Filter bubbles in an agent-based model where agents update their worldviews with LLMs

Jan Lorenz[1], Erkan Gunes[1]

[1]Constructor University, Campus Ring 1, 28759 Bremen, Germany

Abstract

We present the concept of reproducing an agent-based model of opinion dynamics by replacing an abstract numerical opinion space with large-language models (LLMs), capable of generating human-like statements in natural language, to decide how agents change their opinions. We aim to validate by computer simulation whether the mechanism of filter bubble creation through social media will also show up in this more realistic setting. We also discuss different ways to implement parts of the original model using LLM agents and the potential challenges we may face integrating LLMs into an agent-based model of opinion dynamics.

Keywords

agent-based models, large-language models, filter bubbles, opinion dynamics, political space, polarization

1. Introduction

Opinion dynamics is one of the main topics in agent-based modeling of sociopolitical systems [1, 2]. Models of opinion dynamics have been used to understand the mechanisms of societal radicalization and polarization, and the emergence of filter bubbles in various ways [3, 4].

A shortcoming of opinion dynamics models has always been that something as complex as the opinion of a human being has to be boiled down to one number (or a vector of a few numbers) in an abstract numerical opinion space such as the left-right political continuum. Of course, numerical opinion spaces also have applied use. Researchers measure the attitudes of individuals with specific or large-scale international (representative) surveys, where individuals self-assess their attitudes on (standardized) scales, e.g. with numbers from 0 to 10.

In addition, actors in political discourse in the real world use spatial metaphors such as ‘an actor has moved its opinion towards the opinion of another’ or ‘a political actor has moved to the right, left, or center’. So, political science has developed several methods for positioning political actors and texts in low-dimensional numerical opinion spaces. These methods range from expert assessments using theory-driven classification to data-driven approaches measuring positions in latent dimensions of higher-dimensional spaces using roll-call, co-voting data, survey data, or texts using dimensionality reduction and natural language processing.

The well-known models of continuous opinion dynamics (such as the bounded confidence model [5, 6]) make use of such abstract one-(or low-)dimensional opinion spaces and define how individuals change their opinion on that scale when confronted with the opinions of others. However, an attitude change does not occur based on the exchange of numerical numbers but through verbalization of arguments [7], persuasive messages, and emotional emphasis, whether written or spoken language, or other media such as images. While it is plausible that the attitudes of individuals can be mapped reasonably well into a low-dimensional numerical space for descriptive purposes, it is less realistic that individual attitude change can be well operationalized solely based on just the exchange of numbers.

Large Language Models (LLMs) provide a new opportunity to build a new generation of opinion dynamics models. LLMs trained on large bodies of text seem to have embedded the meaning of the antagonisms and compatibilities of opinions as stated in written texts.

ESWC 2025 Workshops and Tutorials Joint Proceedings

jlorenz@constructor.university (J. Lorenz); erkangunes39@gmail.com (E. Gunes)
https://janlo.de (J. Lorenz); https://erkangunes.github.io/ (E. Gunes)
0000-0002-5547-7848 (J. Lorenz); 0009-0005-8106-3314 (E. Gunes)

In this paper, we rebuild the filter bubble model of [8] using agents linking to textual statements. In that way, simulated agents build their worldview as a collection of textual statements instead of linking to abstract news items positioned in a numerical opinion space.

2. The filter bubble model

The filter bubble model of Geschke et al. [8] builds on the concept that three types of filters are operational for the formation of filter bubbles: (i) cognitive filters in the brains of humans which restrict what comes to their attention and consideration, (ii) social filters like the follow links on social media determining what comes to their attention, and (iii) algorithmic filters as implemented in social media compiling their news feed. Here we focus on the model's most interesting systemic insight, which we want to validate using the latent opinion structure encoded in LLMs. To that end, we focus on a simple cognitive filter building on opinion discrepancy and a simulated social network that individuals can use to post opinion content they have in memory. That means, algorithmic filters are not part of the analysis. The core insight of the model is that a social network can be essential for the emergence of filter bubbles. However, the social network triggers the emergence of filter bubbles not by filtering information but by enabling the reposting of existing opinions. The simplified version of the model in NetLogo [9] is provided by Lorenz [10]. The model can be run directly in the browser using NetLogoWeb.

