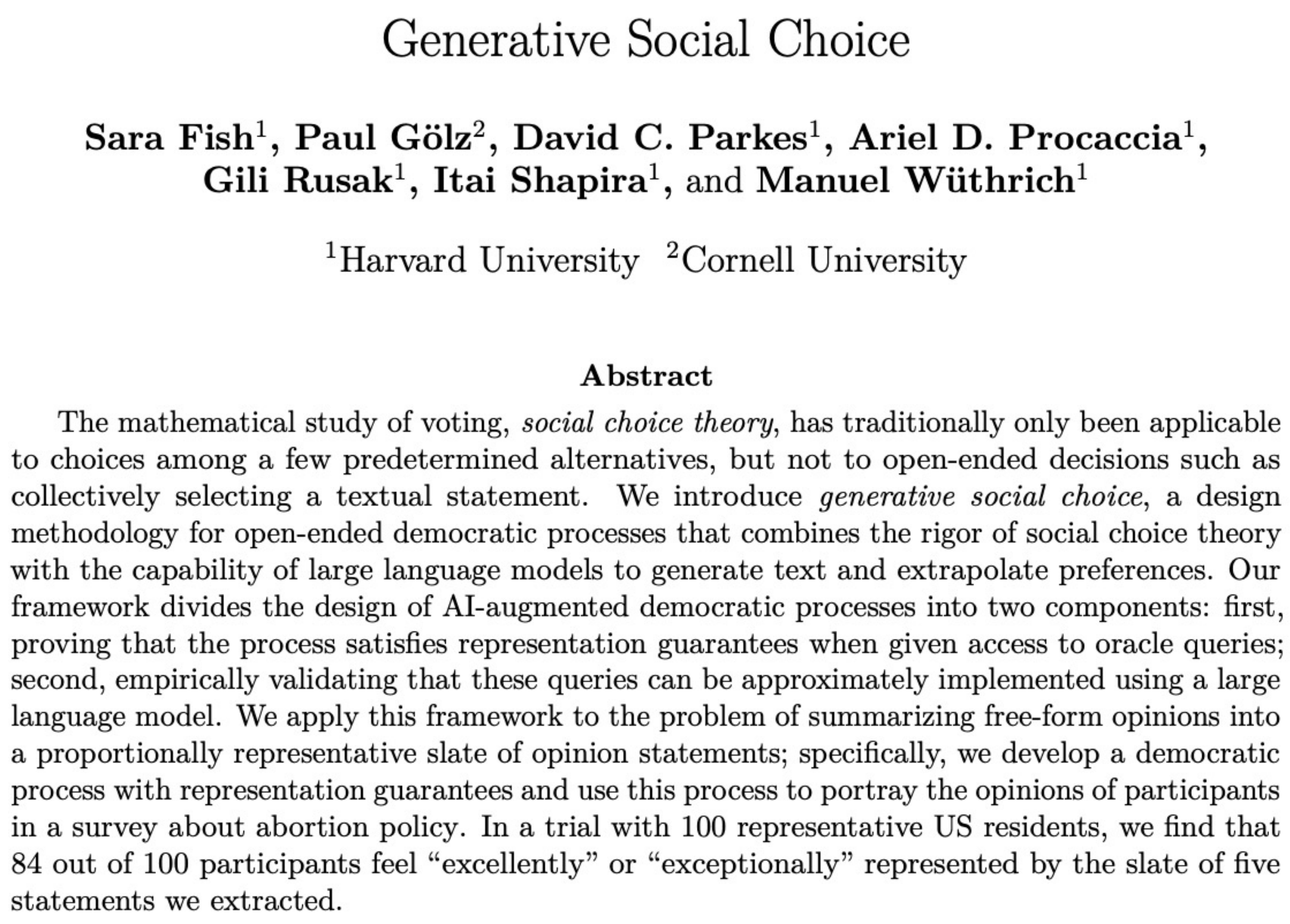
01 Consensus within crowd

02 Consensus within stakeholders

01 Consensus within crowd

- Generative Social Choice.
- Generative Social Choice: The Next Generation.

Generative Social Choice



Generative Social Choice

Sara Fish[1], **Paul Gölz**[2], **David C. Parkes**[1], **Ariel D. Procaccia**[1],
Gili Rusak[1], **Itai Shapira**[1], and **Manuel Wüthrich**[1]

[1]Harvard University [2]Cornell University

Abstract

The mathematical study of voting, *social choice theory*, has traditionally only been applicable to choices among a few predetermined alternatives, but not to open-ended decisions such as collectively selecting a textual statement. We introduce *generative social choice*, a design methodology for open-ended democratic processes that combines the rigor of social choice theory with the capability of large language models to generate text and extrapolate preferences. Our framework divides the design of AI-augmented democratic processes into two components: first, proving that the process satisfies representation guarantees when given access to oracle queries; second, empirically validating that these queries can be approximately implemented using a large language model. We apply this framework to the problem of summarizing free-form opinions into a proportionally representative slate of opinion statements; specifically, we develop a democratic process with representation guarantees and use this process to portray the opinions of participants in a survey about abortion policy. In a trial with 100 representative US residents, we find that 84 out of 100 participants feel "excellently" or "exceptionally" represented by the slate of five statements we extracted.

ACM EC, 2024

[1] Fish S, Gölz P, Parkes D C, et al. Generative Social Choice[C]//EC. 2024.

Motivations

- Voting is a key way in which groups to make common decisions.

- Its most typical setting is that voters express preferences over a finite set of candidates and select one as the outcome.

- But in some complex and subtle decisions, should better choices exist outside the given candidates?

[1] Fish S, Gölz P, Parkes D C, et al. Generative Social Choice[C]//EC. 2024.

Motivations: An Example

Revisited BREXIT, 2016

It’s 2016. Which policy would be the best to address the UK’s deepest problems?

[1] Fish S, Gölz P, Parkes D C, et al. Generative Social Choice[C]//EC. 2024.

Motivations: An Example

Revisited BREXIT, 2016

It’s 2016. Which policy would be the best to address the UK’s deepest problems?

A

Leave the European Union

B

Stay in the European Union

C

Ditch British cuisine for French cuisine

[1] Fish S, Gölz P, Parkes D C, et al. Generative Social Choice[C]//EC. 2024.

Motivations: An Example

Revisited BREXIT, 2016

It’s 2016. Which policy would be the best to address the UK’s deepest problems?

A

Leave the European Union

B

Stay in the European Union

C

Ditch British cuisine for French cuisine

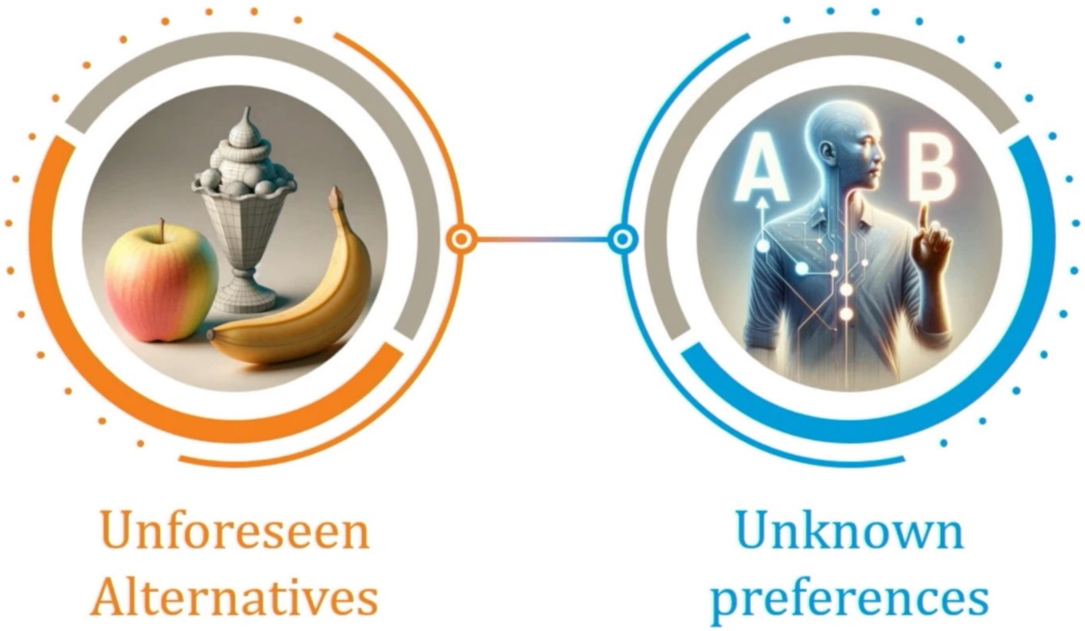
D

Relocate the royal family to California

[1] Fish S, Gölz P, Parkes D C, et al. Generative Social Choice[C]//EC. 2024.

| Generative Social Choice

Challenges:

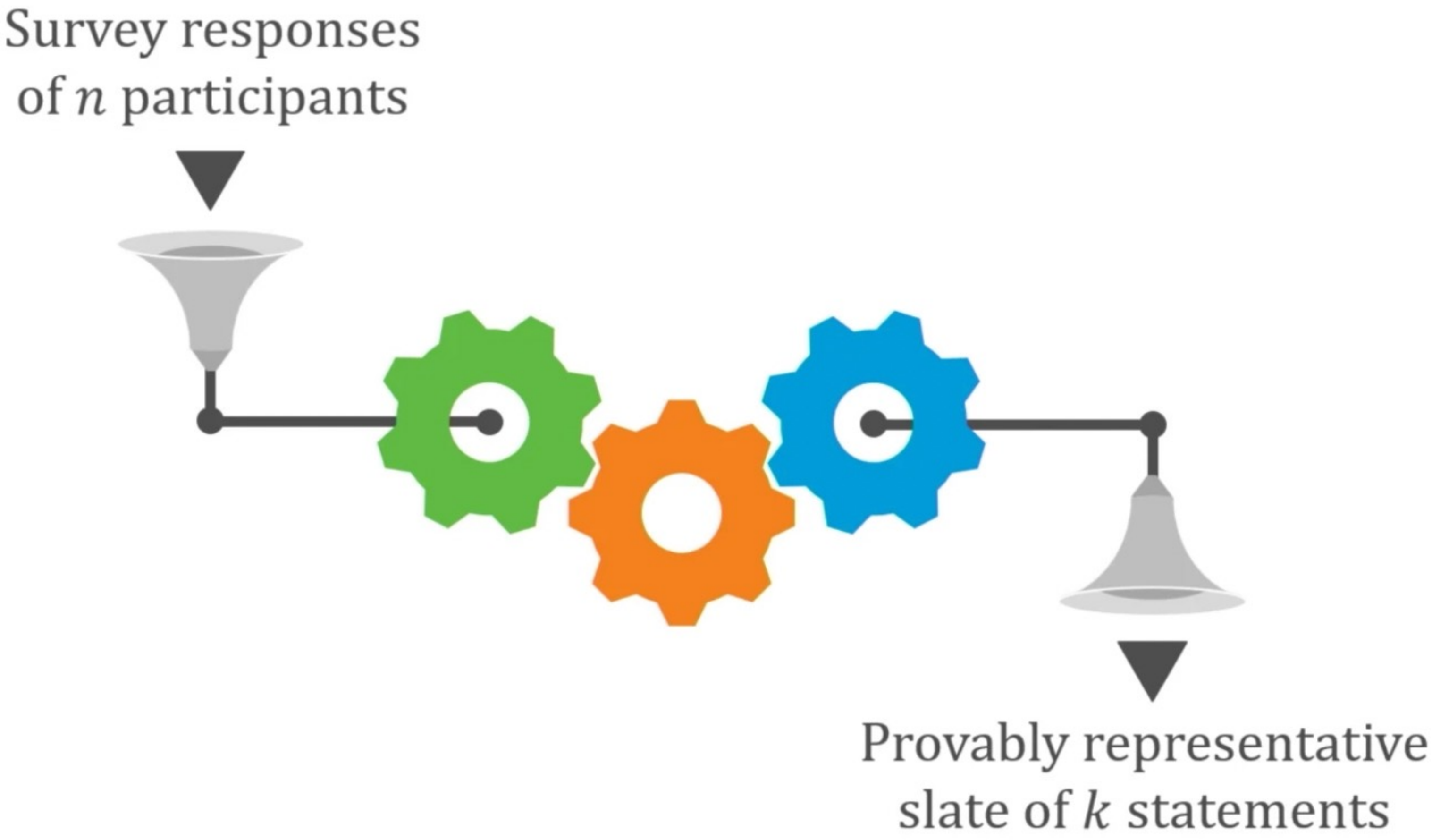


Solution & Goal:

Use LLM to generate alternatives that align with the broader preferences.

[1] Fish S, Gölz P, Parkes D C, et al. Generative Social Choice[C]//EC. 2024.

Generative Social Choice



[1] Fish S, Gölz P, Parkes D C, et al. Generative Social Choice[C]//EC. 2024.

Preliminary

Justified Representation (JR) – Can not guarantee the preference proportion

If there are n voters who need to select k representatives (statements), then any group of voters of size at least n/k, which at least one voter in this group must be satisfied with at least one of the final selected outcomes.

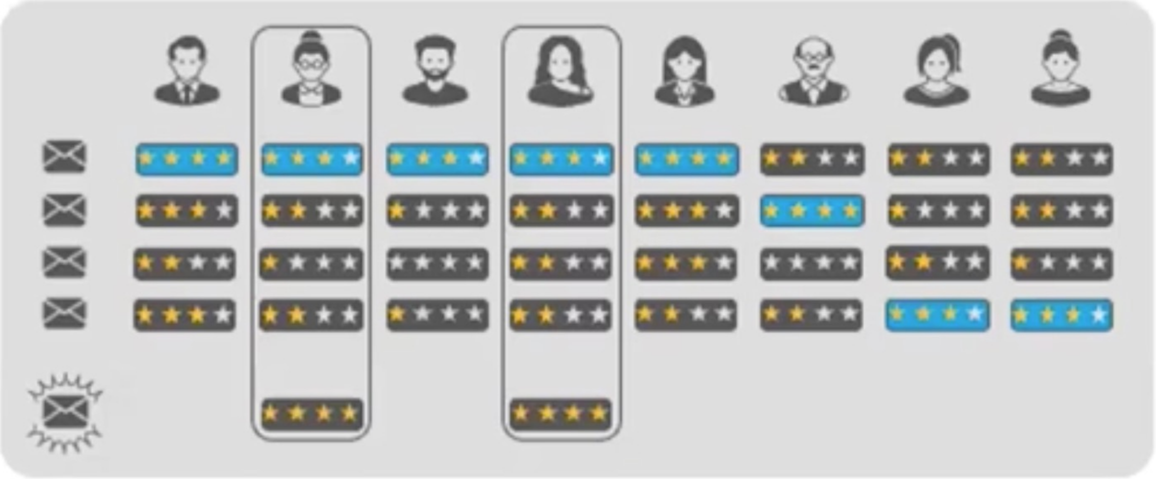
Balanced Justified Representation (BJR)

We need to **map each voter to only one statement**, which satisfies if there are n voters who need to select k statements, then there is no voter subset of size n/k whose maximum level of satisfaction for their matched statements is α and they all have satisfaction at least $\alpha' > \alpha$ for another statement.

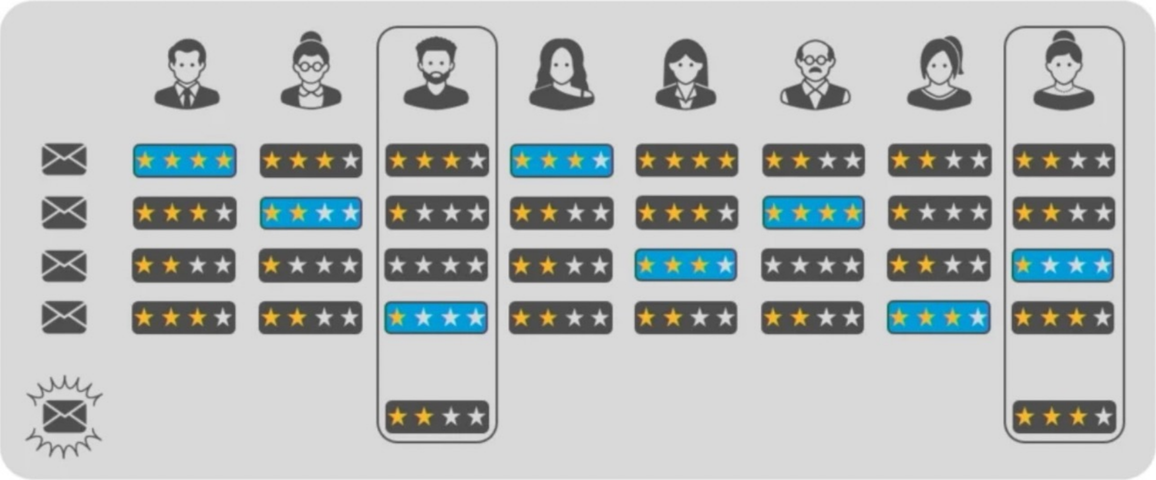
[1] Fish S, Gölz P, Parkes D C, et al. Generative Social Choice[C]//EC. 2024.

Preliminary

Justified Representation (JR)
– Can not guarantee the preference proportion

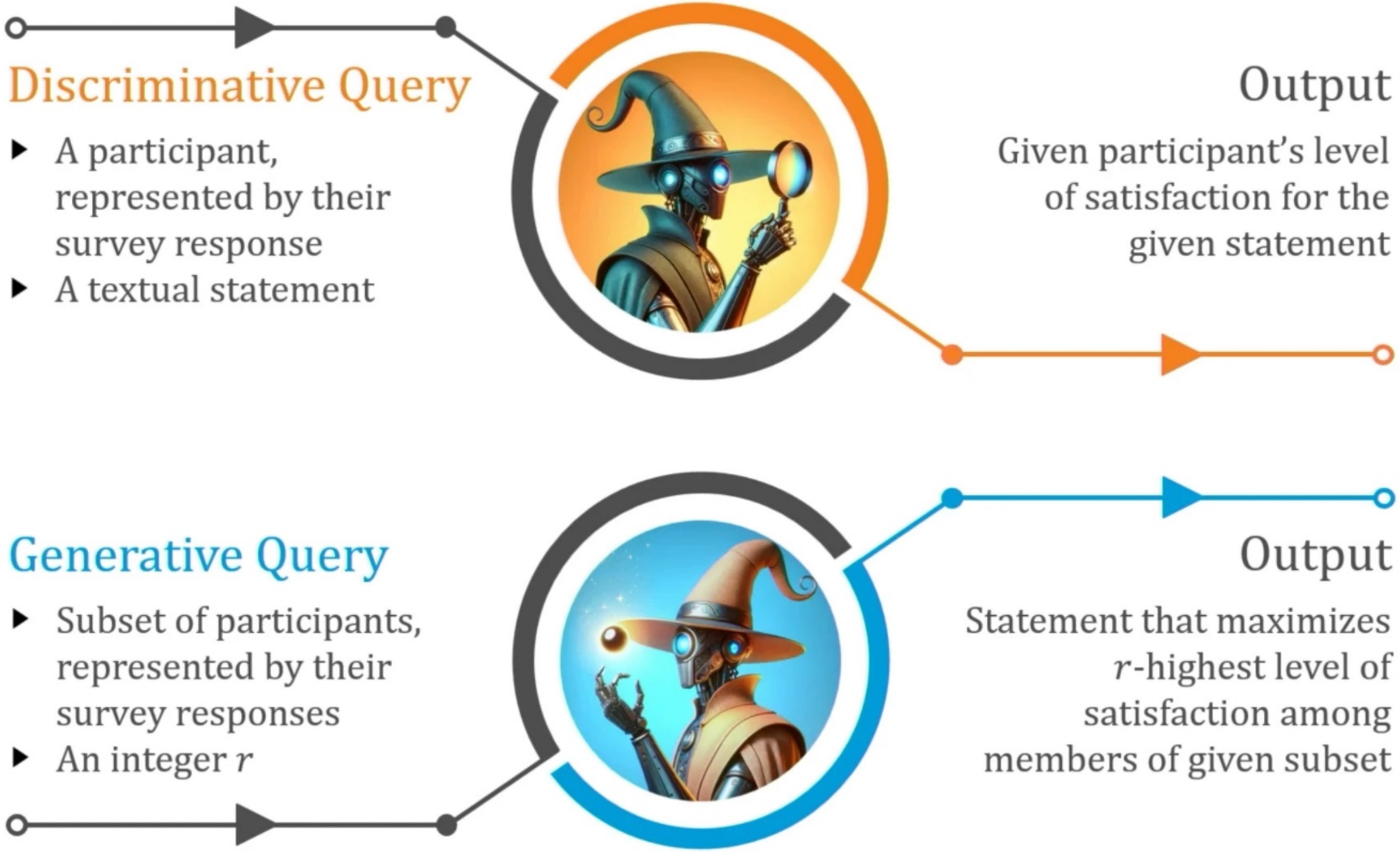


Balanced Justified Representation (BJR)



[1] Fish S, Gölz P, Parkes D C, et al. Generative Social Choice[C]//EC. 2024.

Method: Operators



Discriminative Query

- A participant, represented by their survey response
- A textual statement

Output

Given participant's level of satisfaction for the given statement

Generative Query

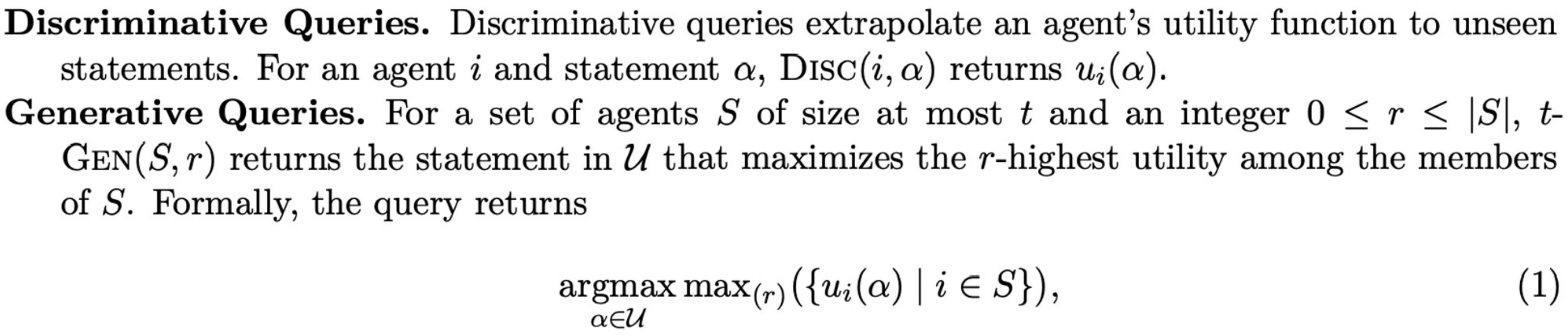
- Subset of participants, represented by their survey responses
- An integer r

Output

Statement that maximizes r-highest level of satisfaction among members of given subset

[1] Fish S, Gölz P, Parkes D C, et al. Generative Social Choice[C]//EC. 2024.

Method: Operators



Discriminative Queries. Discriminative queries extrapolate an agent's utility function to unseen statements. For an agent i and statement α, $\textsc{Disc}(i, \alpha)$ returns $u_i(\alpha)$.

Generative Queries. For a set of agents S of size at most t and an integer $0 \leq r \leq |S|$, t-$\textsc{Gen}(S, r)$ returns the statement in $\mathcal{U}$ that maximizes the r-highest utility among the members of S. Formally, the query returns

$$\operatorname*{argmax}_{\alpha \in \mathcal{U}} \max{}_{(r)}(\{u_i(\alpha) \mid i \in S\}), \tag{1}$$

Discriminative Query:

Use GPT-4 to predict each voter's preference with few-shot demonstrations.

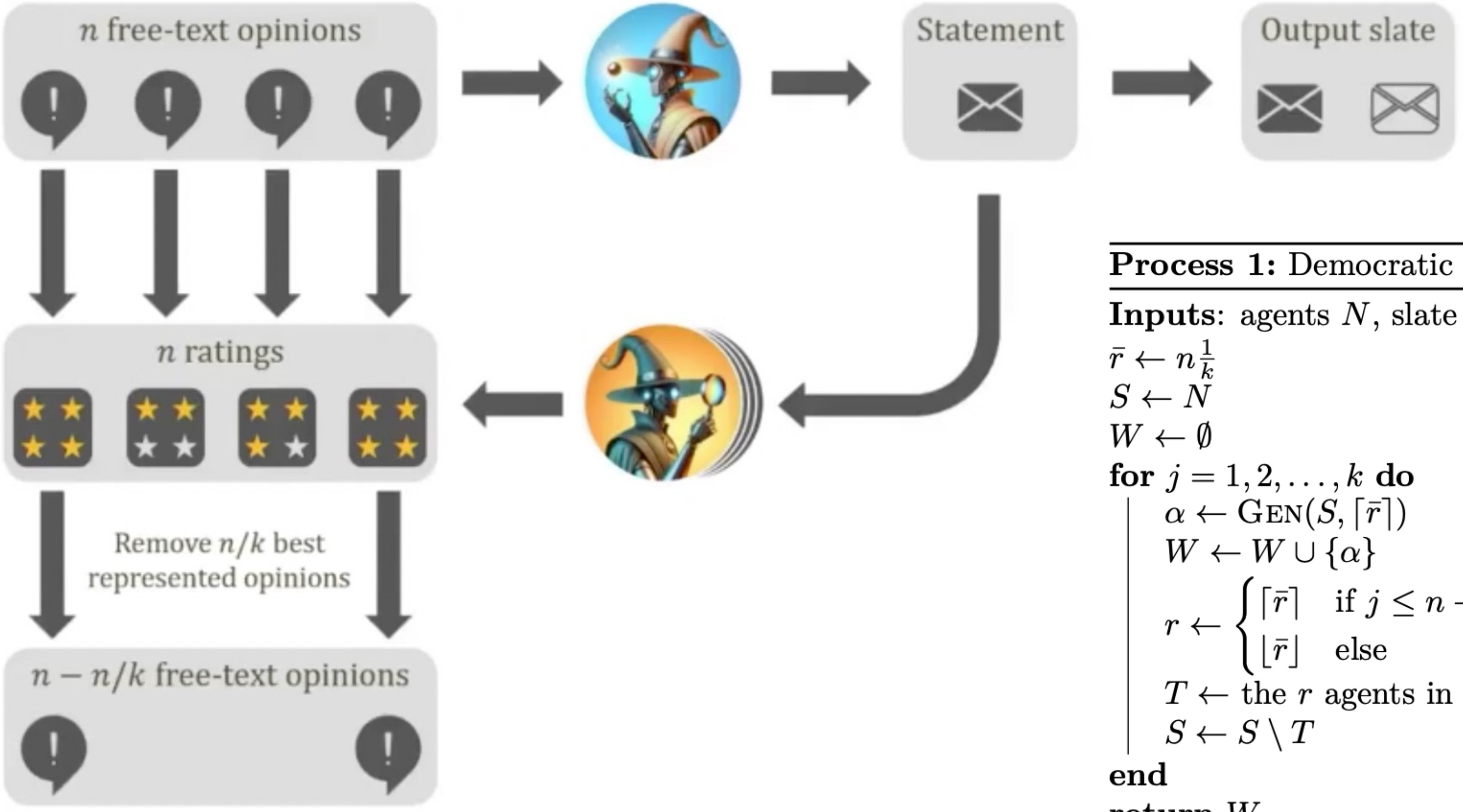
Generative Query:

(1) Use a variation of clustering method to find statements sharing close opinions.

(2) Prompt LLMs to generate an opinion representing these statements.

[1] Fish S, Gölz P, Parkes D C, et al. Generative Social Choice[C]//EC. 2024.

Method: Algorithm


n free-text opinions
Statement
Output slate
n ratings
Remove n/k best represented opinions
$n - n/k$ free-text opinions

Process 1: Democratic Process for Balanced Justified Representation

```
Inputs: agents $N$, slate size $k$
$\bar{r} \leftarrow n\frac{1}{k}$
$S \leftarrow N$
$W \leftarrow \emptyset$
for $j = 1, 2, \ldots, k$ do
    $\alpha \leftarrow \textsc{Gen}(S, \lceil \bar{r} \rceil)$
    $W \leftarrow W \cup \{\alpha\}$
    $r \leftarrow \begin{cases} \lceil \bar{r} \rceil & \text{if } j \leq n - k \cdot \lfloor \bar{r} \rfloor \\ \lfloor \bar{r} \rfloor & \text{else} \end{cases}$
    $T \leftarrow$ the $r$ agents in $S$ with largest $\textsc{Disc}(\cdot, \alpha)$
    $S \leftarrow S \setminus T$
end
return $W$
```

[1] Fish S, Gölz P, Parkes D C, et al. Generative Social Choice[C]//EC. 2024.

Experiment: Building dataset

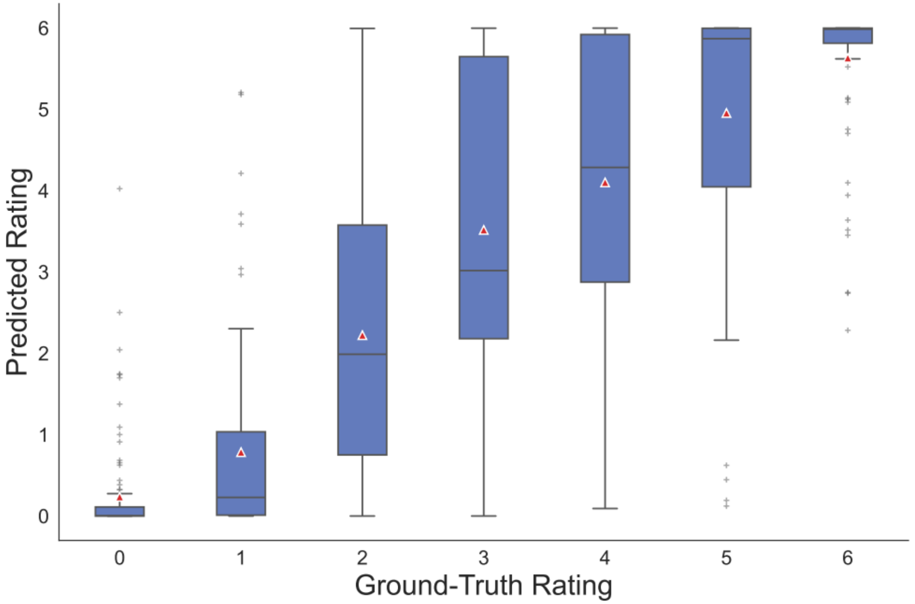
They recruit 100 human annotators to build a dataset on the opinions about the personalization of chatbot.

- Write open-ended opinions on the topic.
- Label their preference with Likert scalar from 1-5 for 6 others opinions.

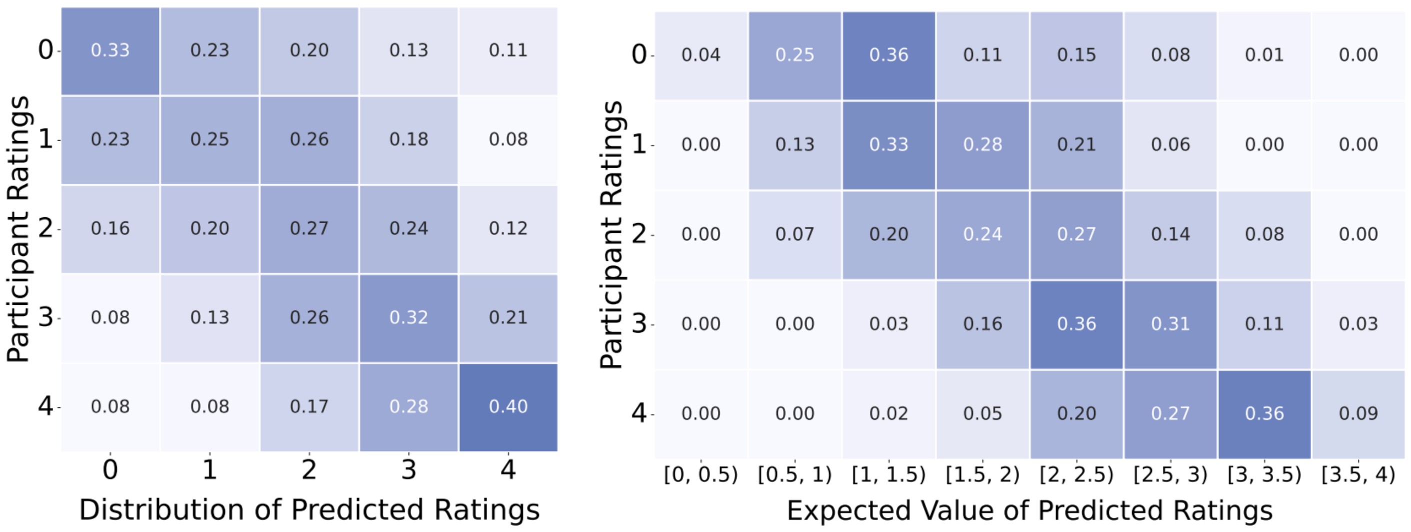
[1] Fish S, Gölz P, Parkes D C, et al. Generative Social Choice[C]//EC. 2024.

Experiment: Build discriminative operator

Since the dataset has preference label of random 6 opinions for 100 annotators, they use 5 opinions as few shots, to predict each voter's preference with GPT-4.



Distribution of discriminative query predictions.



	0	1	2	3	4
0	0.33	0.23	0.20	0.13	0.11
1	0.23	0.25	0.26	0.18	0.08
2	0.16	0.20	0.27	0.24	0.12
3	0.08	0.13	0.26	0.32	0.21
4	0.08	0.08	0.17	0.28	0.40

(a)

	[0, 0.5)	[0.5, 1)	[1, 1.5)	[1.5, 2)	[2, 2.5)	[2.5, 3)	[3, 3.5)	[3.5, 4)
0	0.04	0.25	0.36	0.11	0.15	0.08	0.01	0.00
1	0.00	0.13	0.33	0.28	0.21	0.06	0.00	0.00
2	0.00	0.07	0.20	0.24	0.27	0.14	0.08	0.00
3	0.00	0.00	0.03	0.16	0.36	0.31	0.11	0.03
4	0.00	0.00	0.02	0.05	0.20	0.27	0.36	0.09

(b)

Confusion matrix of discriminative queries.

[1] Fish S, Gölz P, Parkes D C, et al. Generative Social Choice[C]//EC. 2024.

Experiment: Build generative operator

(1) Use a variation of clustering method to find statements sharing close opinions.
(2) Prompt LLMs to generate an opinion representing these statements.

Clustering Steps:

1 **Feature Extraction**

They extract 50 key opinion features from user opinions to create a structured basis for comparison.

2 **Vector Embedding**

LLM rates each voter's agreement (1-7) with features, mapping qualitative opinions into a quantitative 50-dimensional vector space.

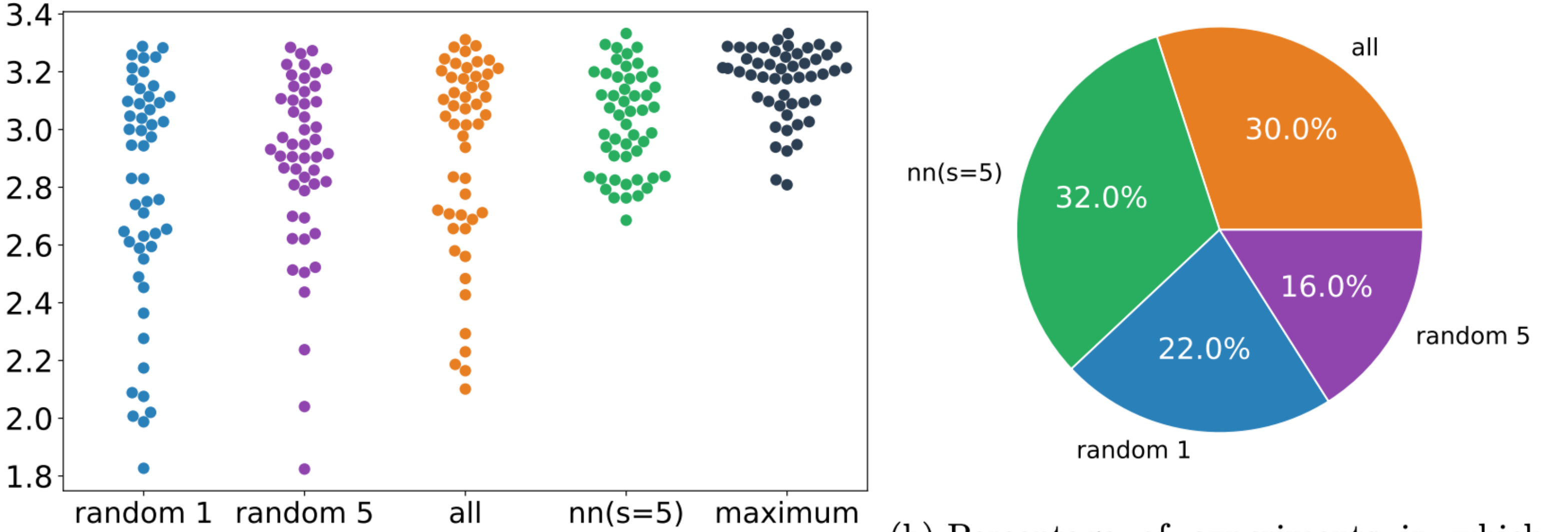
3 **Hybrid Clustering**

They use Balanced K-means to identify cohesive subgroups for generation.