The original filter bubble model has two types of "agents": Individuals and bits of information. Both of these agents are located in a two-dimensional numerical opinion space. All individuals are initialized at random locations. Bits of information are initially absent and appear anew, one in each time step. These new bits of information automatically come to the attention of all individuals (like mass media). That means each individual checks if the information is close in opinion space (details below). If the information is close, the individual integrates the bit into their memory. Technically, 'integration' means creating a link between the individual and the bit of information. In that way, individuals link to more and more bits of information.

Additionally, individuals can perform three more actions: (1) They reposition their opinion by moving to the barycenter of all the bits of information they have in memory (that means the arithmetic mean on each dimension). They do this whenever their bits of information in memory change. (2) They forget one random bit of information whenever their memory is full (the maximal memory is a model parameter). Technically, that means deleting a random link. The baseline value in the simulation is a memory of twenty bits of information. (3) Post one bit of information (from memory) to their contacts in the social network. Social posting is a binary parameter in the model. So, we can run simulations with or without posting on a social network. The social network is initialized with bidirectional friend links with an average number of friends being twenty. Details are in the original paper. When an individual receives a posting from a friend over the social network it has to decide about its integration in memory. The technical procedure is the same as for the integration of new bits of information.

What remains to explain is how individuals decide if they integrate a bit of information when it comes to their attention (either through the central news every time step or through reposting of a friend). Here the cognitive filter applies: An individual integrates a bit of information only if the distance from their current position in opinion space is low. The core parameter is the latitude of acceptance. The baseline value is 0.3 while the opinion space is from -1 to +1 on both dimensions. Euclidean distance is used. More details are in the original paper. In essence, the mechanism is similar to bounded confidence or motivated reasoning [2].

The timeline of a simulation run is as follows: Upon initialization individuals and their social networks are created. In each time step, one new bit of information is created and all individuals check to integrate it. When social posting is enabled also each individual posts one of the bits in memory to their friends. That means when social posting is possible each individual is exposed to one new bit of information and on average twenty already existing ones their friends repost.

Throughout a simulation run, we are interested in monitoring the existence of filter bubbles. Figure 1 helps in understanding the definition used. Filter bubbles tend to exist when the mean distance to