[1] Fish S, Gölz P, Parkes D C, et al. Generative Social Choice[C]//EC. 2024.

Experiment: Build generative operator

Prompt LLMs to generate k opinions representing these clustering statements.



(a) Distribution of the 20th-highest utility obtained by the statements from different sources.

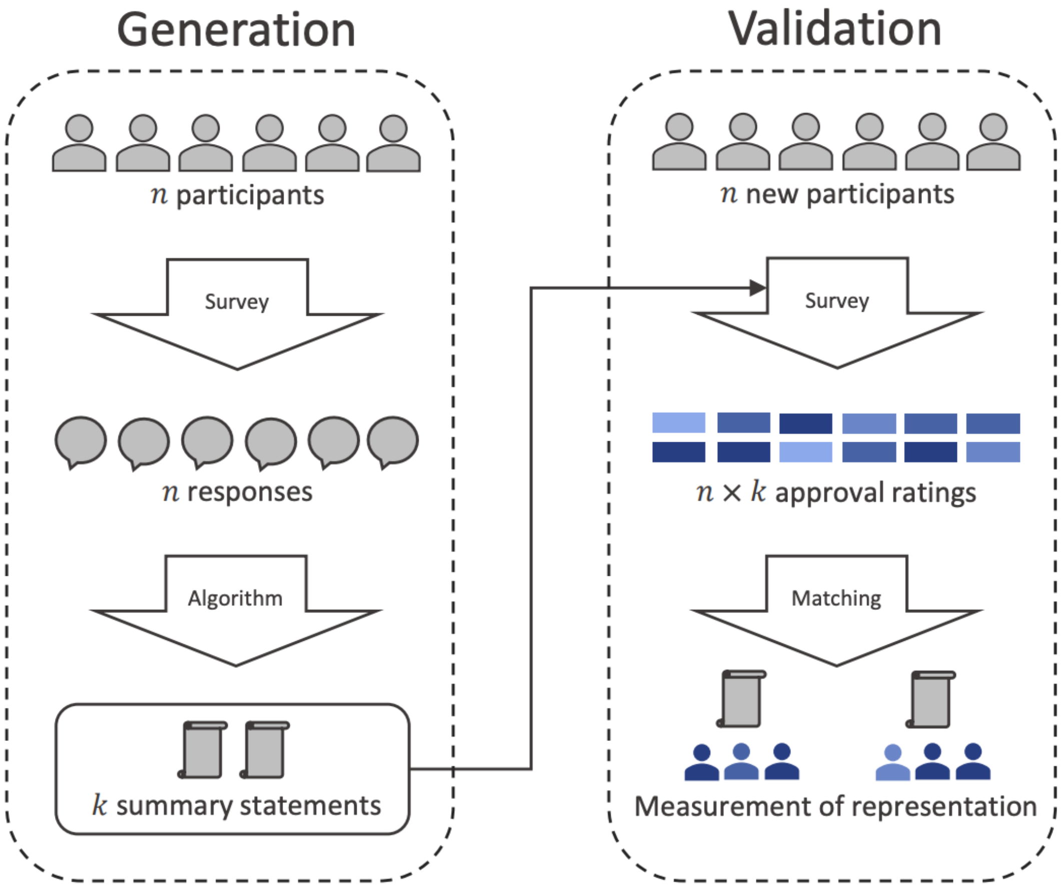
(b) Percentage of experiments in which each statement source obtained a higher 20th-highest utility than all others.

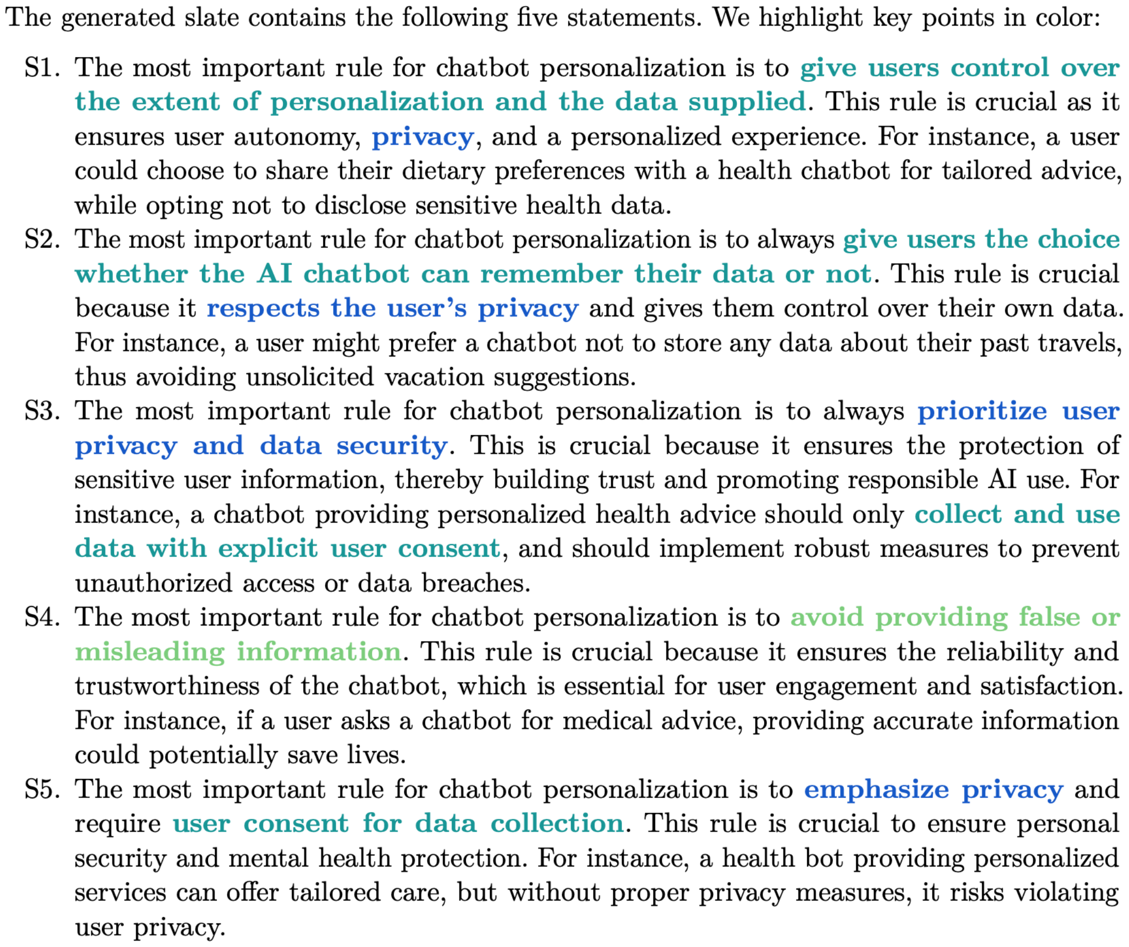
Figure 2: Evaluation of the 20th-highest utility obtained by different generation sources in our experiments. Each of the 50 datapoints corresponds to a random sample of 40 out of the 100 agents.

[1] Fish S, Gölz P, Parkes D C, et al. Generative Social Choice[C]//EC. 2024.

Experiment

They generative 5 statements over 100 opinions, and recruit other 100 annotators for evaluation. Results show that 93% of annotators agree with at least one of the outcomes.





The generated slate contains the following five statements. We highlight key points in color:

S1. The most important rule for chatbot personalization is to **give users control over the extent of personalization and the data supplied**. This rule is crucial as it ensures user autonomy, **privacy**, and a personalized experience. For instance, a user could choose to share their dietary preferences with a health chatbot for tailored advice, while opting not to disclose sensitive health data.

S2. The most important rule for chatbot personalization is to always **give users the choice whether the AI chatbot can remember their data or not**. This rule is crucial because it **respects the user's privacy** and gives them control over their own data. For instance, a user might prefer a chatbot not to store any data about their past travels, thus avoiding unsolicited vacation suggestions.

S3. The most important rule for chatbot personalization is to always **prioritize user privacy and data security**. This is crucial because it ensures the protection of sensitive user information, thereby building trust and promoting responsible AI use. For instance, a chatbot providing personalized health advice should only **collect and use data with explicit user consent**, and should implement robust measures to prevent unauthorized access or data breaches.

S4. The most important rule for chatbot personalization is to **avoid providing false or misleading information**. This rule is crucial because it ensures the reliability and trustworthiness of the chatbot, which is essential for user engagement and satisfaction. For instance, if a user asks a chatbot for medical advice, providing accurate information could potentially save lives.

S5. The most important rule for chatbot personalization is to **emphasize privacy** and require **user consent for data collection**. This rule is crucial to ensure personal security and mental health protection. For instance, a health bot providing personalized services can offer tailored care, but without proper privacy measures, it risks violating user privacy.

[1] Fish S, Gölz P, Parkes D C, et al. Generative Social Choice[C]//EC. 2024.

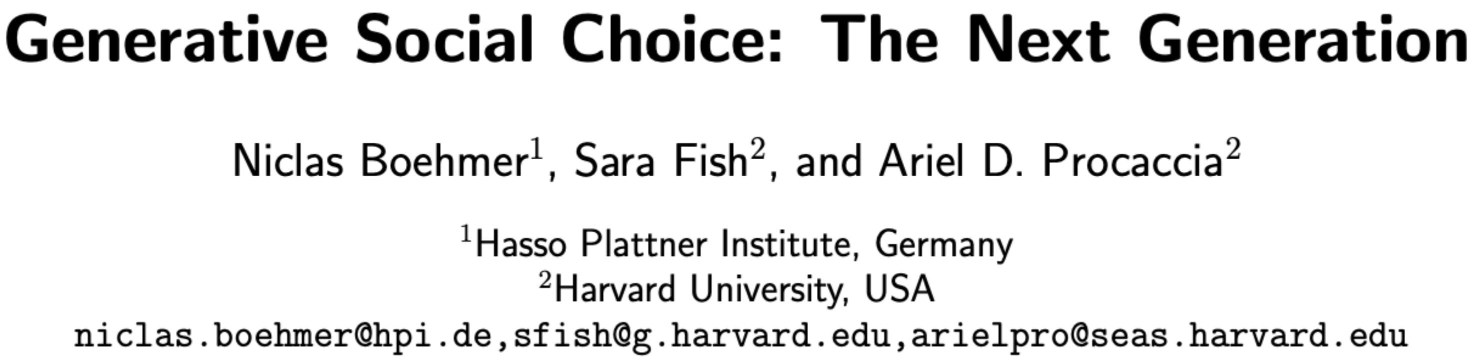
| Generative Social Choice: Take Away

- Methodology: Strict mathematical definition with practical approximation to solve a social science problems.

- Motivations: Utilize LLM’s powerful generation ability to implement some theoretically hard conditions.

[1] Fish S, Gölz P, Parkes D C, et al. Generative Social Choice[C]//EC. 2024.

Generative Social Choice: The Next Generation



Generative Social Choice: The Next Generation

Niclas Boehmer[1], Sara Fish[2], and Ariel D. Procaccia[2]

[1]Hasso Plattner Institute, Germany
[2]Harvard University, USA
niclas.boehmer@hpi.de,sfish@g.harvard.edu,arielpro@seas.harvard.edu

ICML (Oral), 2025

[1] Boehmer N, Fish S, Procaccia A D. Generative Social Choice: The Next Generation[C]//ICML 2025.

Motivations: Problems with GSC

1. **Costs and budgets**

   The statement number k is hard to choose, and the unlimited length of the statements can be a heavy burden for understanding and implementing these decisions.

2. **Approximate queries**

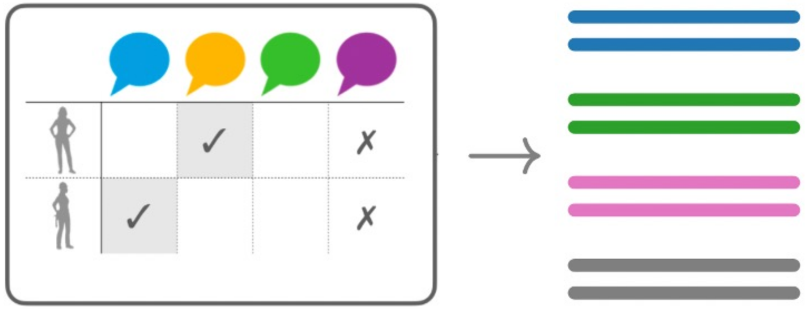
   If the generative and discriminative operators in GSC can not be optimal, the theoretical guarantee in GSC will fail.

How to generate with cost constraints and noisy operator results?

[1] Boehmer N, Fish S, Procaccia A D. Generative Social Choice: The Next Generation[C]//ICML 2025.

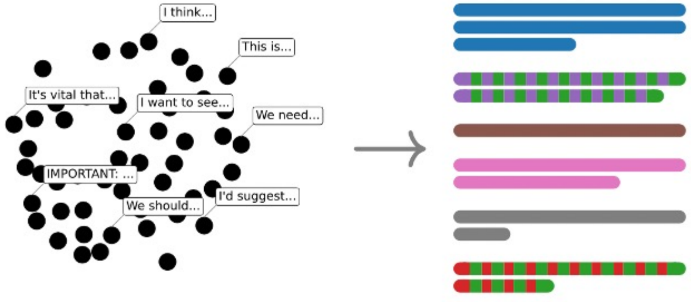
Compared with original GSC

Generative Social Choice



1. User sets number of statements k
2. Each statement represents $1/k$ of users
3. Guarantees only for perfect query results
4. Implementation for structured user data

The Next Generation



1. Algorithm adaptively chooses k
2. Variable statement lengths (support $\propto$ length)
3. Process and guarantees for noisy query results
4. Flexible implementation compatible with unstructured user data

[1] Boehmer N, Fish S, Procaccia A D. Generative Social Choice: The Next Generation[C]//ICML 2025.

Preliminary

(b, d) – cost Balanced Justified Representation ((b, d)-cBJR)

Let $c: \mathcal{U} \rightarrow \mathbb{N}_0$ be a cost function that maps each statement $\alpha \in \mathcal{U}$ to its cost $c(\alpha)$, and let B be the total budget. For $b \in \mathbb{N}_0$ and $d \in \mathbb{R}_{\geq 1}$, a slate W satisfies (b, d)-cBJR if there is a function $\omega: N \rightarrow W$ matching agents to statements in a balanced way, such that no coalition $S \subseteq N$, statement $\alpha \in \mathcal{U}$, and utility threshold $\theta \in [r]$ satisfies (i) $|S| \geq d \cdot \lceil c(\alpha) \cdot n/B \rceil$, (ii) $u_i(\alpha) \geq \theta$ for all $i \in S$, and (iii) $u_i\big(\omega(i)\big) < \theta - b$ for all $i \in S$.

cost Balanced Justified Representation (cBJR)

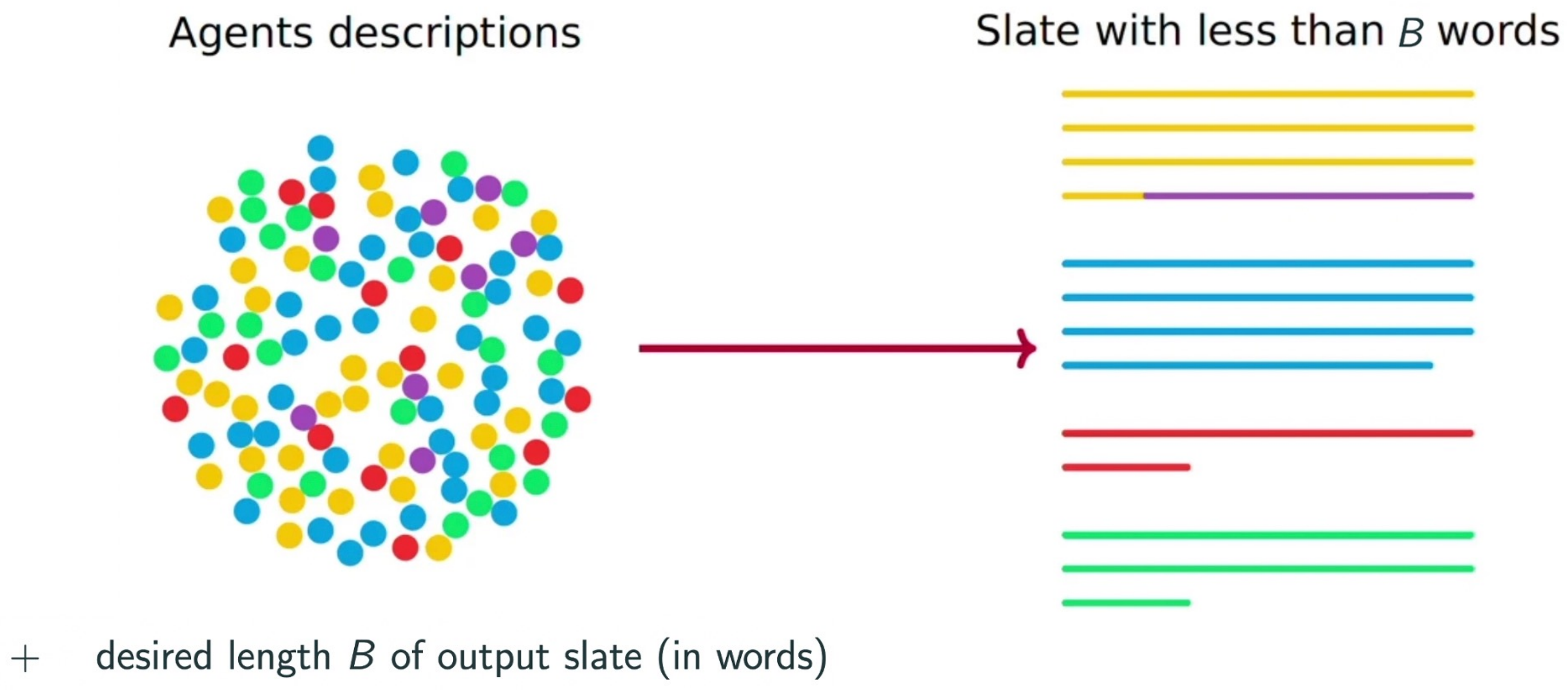
If we set b = 0 and d = 1, the (b, d)-cBJR will become cBJR. That is, we need to map each voter to only one statement, which satisfies if there are n voters who need to select statements with max B tokens, then there is no voter subset of size $|S| \geq \lceil c(\alpha) \cdot n/B \rceil$ whose maximum level of satisfaction for their matched statements is α and they all have satisfaction at least $\alpha' > \alpha$ for another statement.

Use the proportion × cost instead of n / k from BJR

[1] Boehmer N, Fish S, Procaccia A D. Generative Social Choice: The Next Generation[C]//ICML 2025.

Method: Goal

Proportional representation: x% of the users should have control over x% words in the slate.