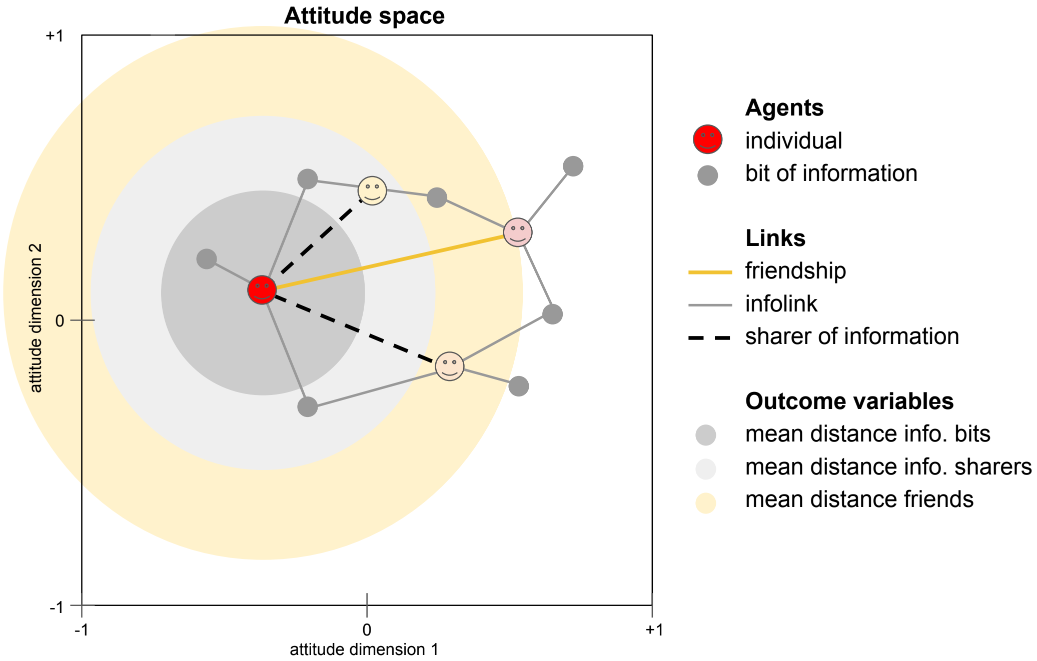


Figure 1: Example of individuals and bits of information in an opinion space (here called attitude space, based on attitude and opinion being considered synonyms here). Friendship links are created upon initialization while info-links are created when the bit of information comes to the attention of the individual and when the distance is close. The link 'sharer of information' is a latent construct to measure the mean distance of info sharers. It does not play an active role in the simulation.

info-sharers is less than the mean distance to the information in memory. Otherwise, people share at least some bits of information with other individuals who have very different opinions. This indicates that at least bridges between opinion bubbles exist.

Figure 2 shows the outcome of two example runs under identical parameters. One without and one with social posting. It is visible that social posting leads to pronounced filter bubbles. In contrast, the absence of social posting leads to some clustering, while all individuals are indirectly connected through shared information.

The generative mechanism of the emergence of filter bubbles is subtle: Whenever a slightly higher density of individuals appears in opinion space by chance, their bits of information from their surroundings appear more often in the news feeds of others. So, the higher density slowly reinforces over time. New bits of information bridging between the emerging groups have a lower chance of getting integrated because they 'drown' in reposted existing bits.

3. Agent interaction with LLMs

Large language models (LLM) are deep neural networks, typically with billions of parameters, that process natural language input from a user—given as a prompt—and return a machine-generated natural language response. Their instruction-following skills make them useful for tasks such as answering questions, summarizing text, and generating human-like dialogue. In an agent-based modeling context, an LLM could be instructed to act like an individual with personal characteristics provided as context within a prompt [11, 12]. In an agent-based model for simulating opinion dynamics, an LLM agent will form opinions on information they observe in their environment, which consists of information bits and other individuals. The individual LLM agent will process new information, e.g., from central news or friends, expressed in natural language, and express updated opinions in natural language.

In this study, we integrate LLMs into our filter bubble model as decision-making individuals. Integrating an LLM into our model as an individual agent opens the doors to analyzing dimensions of social