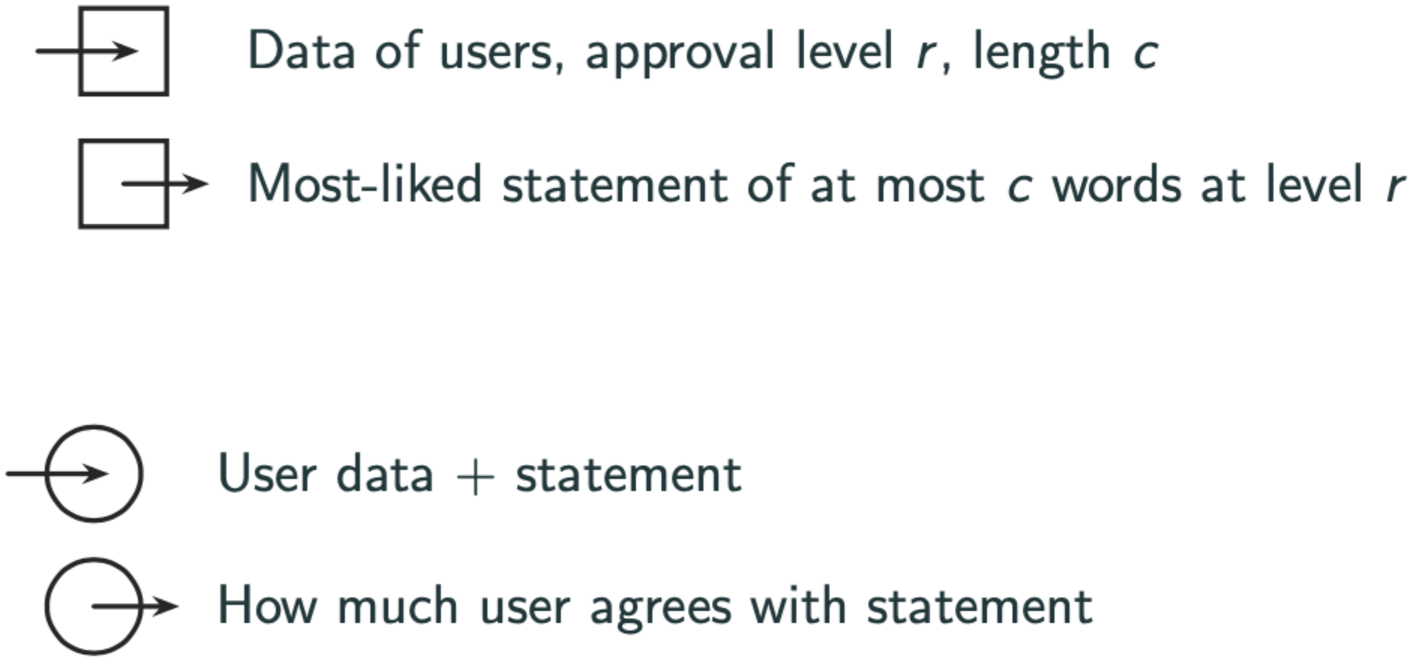
Agents descriptions

Slate with less than B words

+ desired length B of output slate (in words)

[1] Boehmer N, Fish S, Procaccia A D. Generative Social Choice: The Next Generation[C]//ICML 2025.

Method: Operators



Generative Query

Data of users, approval level r, length c

Most-liked statement of at most c words at level r

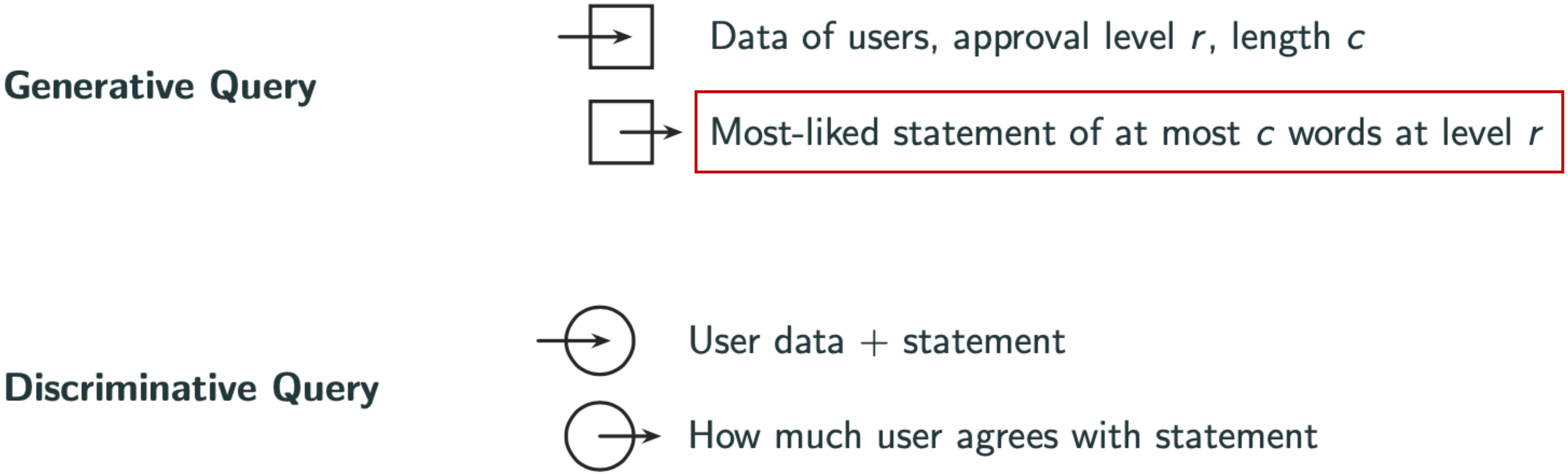
Discriminative Query

User data + statement

How much user agrees with statement

[1] Boehmer N, Fish S, Procaccia A D. Generative Social Choice: The Next Generation[C]//ICML 2025.

Method: Operators



Generative Query

Data of users, approval level r, length c

Most-liked statement of at most c words at level r

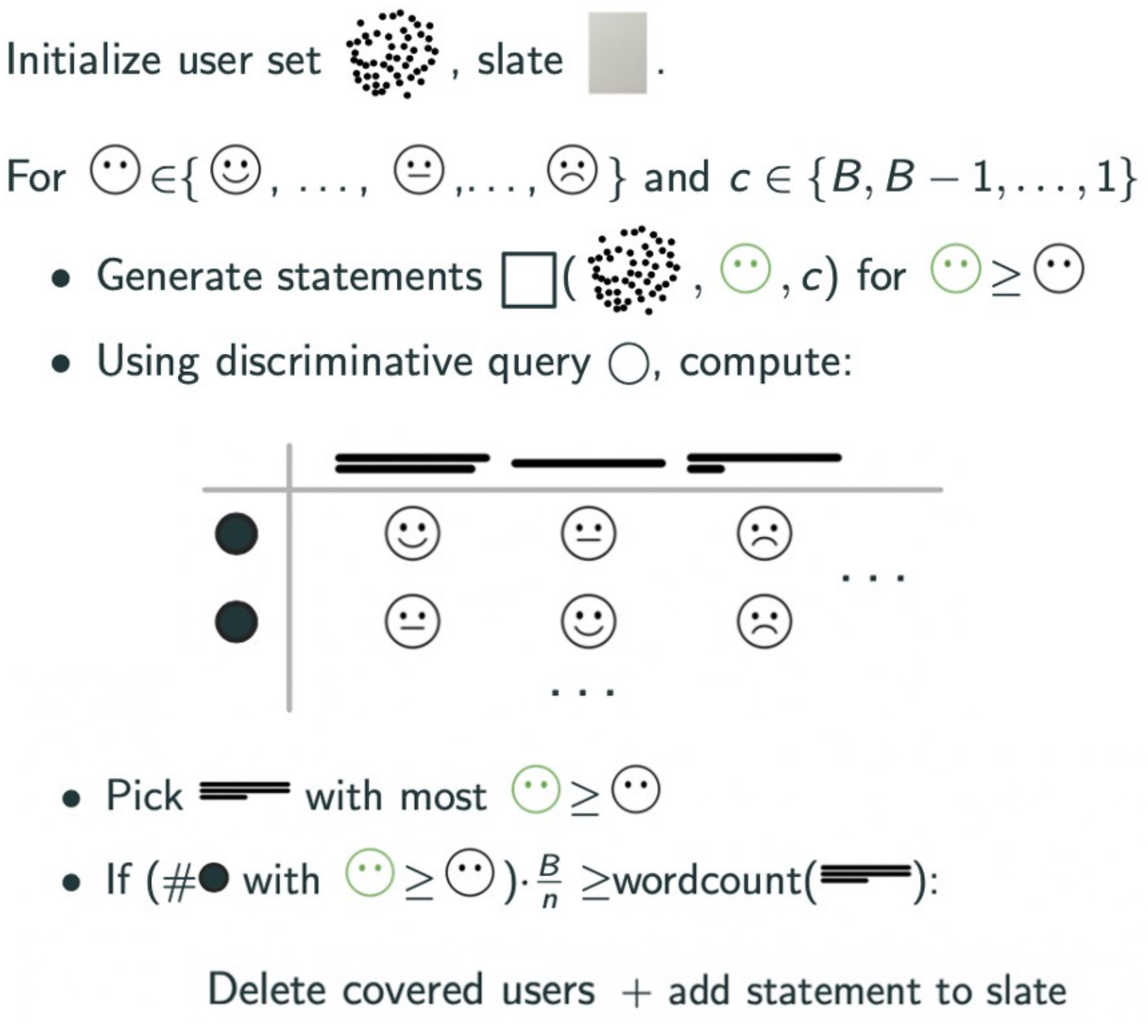
Discriminative Query

User data + statement

How much user agrees with statement

[1] Boehmer N, Fish S, Procaccia A D. Generative Social Choice: The Next Generation[C]//ICML 2025.

Method



Initialize user set ⁘, slate ▭.

For ☺ ∈ {☺, . . . , 😐, . . . , ☹} and $c \in \{B, B-1, \ldots, 1\}$

- Generate statements □(⁘, ☺, c) for ☺ ≥ ☺
- Using discriminative query ○, compute:

	═	═	═	
●	☺	😐	☹	. . .
●	😐	☺	☹	
		. . .		

- Pick ═ with most ☺ ≥ ☺
- If (#● with ☺ ≥ ☺)$\cdot\frac{B}{n}$ ≥wordcount(═):

Delete covered users + add statement to slate

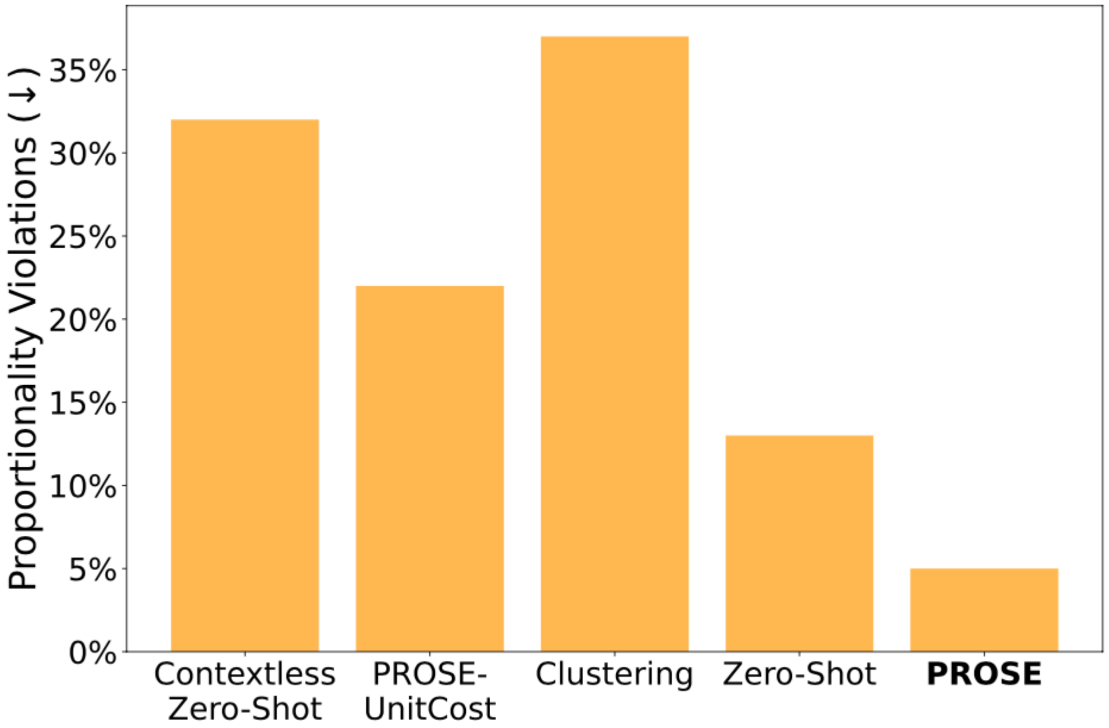
Algorithm 1 DemocraticProcess$_{C,f}(N, B, r)$

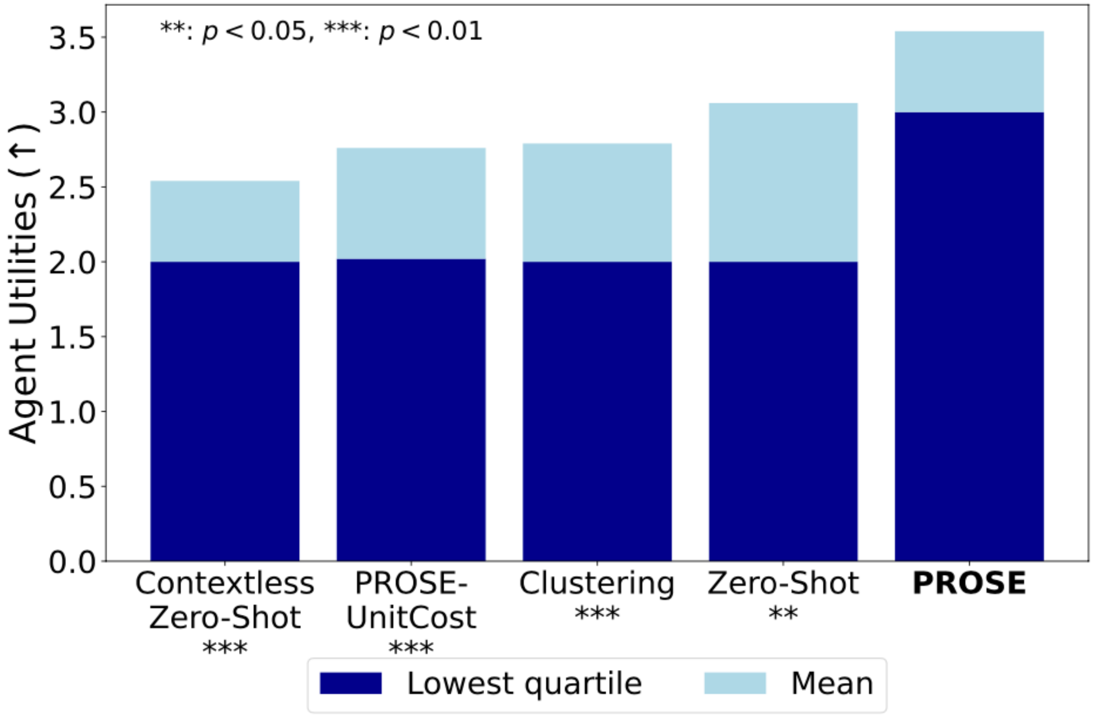
Parameters List C of cost values and function f : $[r] \rightarrow 2^{[r]}$ mapping utility values to subsets of values.

```
$S \leftarrow N, W \leftarrow \emptyset, \ell \leftarrow r$
while $\ell \geq 1$ and $S \neq \emptyset$ do
    $j \leftarrow 1$
    while $B - c(W) \geq C[j]$ and $j \leq |C|$ do
        $U \leftarrow \bigcup_{\ell' \in f(\ell)} \{\textsc{Gen}(S, \ell', C[j])\}$
        $S_\alpha \leftarrow \{i \in S \mid \textsc{Disc}(i, \alpha) \geq \ell\}$ for all $\alpha \in U$
        $\alpha^* \leftarrow \arg\max_{\alpha \in U} |S_\alpha|$
        if $|S_{\alpha^*}| \geq \lceil c(\alpha^*) \cdot n / B \rceil$ then
            $S \leftarrow S \setminus \{\lceil c(\alpha^*) \cdot n / B \rceil$ agents $i$ from $S_{\alpha^*}$ with
            highest $\textsc{Disc}(i, \alpha^*)$ return value$\}$
            $W \leftarrow W \cup \{\alpha^*\}$
        else
            $j \leftarrow j + 1$
    $\ell \leftarrow \ell - 1$
Return $W$
```

[1] Boehmer N, Fish S, Procaccia A D. Generative Social Choice: The Next Generation[C]//ICML 2025.

Experiment





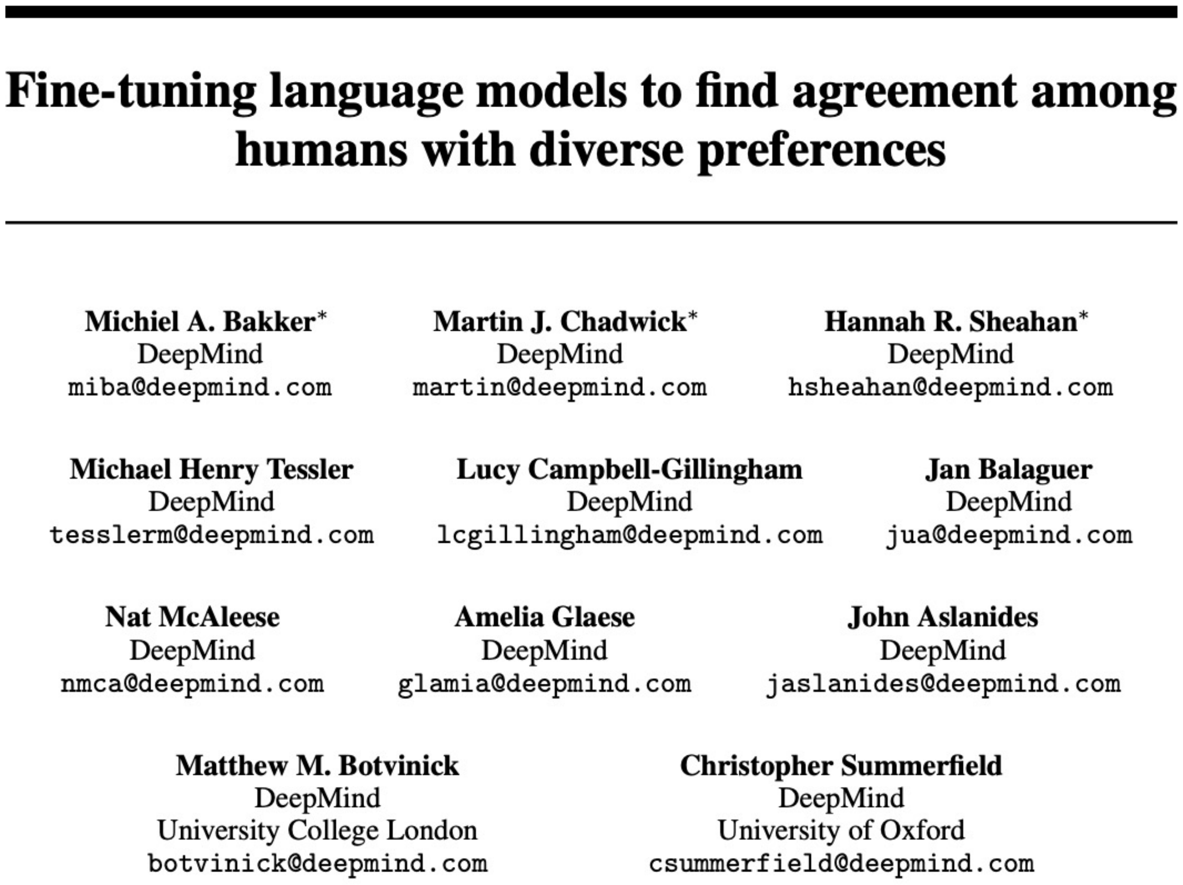
- PROSE achieves better performance of social welfare while keeps align with the opinion proportion.

[1] Boehmer N, Fish S, Procaccia A D. Generative Social Choice: The Next Generation[C]//ICML 2025.

02 Consensus within stakeholders

- Fine-tuning LLMs to Find Agreement among Humans with Diverse Preferences.
- AI can Help Humans Find Common Ground in Democratic Deliberation.
- PoliCon: Evaluating LLMs on Achieving Diverse Political Consensus Objectives.

Fine-tuning LLMs to find agreement among humans with diverse preferences



Fine-tuning language models to find agreement among humans with diverse preferences

Michiel A. Bakker* DeepMind miba@deepmind.com

Martin J. Chadwick* DeepMind martin@deepmind.com

Hannah R. Sheahan* DeepMind hsheahan@deepmind.com

Michael Henry Tessler DeepMind tesslerm@deepmind.com

Lucy Campbell-Gillingham DeepMind lcgillingham@deepmind.com

Jan Balaguer DeepMind jua@deepmind.com

Nat McAleese DeepMind nmca@deepmind.com

Amelia Glaese DeepMind glamia@deepmind.com

John Aslanides DeepMind jaslanides@deepmind.com

Matthew M. Botvinick DeepMind University College London botvinick@deepmind.com

Christopher Summerfield DeepMind University of Oxford csummerfield@deepmind.com

NeurIPS, 2022

[1] Bakker M, Chadwick M, Sheahan H, et al. Fine-tuning language models to find agreement among humans with diverse preferences[J]. NeurIPS 2022.

Motivations

- LLMs have achieved remarkable success with methods like RLHF & DPO.
- However, standard RLHF aligns models to the “average” user or a single labeler, ignoring human’s diverse preferences.
- Can we train an LLM to act as a “Social Welfare Maximizer” that finds consensus text satisfying a diverse group?

[1] Bakker M, Chadwick M, Sheahan H, et al. Fine-tuning language models to find agreement among humans with diverse preferences[J]. NeurIPS 2022.

Methodology

- The key insight is reframing the alignment problem from optimizing for individual satisfaction to optimizing for group agreement.
- It can be implemented by collecting high-quality (question, opinions, consensus) data, and use SFT + RLHF for further finetuning.

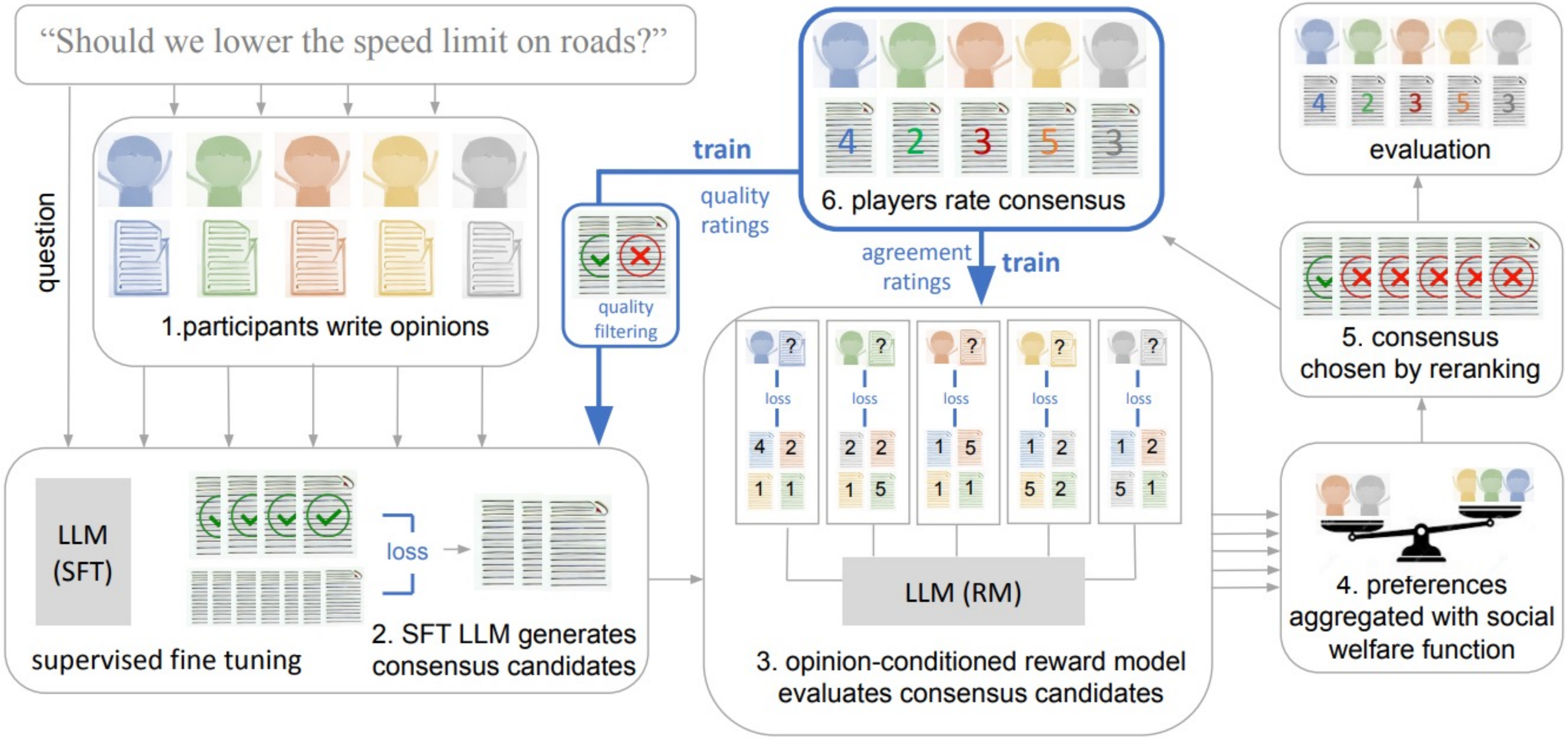
Brute-force but effective!

[1] Bakker M, Chadwick M, Sheahan H, et al. Fine-tuning language models to find agreement among humans with diverse preferences[J]. NeurIPS 2022.

Method: Data collection

Human-in-the-Loop Pipeline

•**Setup:** Small groups (3-5 people) debating UK moral issues.


"Should we lower the speed limit on roads?"
question
1.participants write opinions
LLM (SFT)
supervised fine tuning
loss
2. SFT LLM generates consensus candidates
quality filtering
train
quality ratings
6. players rate consensus
4 2 3 5 3
agreement ratings
train
loss
4 2 1 1
2 2 1 5
1 5 1 1
1 2 5 2
1 2 5 1
LLM (RM)
3. opinion-conditioned reward model evaluates consensus candidates
4. preferences aggregated with social welfare function
5. consensus chosen by reranking
evaluation

[1] Bakker M, Chadwick M, Sheahan H, et al. Fine-tuning language models to find agreement among humans with diverse preferences[J]. NeurIPS 2022.

Method: Two-Stage Iterative Training Approach

1. **Supervised Fine-Tuning (SFT):**
   - Base: Chinchilla 70B.
   - Data: Fine-tuned on high-quality consensus statements (ratings ≥ 6).
   - Role: Generates initial consensus candidates.

2. **Reward Modeling (RM):**
   - Input: Question + Single User Opinion + Candidate Statement.
   - Output: Predicts scalar agreement score for *each* individual.
   - Training: Trained on pairwise preferences via human Likert ratings.

Use RM and the social welfare function to generate high-quality data for SFT iteratively.

[1] Bakker M, Chadwick M, Sheahan H, et al. Fine-tuning language models to find agreement among humans with diverse preferences[J]. NeurIPS 2022.

Method: SWF

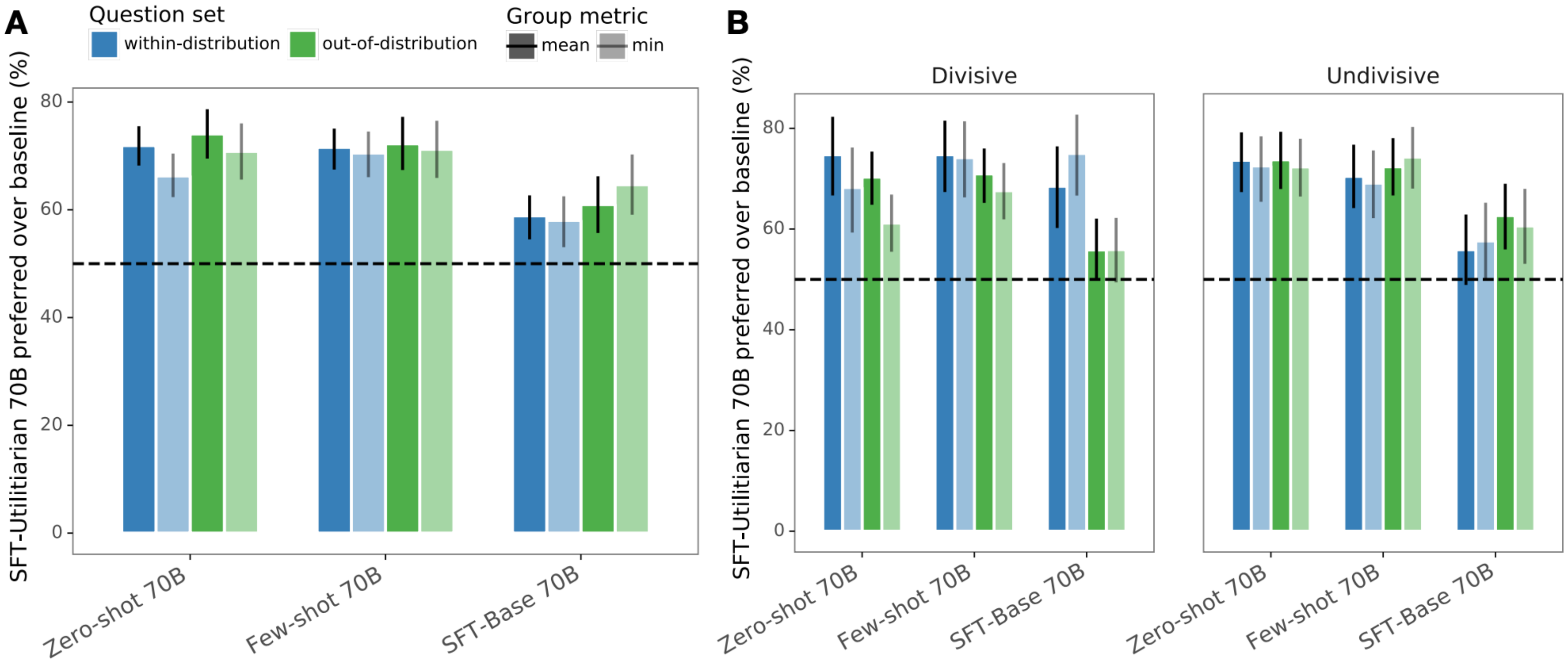
For each consensus candidate, the RM predicts agreement scores for all group members, which are then combined using different aggregation strategies:

$$W_\alpha(u_1, \ldots, u_n) = \begin{cases} \left[\frac{1}{n}\sum_{i=1}^{n} u_i^{1-\alpha}\right]^{\frac{1}{1-\alpha}} & \text{if } \alpha \geq 0, \alpha \neq 1 \\ \sqrt[n]{\prod_{i=1}^{n} u_i} & \text{if } \alpha = 1 \end{cases}$$

During training, α (inequality aversion controller) was sampled from a log-normal distribution to ensure the model could generalize across different welfare functions. Then use a mix-data approach for training.

[1] Bakker M, Chadwick M, Sheahan H, et al. Fine-tuning language models to find agreement among humans with diverse preferences[J]. NeurIPS 2022.

Experiment Findings



- SFT-Utilitarian consistently preferred over baselines on both within-distribution and out-of-distribution questions, for both divisive and undivisive topics.

[1] Bakker M, Chadwick M, Sheahan H, et al. Fine-tuning language models to find agreement among humans with diverse preferences[J]. NeurIPS 2022.

Experiment Findings: Resolving Divisive Conflicts

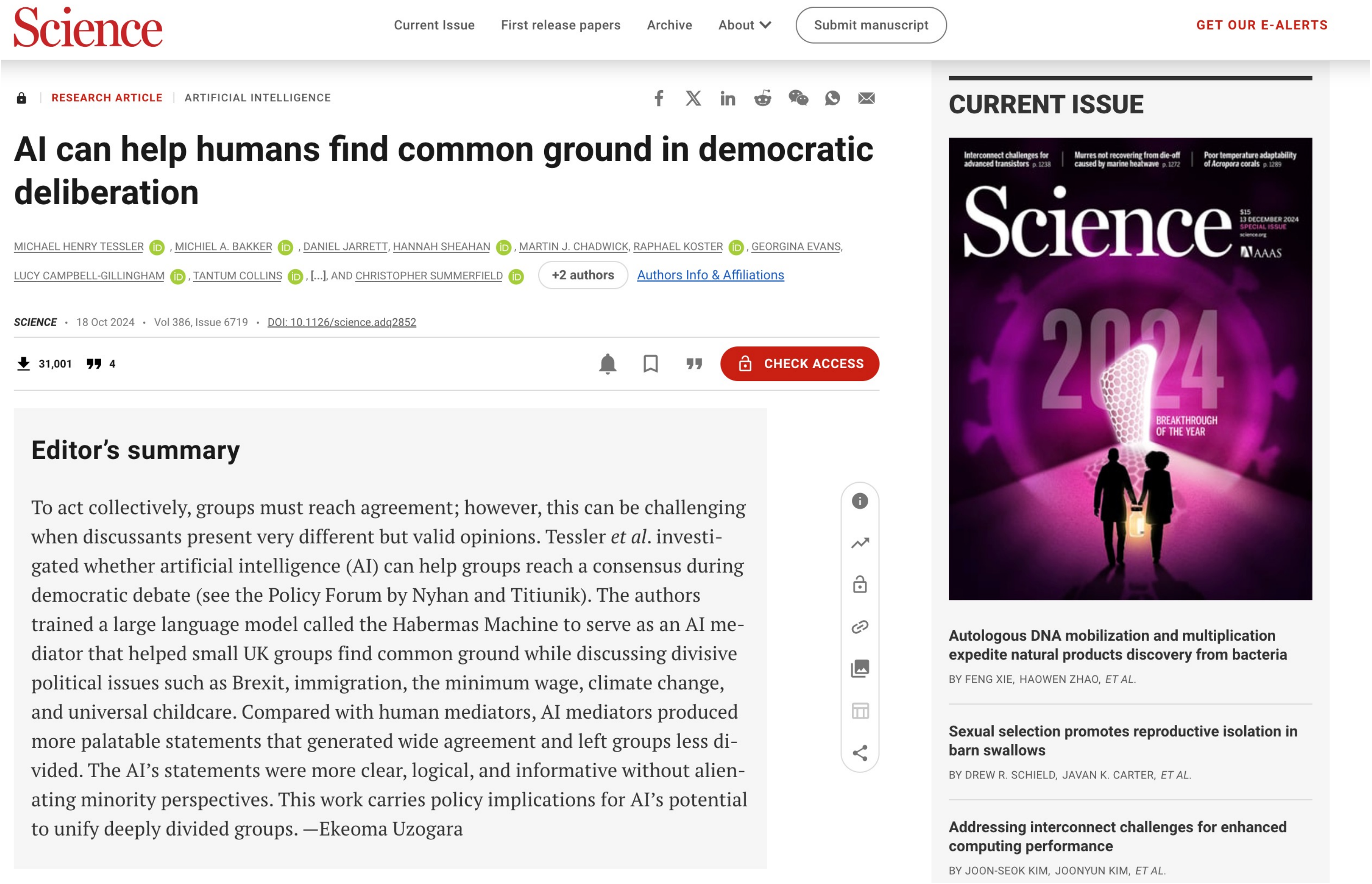
- **Depolarization:** In 65.6% of divisive rounds, AI candidates were less polarizing than original statements.

- **Unanimous Support:** Achieved “somewhat agree” or better from *all* participants in 40.8% of divisive cases.

- Super-Human Performance:
  - Best AI candidates preferred 65% of the time over human-written text.
  - AI mean quality exceeded human mean quality 78% of the time.

[1] Bakker M, Chadwick M, Sheahan H, et al. Fine-tuning language models to find agreement among humans with diverse preferences[J]. NeurIPS 2022.