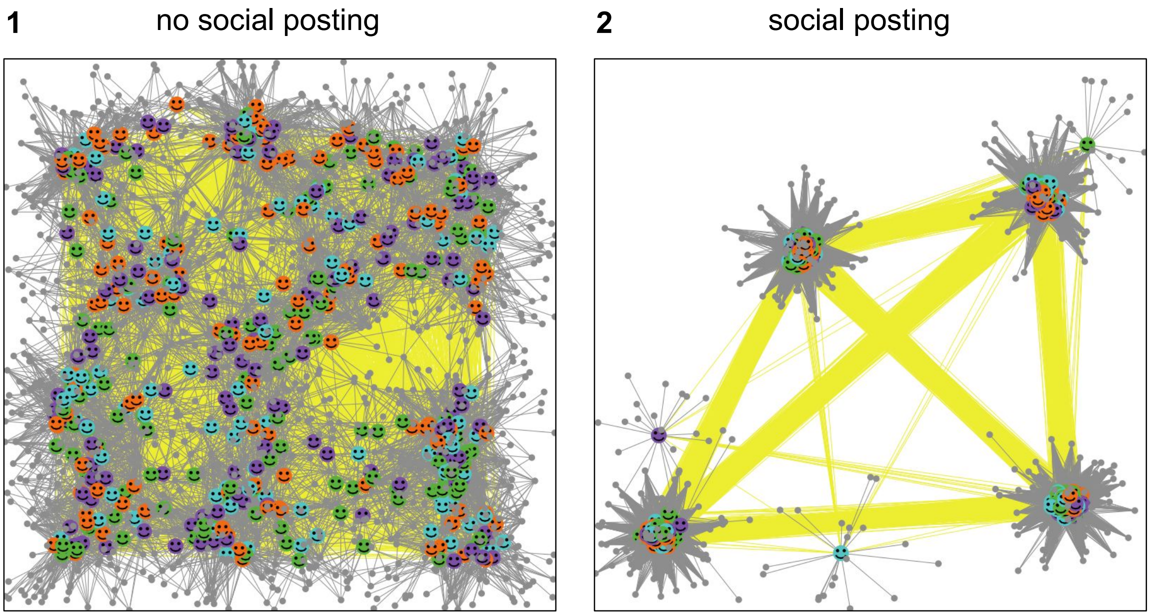


Figure 2: Examples of evolving positions of individuals and their info-bits in simulation runs. (1) Without social posting, only new bits of information. (2) Same parameters but with social posting.

interaction that we could not model before. It introduces complexity which makes the model a more realistic representation of the social processes we are modeling.

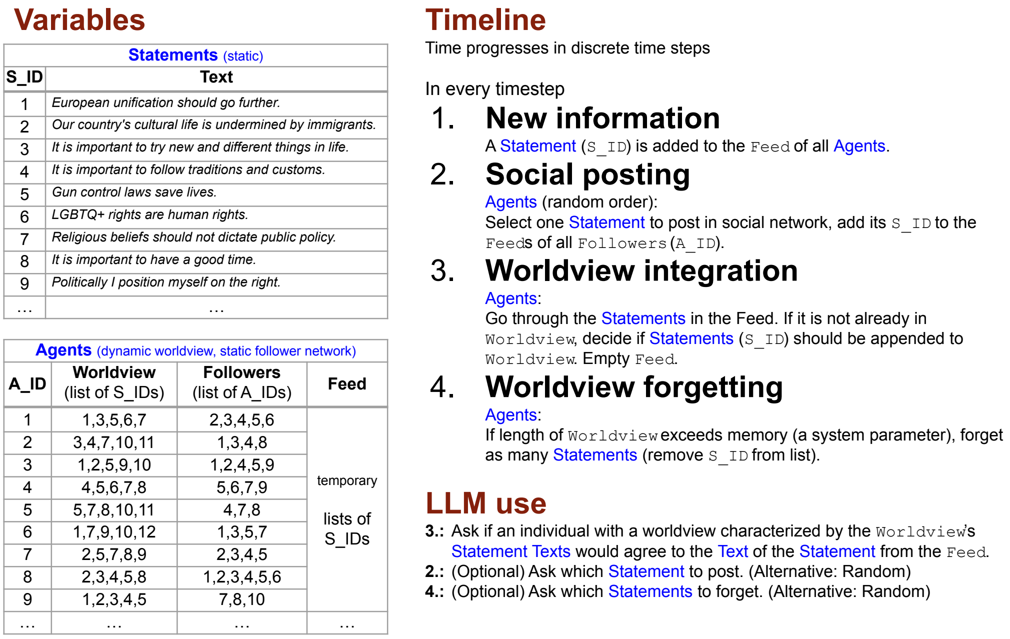
For the LLM version of the filter bubble model, we would keep the structure that an individual's opinion (or worldview) is determined by linking to a certain number of statements (where the maximal number is determined by memory). Instead of computing the barycenter of numerical values, we concatenate a list of statements to build a list representing the beliefs and convictions of that individual. This list can be given to an LLM agent as context. We will explore the following options related to the original filter bubble model:

Initialization and sources of statements. A first exploration builds on a corpus of statements taken from the questionnaires of the European Social Survey as outlined in Table 1. We can initialize agents by linking them to random statements until their memory is filled. Further options are: Using real-world news items or using LLMs as information generators.

Integration of new statements. This will be the core procedure. The agent-based model will confront the individual with a new statement (either selected randomly from the corpus and broadcasted to all or through a posting of a friend) and has to decide if a link is added. This should happen when the new statement fits the currently held statements. This could be implemented by asking the LLM agent to make a binary decision, e.g. agree or disagree with the statement, or by letting the agent express an opinion and then classifying the opinion into categories corresponding to integration and no integration.

Forgetting of statements. When memory is full, individuals need to drop a link to a statement. This can happen by random selection as in the original model, or by letting an LLM agent update the list of statements in the memory and observe which statement it removes to open space for the newly integrated statement.

Posting of statements in their social network. The easiest way is to select a random statement, as in the original model.



Variables

Statements (static)	
S_ID	**Text**
1	*European unification should go further.*
2	*Our country's cultural life is undermined by immigrants.*
3	*It is important to try new and different things in life.*
4	*It is important to follow traditions and customs.*
5	*Gun control laws save lives.*
6	*LGBTQ+ rights are human rights.*
7	*Religious beliefs should not dictate public policy.*
8	*It is important to have a good time.*
9	*Politically I position myself on the right.*
…	…

Agents (dynamic worldview, static follower network)			
A_ID	**Worldview** (list of S_IDs)	**Followers** (list of A_IDs)	**Feed**
1	1,3,5,6,7	2,3,4,5,6	temporary lists of S_IDs
2	3,4,7,10,11	1,3,4,8	
3	1,2,5,9,10	1,2,4,5,9	
4	4,5,6,7,8	5,6,7,9	
5	5,7,8,10,11	4,7,8	
6	1,7,9,10,12	1,3,5,7	
7	2,5,7,8,9	2,3,4,5	
8	2,3,4,5,8	1,2,3,4,5,6	
9	1,2,3,4,5	7,8,10	
…	…	…	…

Timeline

Time progresses in discrete time steps

In every timestep

1. **New information**
   A Statement (`S_ID`) is added to the `Feed` of all Agents.
2. **Social posting**
   Agents (random order):
   Select one Statement to post in social network, add its `S_ID` to the `Feed`s of all `Followers` (`A_ID`).
3. **Worldview integration**
   Agents:
   Go through the Statements in the Feed. If it is not already in `Worldview`, decide if Statements (`S_ID`) should be appended to `Worldview`. Empty `Feed`.
4. **Worldview forgetting**
   Agents:
   If length of `Worldview` exceeds memory (a system parameter), forget as many Statements (remove `S_ID` from list).

LLM use

3.: Ask if an individual with a worldview characterized by the `Worldview`'s Statement Texts would agree to the Text of the Statement from the `Feed`.
2.: (Optional) Ask which Statement to post. (Alternative: Random)
4.: (Optional) Ask which Statements to forget. (Alternative: Random)

Figure 3: Variables, timeline, and use of LLMs in the agent-based model.

Table 1
Examples of how a corpus of statements (beliefs or convictions) can be generated from questionnaires of the European Social Survey.

ESS ID	Section	Statement (slightly rephrased from question to statement)
psppsgva	Politics	The political system in our country allows people like me to have a say in what the government does.
psppsgva	Politics	The political system in our country does not allow people like me to have a say in what the government does.
lrscale	Politics	Politically I position myself on the left.
lrscale	Politics	Politically I position myself on the right.
euftf	Politics	European unification should go further.
euftf	Politics	European unification has already gone too far.
imueclt	Politics	Our country's cultural life is undermined by immigrants.
imueclt	Politics	Our country's cultural life is enriched by immigrants.
ipshabta	Human Values	It is important to show abilities and to be admired.
impdiffa	Human Values	It is important to try new and different things in life.
ipgdtima	Human Values	It is important to have a good time.
imptrada	Human Values	It is important to follow traditions and customs.

Figure 3 summarizes the variables and timeline of the new agent-based model and outlines the use of LLMs.