AI can help humans find common ground in democratic deliberation



RESEARCH ARTICLE | ARTIFICIAL INTELLIGENCE

AI can help humans find common ground in democratic deliberation

MICHAEL HENRY TESSLER, MICHIEL A. BAKKER, DANIEL JARRETT, HANNAH SHEAHAN, MARTIN J. CHADWICK, RAPHAEL KOSTER, GEORGINA EVANS, LUCY CAMPBELL-GILLINGHAM, TANTUM COLLINS, [...], AND CHRISTOPHER SUMMERFIELD +2 authors Authors Info & Affiliations

SCIENCE • 18 Oct 2024 • Vol 386, Issue 6719 • DOI: 10.1126/science.adq2852

Editor’s summary

To act collectively, groups must reach agreement; however, this can be challenging when discussants present very different but valid opinions. Tessler *et al.* investigated whether artificial intelligence (AI) can help groups reach a consensus during democratic debate (see the Policy Forum by Nyhan and Titiunik). The authors trained a large language model called the Habermas Machine to serve as an AI mediator that helped small UK groups find common ground while discussing divisive political issues such as Brexit, immigration, the minimum wage, climate change, and universal childcare. Compared with human mediators, AI mediators produced more palatable statements that generated wide agreement and left groups less divided. The AI’s statements were more clear, logical, and informative without alienating minority perspectives. This work carries policy implications for AI’s potential to unify deeply divided groups. —Ekeoma Uzogara

CURRENT ISSUE

Autologous DNA mobilization and multiplication expedite natural products discovery from bacteria
BY FENG XIE, HAOWEN ZHAO, *ET AL.*

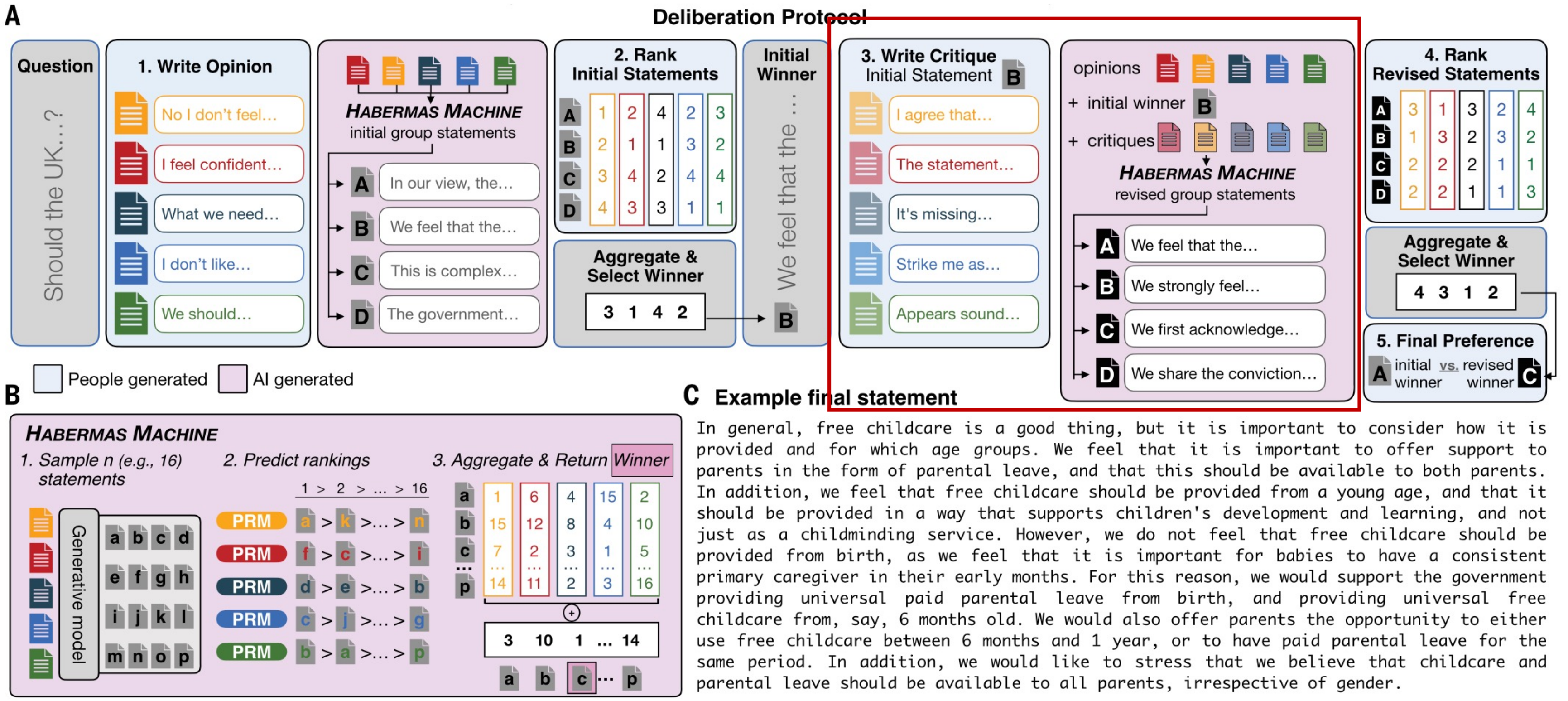
Sexual selection promotes reproductive isolation in barn swallows
BY DREW R. SCHIELD, JAVAN K. CARTER, *ET AL.*

Addressing interconnect challenges for enhanced computing performance
BY JOON-SEOK KIM, JOONYUN KIM, *ET AL.*

Science, 2024

[1] Tessler M H, Bakker M A, Jarrett D, et al. AI can help humans find common ground in democratic deliberation[J]. Science, 2024.

Method



In general, free childcare is a good thing, but it is important to consider how it is provided and for which age groups. We feel that it is important to offer support to parents in the form of parental leave, and that this should be available to both parents. In addition, we feel that free childcare should be provided from a young age, and that it should be provided in a way that supports children's development and learning, and not just as a childminding service. However, we do not feel that free childcare should be provided from birth, as we feel that it is important for babies to have a consistent primary caregiver in their early months. For this reason, we would support the government providing universal paid parental leave from birth, and providing universal free childcare from, say, 6 months old. We would also offer parents the opportunity to either use free childcare between 6 months and 1 year, or to have paid parental leave for the same period. In addition, we would like to stress that we believe that childcare and parental leave should be available to all parents, irrespective of gender.

Fig. 1. Overview of methods. (**A**) Mediated deliberation procedure. **1.** Participants, organized into small groups, privately wrote an opinion statement in response to a question. The Habermas Machine (HM) generated candidate initial group statements from the group's individual opinions. **2.** Participants ranked these initial statements. The top-ranked statement, on the basis of aggregated rankings, was returned to the group. **3.** Participants privately wrote critiques of the initial winner. The HM generated revised group statements from the group's critiques (along with the initial opinions and initial group winner). **4.** Participants ranked these revised statements, and the winner was again selected through aggregated rankings. **5.** Participants made a final preference judgement between the initial and revised winning statements. A deliberation round for a single question lasted approximately 15 min. (**B**) The HM produces a group statement through a simulated election. **1.** A generative model samples many candidate group statements. **2.** A personalized reward model produces predicted rankings for each person in the group. **3.** The top-ranked statement, on the basis of aggregated rankings, is returned. (**C**) Example top-ranked revised group opinion statement, from the virtual citizens' assembly (see SM 6 for full example, including the opinions and critiques).

[1] Tessler M H, Bakker M A, Jarrett D, et al. AI can help humans find common ground in democratic deliberation[J]. Science, 2024.

Habermas Machine: Research Questions

RQ1: Does AI-mediated deliberation help people find common ground?

RQ2: Does AI-mediated deliberation leave groups less divided?

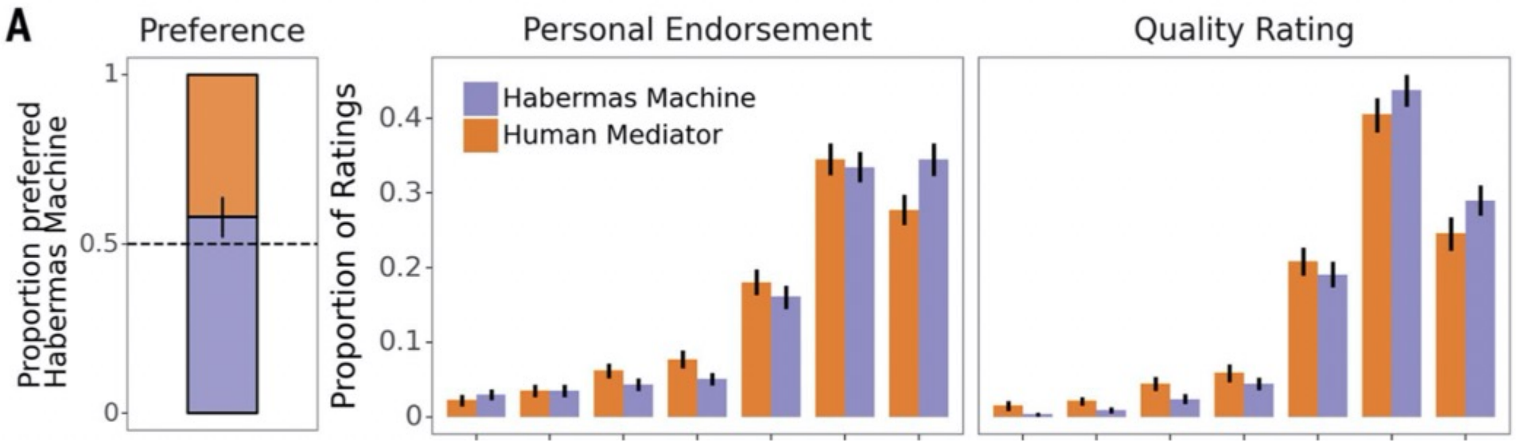
RQ3: Does the Habermas Machine represent all viewpoints equally?

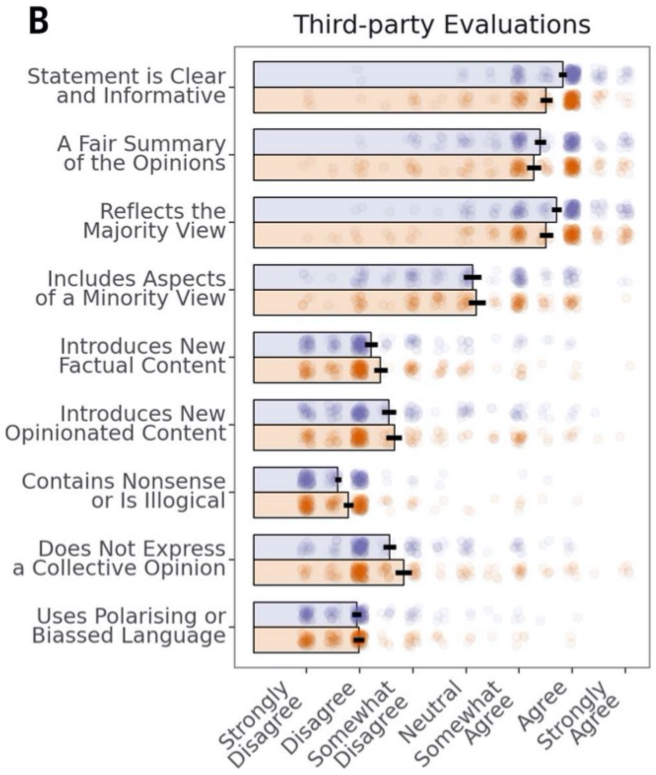
RQ4: Can AI-mediation support deliberation in a citizens’ assembly?

[1] Tessler M H, Bakker M A, Jarrett D, et al. AI can help humans find common ground in democratic deliberation[J]. Science, 2024.

Experiment Findings

RQ1: Does AI-mediated deliberation help people find common ground?



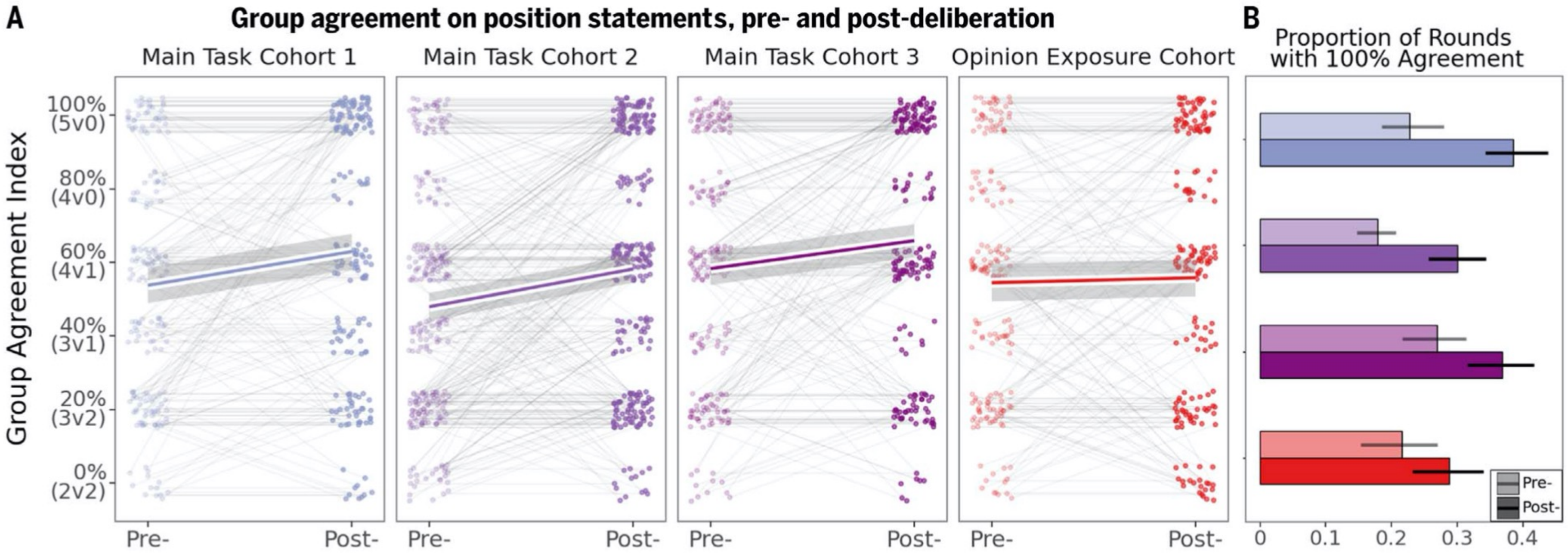


- Participants preferred AI-generated statements over human-written ones.
- AI statements were rated as more: clear, informative, fair, and reflect the majority view.

[1] Tessler M H, Bakker M A, Jarrett D, et al. AI can help humans find common ground in democratic deliberation[J]. Science, 2024.

Experiment Findings

RQ2: Does AI-mediated deliberation leave groups less divided?

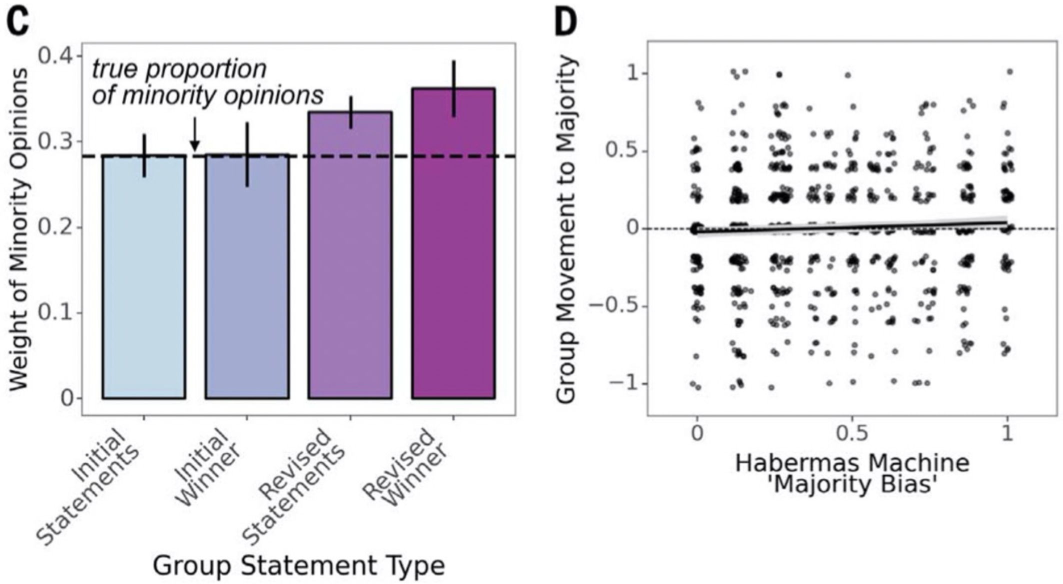


- Group agreement on the position statements increased from pre- to post-deliberation when interacting with the HM but not when participants viewed each other's opinions unmediated.
- The proportion of groups that achieved unanimous agreement increases from pre- to post-deliberation.

[1] Tessler M H, Bakker M A, Jarrett D, et al. AI can help humans find common ground in democratic deliberation[J]. Science, 2024.

Experiment Findings

RQ3: Does the Habermas Machine represent all viewpoints equally?



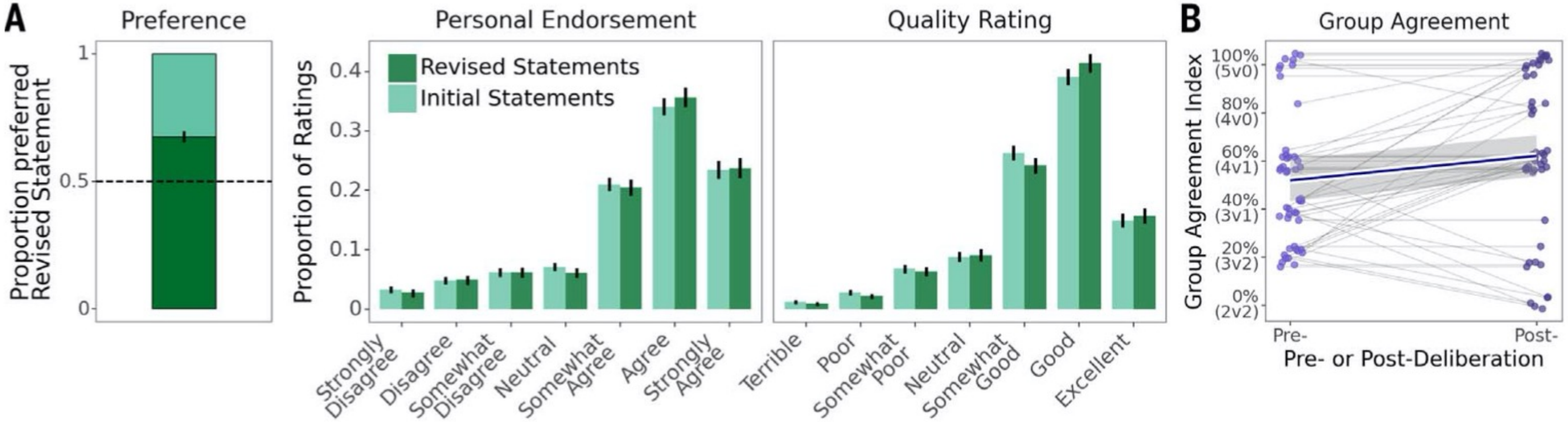
- Initial group statements proportionately represent minority opinions, and revised statements overweight them relative to the true proportion of minority opinions (dotted line).
- There is no evidence for an association between the tendency for the HM to produce majority-leaning group statements and groups moving in the direction of the majority.

[1] Tessler M H, Bakker M A, Jarrett D, et al. AI can help humans find common ground in democratic deliberation[J]. Science, 2024.

Experiment Findings

RQ4: Can AI-mediation support deliberation in a citizens' assembly?

The team conducted an experiment using a virtual citizens' assembly, recruiting a demographically representative sample of the UK population to participate in the virtual deliberation.



- Revised statements were preferred over the initial statements in the final preference judgement, and group statements were positively endorsed and high quality.
- Group agreement index increased from before to after the deliberation.

[1] Tessler M H, Bakker M A, Jarrett D, et al. AI can help humans find common ground in democratic deliberation[J]. Science, 2024.

SFT-U & Habermas Machine: Take Away

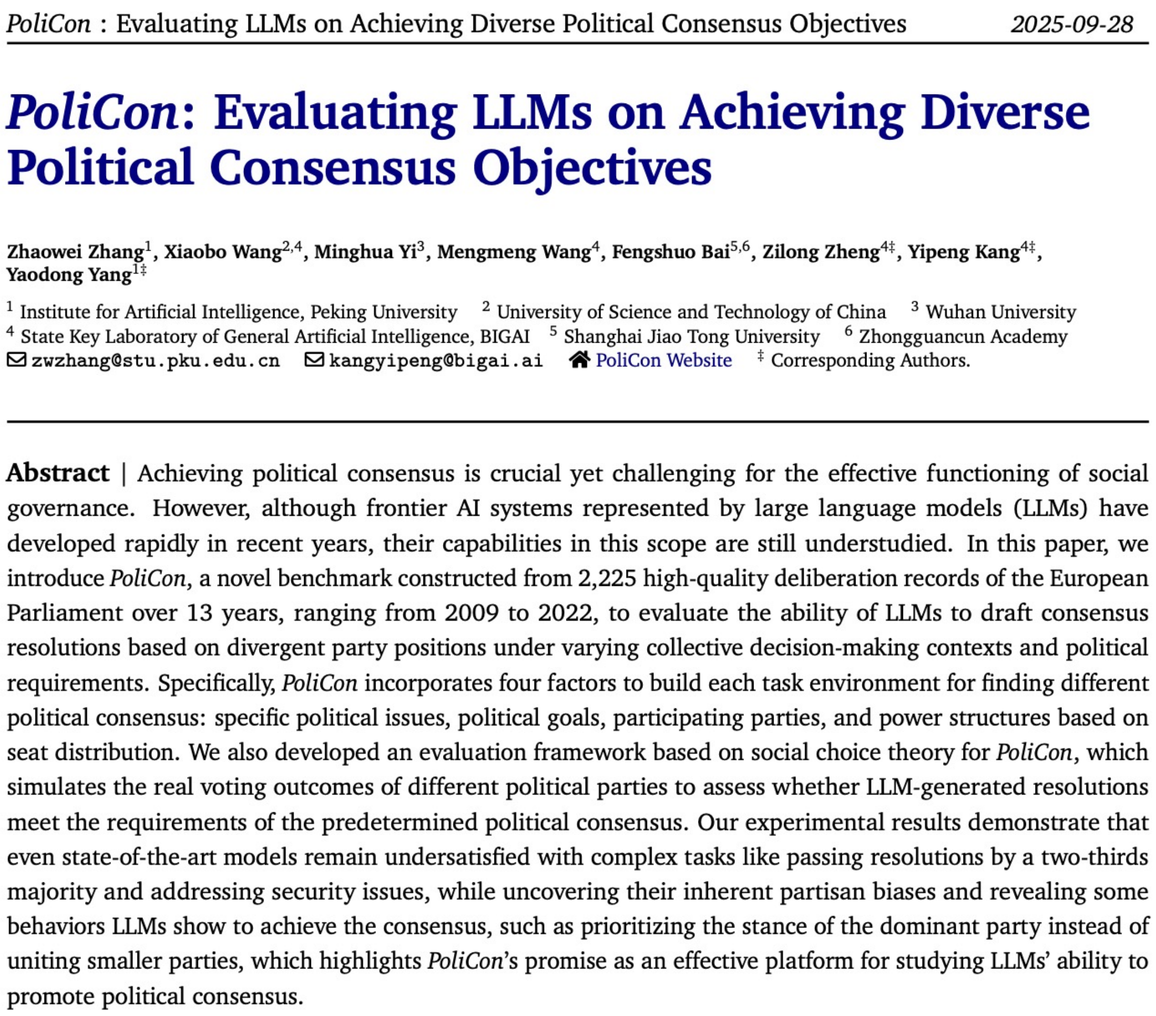
- AI has the potential to outperform human mediators, enabling fair, high-quality, and time-efficient deliberation that scales to large populations.

- AI-mediated discussion reduces polarization, helping diverse groups converge toward shared stances and significantly increasing unanimous agreement.

- Alignment is redefined as optimizing collective social welfare functions to mathematically arbitrate diverse values and balance minority critiques.

[1] Bakker M, Chadwick M, Sheahan H, et al. Fine-tuning language models to find agreement among humans with diverse preferences[J]. NeurIPS 2022.
[2] Tessler M H, Bakker M A, Jarrett D, et al. AI can help humans find common ground in democratic deliberation[J]. Science, 2024.

PoliCon: Evaluating LLMs on Achieving Diverse Political Consensus Objectives



PoliCon: Evaluating LLMs on Achieving Diverse Political Consensus Objectives

Zhaowei Zhang[1], **Xiaobo Wang**[2,4], **Minghua Yi**[3], **Mengmeng Wang**[4], **Fengshuo Bai**[5,6], **Zilong Zheng**[4‡], **Yipeng Kang**[4‡], **Yaodong Yang**[1‡]

[1] Institute for Artificial Intelligence, Peking University [2] University of Science and Technology of China [3] Wuhan University
[4] State Key Laboratory of General Artificial Intelligence, BIGAI [5] Shanghai Jiao Tong University [6] Zhongguancun Academy
✉ zwzhang@stu.pku.edu.cn ✉ kangyipeng@bigai.ai 🏠 PoliCon Website ‡ Corresponding Authors.

Abstract | Achieving political consensus is crucial yet challenging for the effective functioning of social governance. However, although frontier AI systems represented by large language models (LLMs) have developed rapidly in recent years, their capabilities in this scope are still understudied. In this paper, we introduce *PoliCon*, a novel benchmark constructed from 2,225 high-quality deliberation records of the European Parliament over 13 years, ranging from 2009 to 2022, to evaluate the ability of LLMs to draft consensus resolutions based on divergent party positions under varying collective decision-making contexts and political requirements. Specifically, *PoliCon* incorporates four factors to build each task environment for finding different political consensus: specific political issues, political goals, participating parties, and power structures based on seat distribution. We also developed an evaluation framework based on social choice theory for *PoliCon*, which simulates the real voting outcomes of different political parties to assess whether LLM-generated resolutions meet the requirements of the predetermined political consensus. Our experimental results demonstrate that even state-of-the-art models remain undersatisfied with complex tasks like passing resolutions by a two-thirds majority and addressing security issues, while uncovering their inherent partisan biases and revealing some behaviors LLMs show to achieve the consensus, such as prioritizing the stance of the dominant party instead of uniting smaller parties, which highlights *PoliCon*'s promise as an effective platform for studying LLMs' ability to promote political consensus.

ICLR 2026

[1] Zhang Z, Wang X, Yi M, et al. PoliCon: Evaluating LLMs on Achieving Diverse Political Consensus Objectives[J]. ICLR 2026.

Motivations

- Achieving political consensus is crucial yet challenging for the effective functioning of social governance.
- However, although frontier AI systems represented by LLMs have developed rapidly in recent years, their capabilities in this scope are still understudied.
- We need a comprehensive benchmark to provide a research platform for investigating this field.

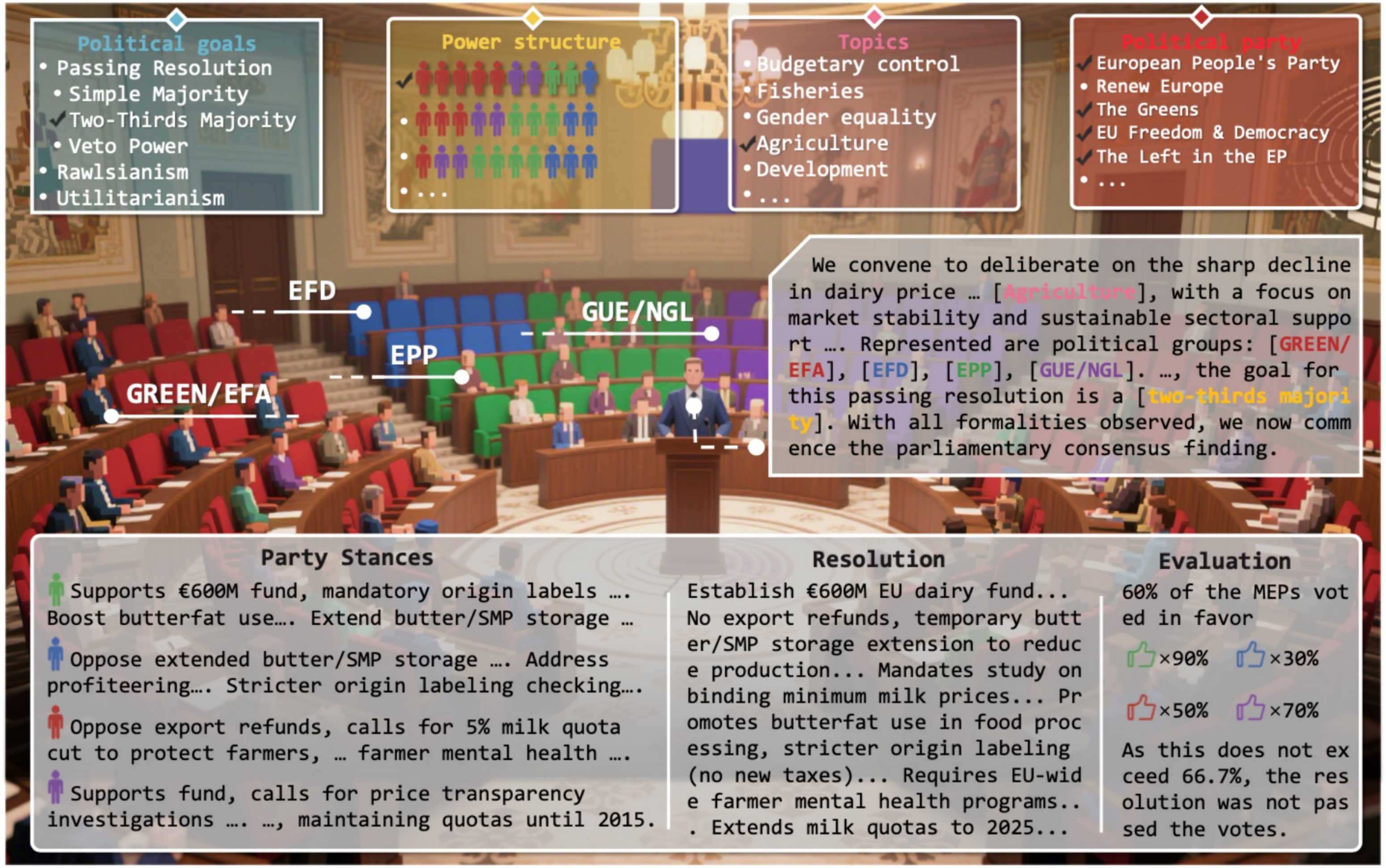
[1] Zhang Z, Wang X, Yi M, et al. PoliCon: Evaluating LLMs on Achieving Diverse Political Consensus Objectives[J]. ICLR 2026.

Contributions

- We introduce PoliCon, a benchmark constructed from 2,225 high-quality deliberation records of the European Parliament over 13 years.
- We define the components of finding political consensus and set relevant tasks.
- We develop an open-ended evaluation framework that can simulate the proportion of MEPs who vote in favor in each party.

[1] Zhang Z, Wang X, Yi M, et al. PoliCon: Evaluating LLMs on Achieving Diverse Political Consensus Objectives[J]. ICLR 2026.

Components of Finding Political Consensus



We convene to deliberate on the sharp decline in dairy price … [Agriculture], with a focus on market stability and sustainable sectoral support …. Represented are political groups: [GREEN/EFA], [EFD], [EPP], [GUE/NGL]. …, the goal for this passing resolution is a [two-thirds majority]. With all formalities observed, we now commence the parliamentary consensus finding.

Party Stances

Supports €600M fund, mandatory origin labels …. Boost butterfat use…. Extend butter/SMP storage …

Oppose extended butter/SMP storage …. Address profiteering…. Stricter origin labeling checking….

Oppose export refunds, calls for 5% milk quota cut to protect farmers, … farmer mental health ….

Supports fund, calls for price transparency investigations …. …, maintaining quotas until 2015.

Resolution

Establish €600M EU dairy fund... No export refunds, temporary butter/SMP storage extension to reduce production... Mandates study on binding minimum milk prices... Promotes butterfat use in food processing, stricter origin labeling (no new taxes)... Requires EU-wide farmer mental health programs... Extends milk quotas to 2025...

Evaluation

60% of the MEPs voted in favor

👍×90% 👍×30%

👍×50% 👍×70%

As this does not exceed 66.7%, the resolution was not passed the votes.

An example scenario in PoliCon. In each task, PoliCon builds a collective decision-making environment with varying political goals, power structures, issues, and participating parties. The tested LLM then attempts to achieve a consensus resolution based on these setups and the divergent party positions. The outcome is evaluated first via a simulated vote and then mapped to a quantitative score according to the specific environment setting by PoliCon's evaluation framework.

- Political Goals.
- Power Structure.
- Topics.
- Stakeholders.

[1] Zhang Z, Wang X, Yi M, et al. PoliCon: Evaluating LLMs on Achieving Diverse Political Consensus Objectives[J]. ICLR 2026.

Dataset details

Political Goals

1. Voting Mechanism
   - Simple Majority (SM).
   - Two-Thirds Majority (2/3M).
   - Veto Power (VP).
2. Social Choice Theory
   - Rawlsianism (Rawls).
   - Utilitarianism (Util).

Topics

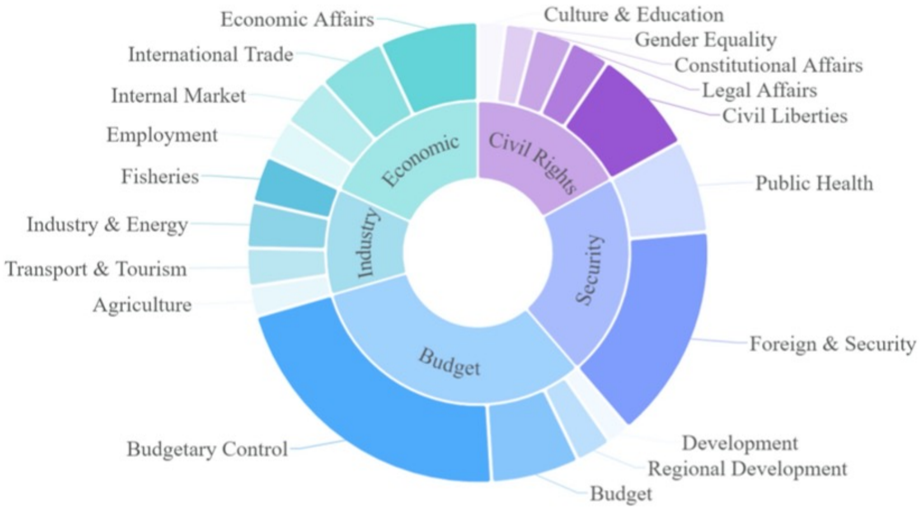
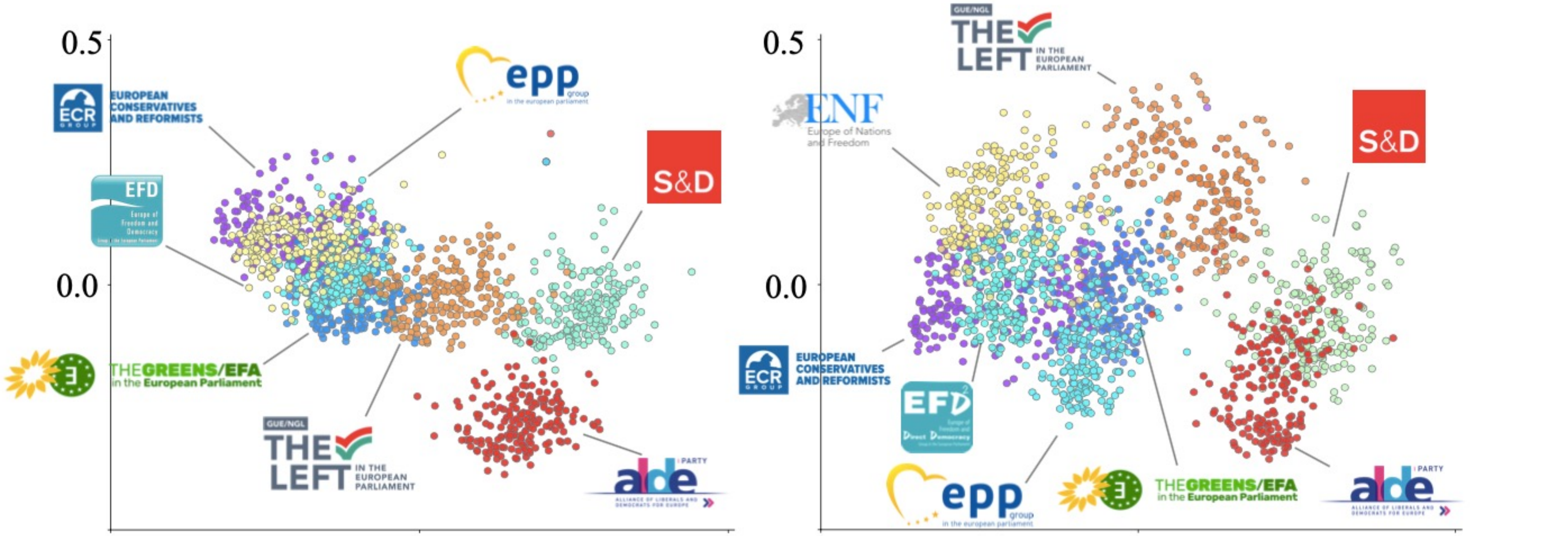


Figure 2: The 5 coarse-grained and 19 fine-grained topic categories of issues in *EuroCon*, whose definitions can be found in Appendix B.1. The shade of the color indicates the proportion of the fine-grained topic within the coarse-grained topic; the darker the color, the higher the proportion.

[1] Zhang Z, Wang X, Yi M, et al. PoliCon: Evaluating LLMs on Achieving Diverse Political Consensus Objectives[J]. ICLR 2026.

Dataset details: Stakeholders

The stances of stakeholders for each issue is diverse enough.