One key question when LLMs are used to simulate individuals in an agent-based model is whether their behavior matches or converges towards human behavior in similar contexts. Growing evidence suggests synthetic survey samples generated using LLMs could be remarkably similar to data from human samples [13], which suggests they could be useful for mimicking human decision-making processes. Research also indicates that LLM responses can be highly sensitive to both the information content and the structure of the prompt [14], making prompt engineering and validation key to achieving the desired realism in an agent-based model with LLM agents. Given that, we will need to do some small-scale tests to see how sensitive an average LLM agent's decision will be to potential prompt designs we may use in the simulation. Another important consideration will be cost management when using proprietary LLMs, which charge by the length of input tokens and output tokens. When we have 100 individuals, and each individual has on average twenty friends we need a hundred LLM calls to check the integration of the new statement for all agents, and 2,000 LLM calls to check the integration of the twenty postings received from friends for all agents. So, 2,100 calls in each time step. These challenges highlight the need for systematic testing to refine prompt structures and cost-efficient strategies before scaling up the simulation.

Appendix Section A shows a first promising example prompt.

4. Research Directions and Expected Results

After successful microscopic validation of individual LLM actions and implementation of a running simulation model, we want to explore if similar formations of filter bubbles emerge. This can be quantified for example by using measures of network modularity in the bipartite network of statements and agents, or by the number of statements covered by the society. This should inform us about the validity of the bubble-creating capacity of social media as shown in the original abstract model in a world of text-based opinions.

Acknowledgments

The work has benefited from funding in the Horizon Europe project ”European City Squared: Computational Social Science Simulation for Democracy” (EuropeanCity2 Grant agreement ID: 101178170, doi: 10.3030/101178170).

Declaration on Generative AI

The authors have not employed any Generative AI tools despite those directly mentioned in Appendix Section A as part of the research.

References

[1] A. Flache, M. Mäs, T. Feliciani, E. Chattoe-Brown, G. Deffuant, S. Huet, J. Lorenz, Models of Social Influence: Towards the Next Frontiers, Journal of Artificial Societies and Social Simulation 20 (2017) 2. doi:10.18564/jasss.3521.

[2] J. Lorenz, M. Neumann, T. Schröder, Individual attitude change and societal dynamics: Computational experiments with psychological theories., Psychological Review 128 (2021) 623–642. doi:10.1037/rev0000291.

[3] M. A. Keijzer, M. Mäs, A. Flache, Communication in Online Social Networks Fosters Cultural Isolation, Complexity 2018 (????) 9502872. doi:10.1155/2018/9502872.

[4] M. A. Keijzer, M. Mäs, The complex link between filter bubbles and opinion polarization, Data Science 5 (2022) 139. URL: https://publikationen.bibliothek.kit.edu/1000164788.

[5] G. Deffuant, D. Neau, F. Amblard, G. Weisbuch, Mixing Beliefs among Interacting Agents, Advances in Complex Systems 3 (2000) 87–98. doi:10.1142/S0219525900000078.

[6] R. Hegselmann, U. Krause, Opinion Dynamics and Bounded Confidence, Models, Analysis and Simulation, Journal of Artificial Societies and Social Simulation 5 (2002) 2. URL: http://jasss.soc.surrey.ac.uk/5/3/2.html.

[7] S. Banisch, E. Olbrich, Opinion polarization by learning from social feedback, The Journal of Mathematical Sociology 43 (2019) 76–103. Publisher: Taylor & Francis.

[8] D. Geschke, J. Lorenz, P. Holtz, The triple-filter bubble: Using agent-based modelling to test a meta-theoretical framework for the emergence of filter bubbles and echo chambers, British Journal of Social Psychology 58 (2019) 129–149. doi:10.1111/bjso.12286.

[9] U. Wilensky, NetLogo, Technical Report, Center for Connected Learning and Computer-Based Modeling, Northwestern University. Evanston, IL., 1999. URL: http://ccl.northwestern.edu/netlogo/.

[10] J. Lorenz, janlorenz/ABMExamples: v1.1, 2024. URL: https://zenodo.org/records/12764221. doi:10.5281/zenodo.12764221.

[11] P. Törnberg, D. Valeeva, J. Uitermark, C. Bail, Simulating Social Media Using Large Language Models to Evaluate Alternative News Feed Algorithms, 2023. URL: http://arxiv.org/abs/2310.05984. doi:10.48550/arXiv.2310.05984, arXiv:2310.05984 [cs].

[12] L. Zhang, Y. Hu, W. Li, Q. Bai, P. Nand, LLM-AIDSim: LLM-Enhanced Agent-Based Influence Diffusion Simulation in Social Networks, Systems 13 (2025) 29. URL: https://www.mdpi.com/2079-8954/13/1/29. doi:10.3390/systems13010029, number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

[13] L. P. Argyle, E. C. Busby, N. Fulda, J. R. Gubler, C. Rytting, D. Wingate, Out of One, Many: Using Language Models to Simulate Human Samples, Political Analysis 31 (2023) 337–351. URL: https://www.cambridge.org/core/journals/political-analysis/article/abs/out-of-one-many-using-language-models-to-simulate-human-samples/035D7C8A55B237942FB6DBAD7CAA4E49. doi:10.1017/pan.2023.2.

[14] M. Sclar, Y. Choi, Y. Tsvetkov, A. Suhr, Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting, 2023. URL: https://arxiv.org/abs/2310.11324. doi:10.48550/ARXIV.2310.11324, version Number: 2.

A. First prompt exploration

The following prompt has been tested with ChatGTP4o. From the 25 statements composing the example world view, the first two have been taken from the European Social Survey, the other 23 have been generated using the LLM using the first two as seeds. Those are statements representing a center-left (American liberal) worldview.

The LLM agrees with the new statement and drops the least fitting statement.

System prompt “You are an individual with a memory of opinion statements, and placed in a network of other individuals like you.

You will be presented a new opinion statement and make a decision on whether you agree or disagree with the statement. Your response on that question can either be‘agree' or ‘disagree'. No explanation of why you agree or disagree is needed.

If you agree with the statement, then it will be stored in your memory of statements that represent your world view. However, you have limited memory. You can hold up to 20 opinion statements in your memory. Once you reach 20 opinion statements in the memory, you need to drop one to open space for a new opinion statement. You also need to decide on which existing statements you will drop from memory to open space for the new statement. If you disagree with the statement, then you do not need to update the list of statements in the memory, and return‘None' as your response. The list of statements in your memory will be presented along with the new opinion statement.

You are part of a network of individuals. You can also decide to share one of the statements in your memory with other individuals in your network. If you decide to do so, you need to tell which statement you would want to share with your network.

Your response must be in json format with three entries representing the three decisions you must make. The three entries must be respectively named:‘response', ‘drop_statement', and ‘share_statement'.”

User prompt

```
What is your opinion on that statement?
"The political system in our country allows people like me to have a say in what
the government does."
```

```
Here is the list of opinion statements in your memory:
1. "The government should take measures to reduce differences in income levels."
2. "Women should be protected by men."
3. "Healthcare is a right, not a privilege."
4. "Wealthy individuals should pay higher taxes to support public services."
5. "Diversity makes societies stronger."
6. "Corporations should be held accountable for their environmental impact."
7. "No one working full time should live in poverty."
8. "Traditional gender roles limit personal freedom."
9. "Public education needs more investment, not privatization."
10. "Immigrants contribute positively to our society and economy."
11. "Gun control laws save lives."
12. "Climate change is a global crisis and demands urgent action."
13. "Workers deserve a living wage, not just a minimum wage."
14. "LGBTQ+ rights are human rights."
15. "The government should ensure access to affordable childcare."
16. "Racism is still a structural issue in modern societies."
17. "A strong welfare system reflects a compassionate society."
18. "Freedom of speech doesn’t mean freedom from consequences."
19. "The rich should not have more political influence than the poor."
20. "Police reform is necessary to ensure justice for all."
21. "Universal basic income could be part of the future of work."
22. "Feminism benefits everyone, not just women."
23. "Mental health should be treated as seriously as physical health."
24. "Religious beliefs should not dictate public policy."
25. "We need more cooperation and less competition in international politics."
```

LLM Response

```
{
  "response": "agree",
  "drop_statement": "Women should be protected by men.",
  "share_statement": "The rich should not have more political influence than
                      the poor."
}
```