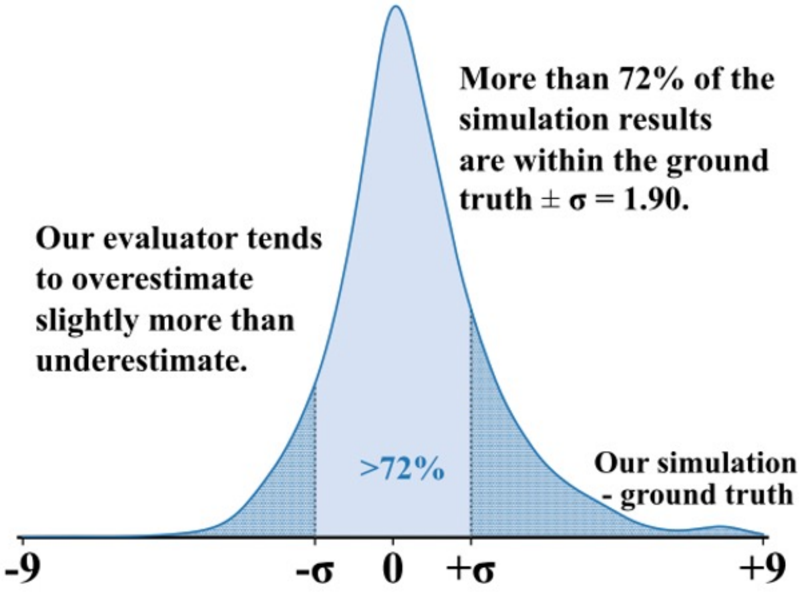
(a) Party Stances for the 7th Parliament

(b) Party Stances for the 8th Parliament

Figure 3: Semantic representation distribution of party stances (indicated by their symbols) in the 7th (2009-2014) and 8th (2014-2019) terms of the European Parliament in *EuroCon*.

[1] Zhang Z, Wang X, Yi M, et al. PoliCon: Evaluating LLMs on Achieving Diverse Political Consensus Objectives[J]. ICLR 2026.

Experiment: Effectiveness of Evaluator



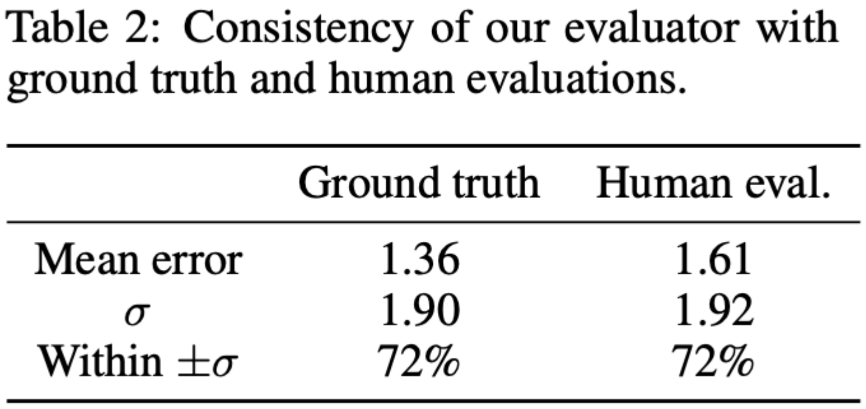


Table 2: Consistency of our evaluator with ground truth and human evaluations.

	Ground truth	Human eval.
Mean error	1.36	1.61
σ	1.90	1.92
Within $\pm\sigma$	72%	72%

- Our evaluator shows high consistency with both real-world voting and human annotators.
- Interestingly, the results of ground truth voting results is nearly the same with humans.

[1] Zhang Z, Wang X, Yi M, et al. PoliCon: Evaluating LLMs on Achieving Diverse Political Consensus Objectives[J]. ICLR 2026.

Experiment Findings

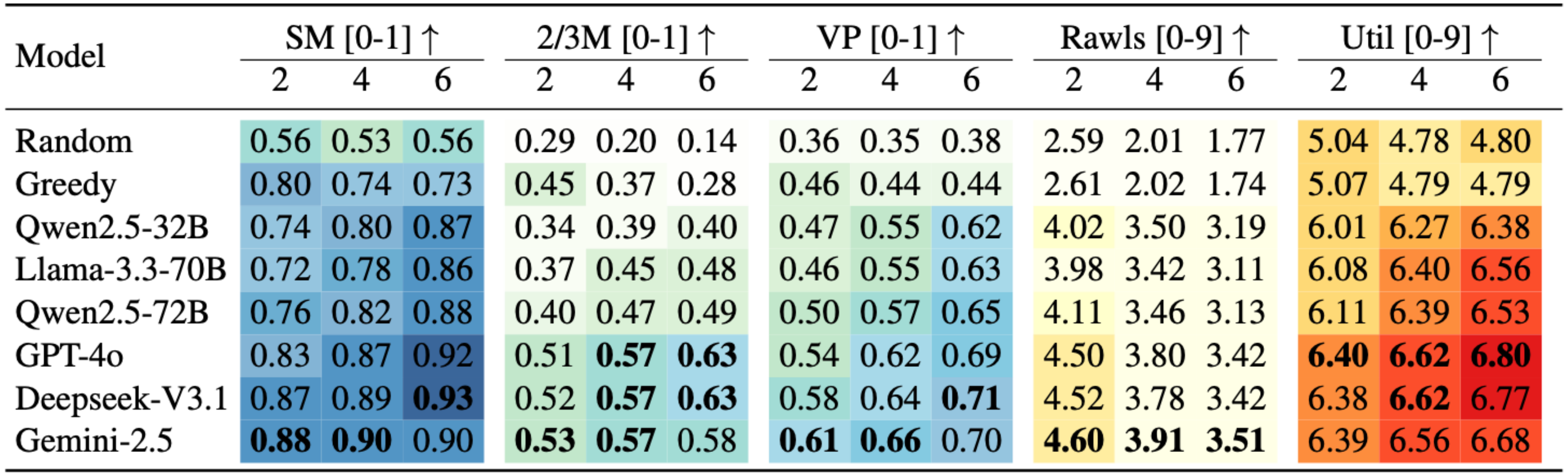


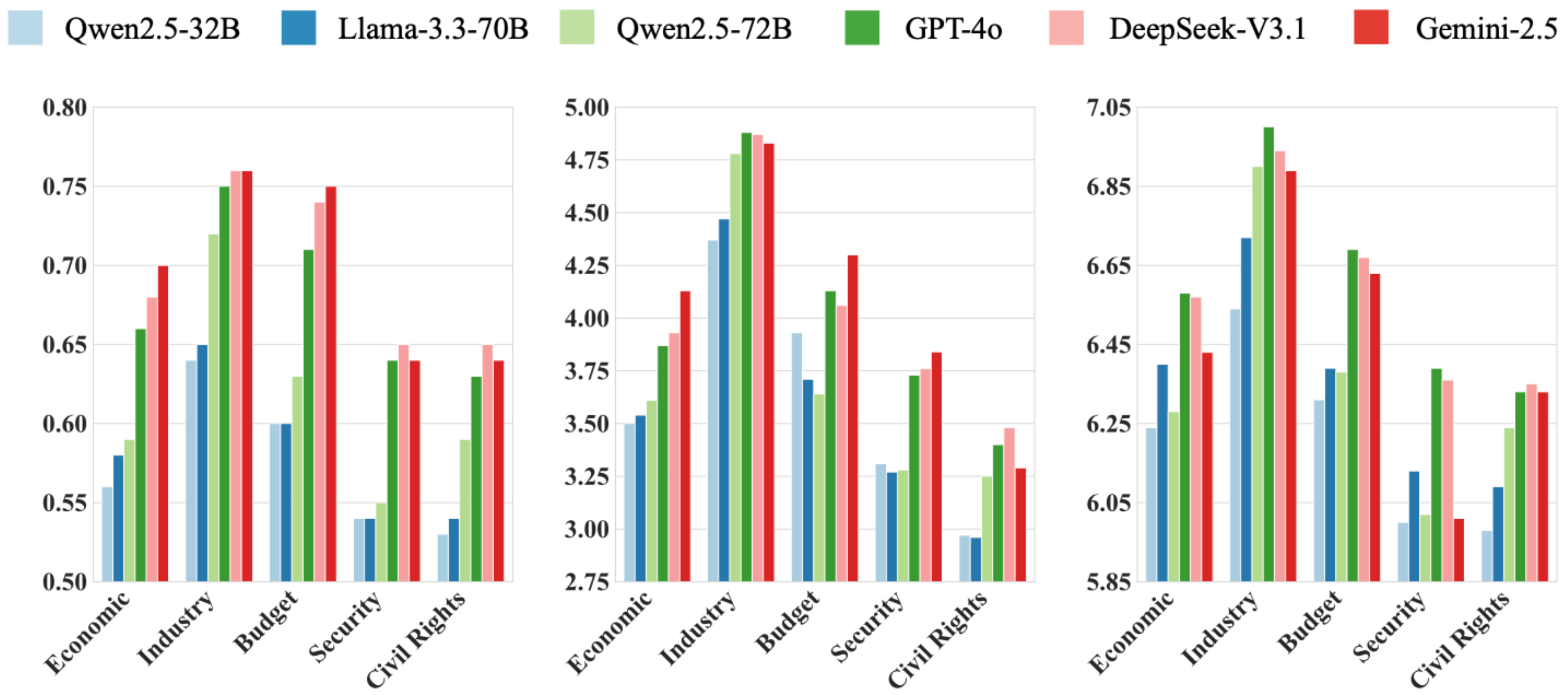
Table 1: Performance of different LLMs on *PoliCon*. The values in square brackets indicate the range of each metric, and all metrics follow the principle that higher values are better. The background color of the table cells deepens as the performance improves. The blue color scheme represents metrics in the 0-1 range, while the red color scheme represents metrics in the 0-9 range.

Model	SM [0-1] ↑			2/3M [0-1] ↑			VP [0-1] ↑			Rawls [0-9] ↑			Util [0-9] ↑		
	2	4	6	2	4	6	2	4	6	2	4	6	2	4	6
Random	0.56	0.53	0.56	0.29	0.20	0.14	0.36	0.35	0.38	2.59	2.01	1.77	5.04	4.78	4.80
Greedy	0.80	0.74	0.73	0.45	0.37	0.28	0.46	0.44	0.44	2.61	2.02	1.74	5.07	4.79	4.79
Qwen2.5-32B	0.74	0.80	0.87	0.34	0.39	0.40	0.47	0.55	0.62	4.02	3.50	3.19	6.01	6.27	6.38
Llama-3.3-70B	0.72	0.78	0.86	0.37	0.45	0.48	0.46	0.55	0.63	3.98	3.42	3.11	6.08	6.40	6.56
Qwen2.5-72B	0.76	0.82	0.88	0.40	0.47	0.49	0.50	0.57	0.65	4.11	3.46	3.13	6.11	6.39	6.53
GPT-4o	0.83	0.87	0.92	0.51	**0.57**	**0.63**	0.54	0.62	0.69	4.50	3.80	3.42	**6.40**	**6.62**	**6.80**
Deepseek-V3.1	0.87	0.89	**0.93**	0.52	**0.57**	**0.63**	0.58	0.64	**0.71**	4.52	3.78	3.42	6.38	**6.62**	6.77
Gemini-2.5	**0.88**	**0.90**	0.90	**0.53**	**0.57**	0.58	**0.61**	**0.66**	0.70	**4.60**	**3.91**	**3.51**	6.39	6.56	6.68

- Deepseek-V3.1 and Gemini-2.5 perform the best, and the performance is relatively related to the model parameters.
- Even state-of- the-art models remain undersatisfied with complex tasks like passing resolutions by a two-thirds majority and achieving Rawlsianism.

[1] Zhang Z, Wang X, Yi M, et al. PoliCon: Evaluating LLMs on Achieving Diverse Political Consensus Objectives[J]. ICLR 2026.

Experiment Findings



(a) PR Analysis (b) Rawls Analysis (c) Util Analysis

Figure 6: The average results of the six evaluated LLMs of the five coarse-grained topics on passing resolution (PR, including SM, 2/3M, and VP), Rawls, and Util political goals.

- LLMs perform bad on harder topics like security and civil rights.

[1] Zhang Z, Wang X, Yi M, et al. PoliCon: Evaluating LLMs on Achieving Diverse Political Consensus Objectives[J]. ICLR 2026.

Experiment Findings

We also analyze whether a common strategy exists for LLMs to achieve political consensus under various power structures.

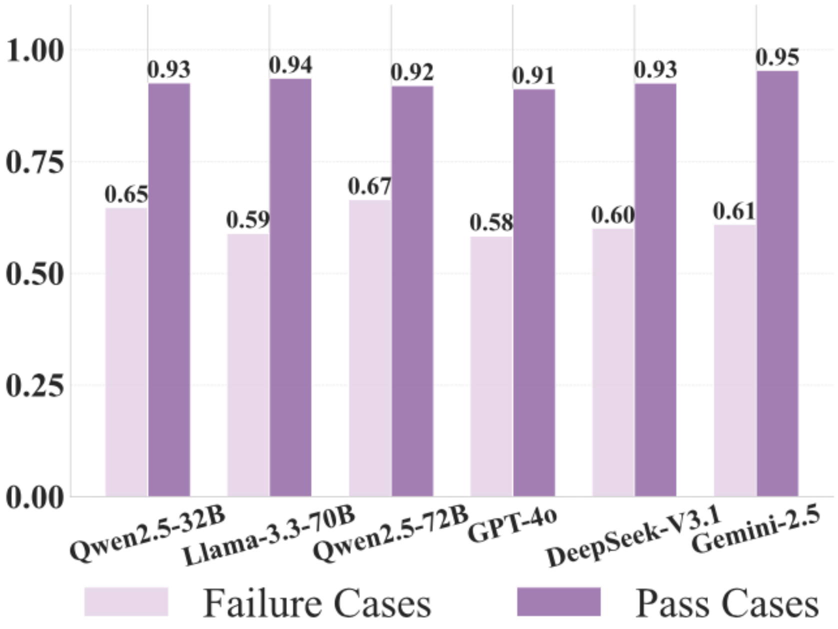


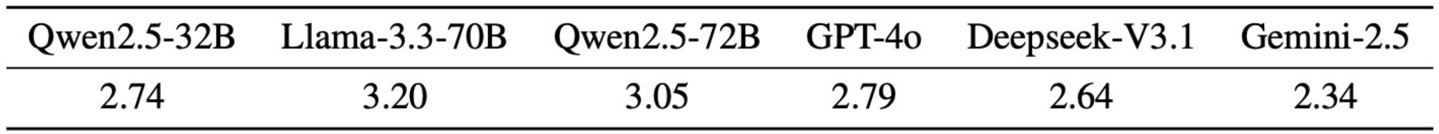
Figure 5: The average contribution ratio of the largest party to other parties in failed and passed cases across SM and 2/3M.

LLMs tend to prioritizing the stance of the dominant party instead of uniting smaller parties.

[1] Zhang Z, Wang X, Yi M, et al. PoliCon: Evaluating LLMs on Achieving Diverse Political Consensus Objectives[J]. ICLR 2026.

Experiment Findings

Table 3: Scores of different LLMs regarding the degree of bias between political parties.



Qwen2.5-32B	Llama-3.3-70B	Qwen2.5-72B	GPT-4o	Deepseek-V3.1	Gemini-2.5
2.74	3.20	3.05	2.79	2.64	2.34

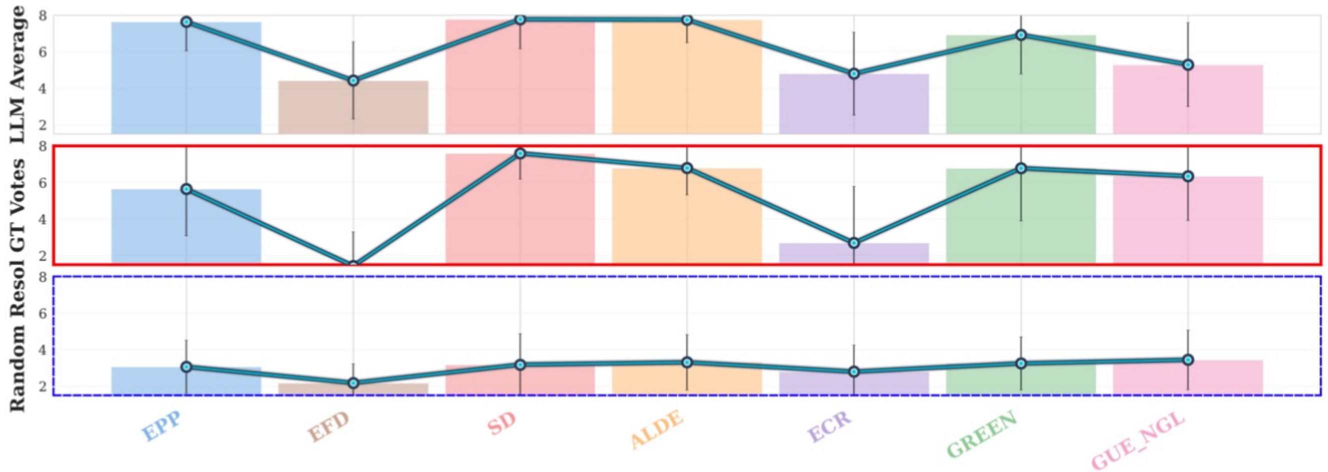


Figure 7: Partisan bias of the tested LLMs. (Top) Average scores from the tested LLMs on different parties. (Middle) Ground truth votes of different parties. (Bottom) Scores of random assignment.

- Generally, as LLMs' bias towards parties decreases, the corresponding performance increases.
- Scores across different parties still resemble the distribution of real-world voting results, which indicates that the tested models are somehow influenced by the intrinsic party bias.

[1] Zhang Z, Wang X, Yi M, et al. PoliCon: Evaluating LLMs on Achieving Diverse Political Consensus Objectives[J]. ICLR 2026.

PoliCon: Takeaway

- LLMs retain partisan biases mirroring real world, which persist even when prompts explicitly randomize power structures or party roles.

- Models tend to “shortcut” consensus by exclusively prioritizing the dominant party's stance, frequently neglecting minority interests and failing to build broader coalitions.

- LLMs struggle significantly with complex tasks like strict majority voting or Rawlsian fairness.

[1] Zhang Z, Wang X, Yi M, et al. PoliCon: Evaluating LLMs on Achieving Diverse Political Consensus Objectives[J]. ICLR 2026.

Conclusions

- AI is Reshaping “Social Choice” Mechanisms, Shifting from Finite Options to Generative Consensus.

- AI Mediators Outperform Humans in Reducing Group Polarization and Finding Common Ground.

- Despite Immense Potential, Current Models Still Exhibit Bias and Limitations in Complex Consensus Finding Scenarios.

[1] Zhang Z, Wang X, Yi M, et al. PoliCon: Evaluating LLMs on Achieving Diverse Political Consensus Objectives[J]. ICLR 2026.

Discussions

1. **Dynamic Deliberation**

   Generate the consensus in a human-in-the loop deliberation process.

2. **Interaction Cost Optimization**

   Use a preference approximation, and calculate the theoretical upper regret bound.

3. **Unbiased Persuasion**

   In reality, a consensus is usually accepted by "persuasion" instead of fully agree the whole context.

Thank You!

Q&A