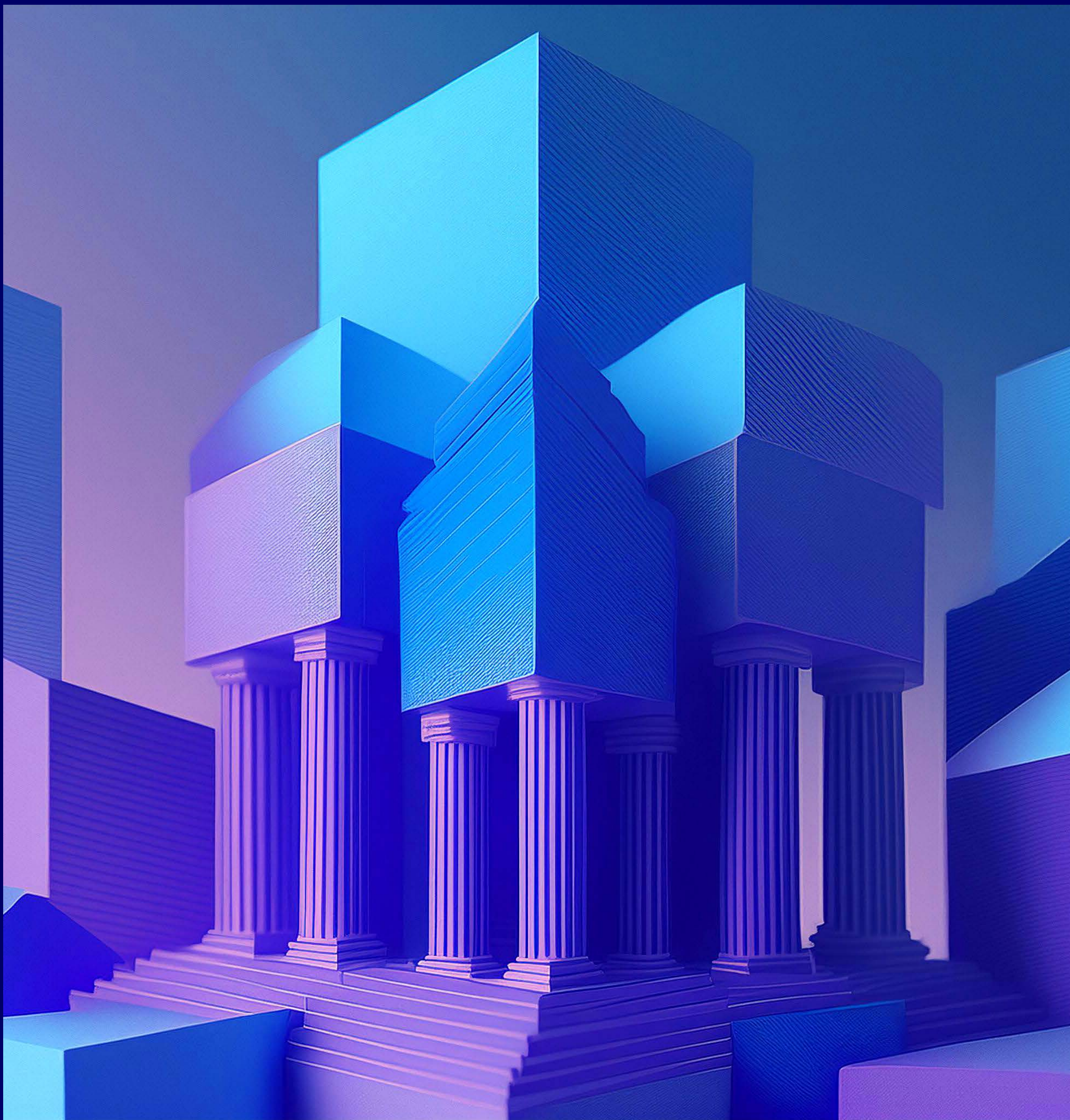
IE UNIVERSITY
CENTER FOR THE GOVERNANCE OF CHANGE

AI4 DEMOCRACY



This image was created using the AI tool Adobe Firefly.

HOW CAN **AI** BE USED TO INFORM **POLICYMAKING?**

JUNE 2024

ABSTRACT

LLMs offer new capacities of particular relevance to soliciting public input when used to process large volumes of qualitative inputs and produce aggregate descriptions in natural language.

In this paper, we discuss the use of an LLM-based collective decision-making tool, Talk to the City, to solicit, analyze, and organize public opinion, drawing on three current applications of the tool at varying scales: union decision making, coordination within DAOs, and nation-state consultations. We highlight the ways in which current-generation LLM tools can help leaders understand the needs of their constituents, review what measures are necessary to mitigate the flaws in these existing tools, and explore what future progress in foundation models would be most beneficial for the progress of tools like Talk to the City.

We conclude that rapidly advancing AI capabilities offer substantial potential for informing and refining the process of governance, but demand strong and careful governance to mitigate their risks and take full advantage of their benefits. If applied carefully and with an understanding of the social context of its use, AI-driven technology for democratic decision-making has the potential to support collective agency in ways that systematically feed back into AI governance and AI safety institutions, creating a virtuous circle of improving AI's impact on society.

WRITTEN BY
DEGER TURAN AND
COLLEEN MCKENZIE.

INTRODUCTION

Rapid advances in AI capabilities necessitate strong and careful governance of these new technologies and their effects. Their risks suggest that the current pace of technological development is both a source of concern and a source of hope for democracy. Concerns about our existing democratic processes and institutions being too slow and inefficient to address the many crises humanity is facing may result in degraded trust in the efficacy of our democratic processes, and without this trust, we risk losing the broad participation and agreement on legitimacy of outcomes that make democracy functional.

But we believe that the benefits of new technology outweigh its risks to our collective governance processes. New AI tools may help us escape this spiral of declining trust by allowing us to consult the public in much larger numbers, at a much faster pace, and in much more inclusive and transparent ways that capture diversity and nuance of opinion. Building tools well-suited for our social and political systems will require solving significant problems of sociotechnical process design,[1] political adoption, and public understanding, in addition to the technical work of building the tools themselves. But now that a growing number of AI pioneers have been popularizing this idea,[2] we believe that with the appropriate safety precautions, we are on the cusp of a new paradigm for collective governance, propelled by the design and development of open-source prototypes.

Early AI tools of this type include those for informing and refining the process of governance, making policy development more informed, refined, and representative of citizens' preferences.

We believe LLMs offer new capacities of particular relevance to soliciting public input—essential for democratic decision-making—when used to process large volumes of qualitative inputs and produce aggregate descriptions in natural language.

1 For background on sociotechnical systems, see Fairness and Abstraction in Sociotechnical Systems (Selbst et al, 2019)

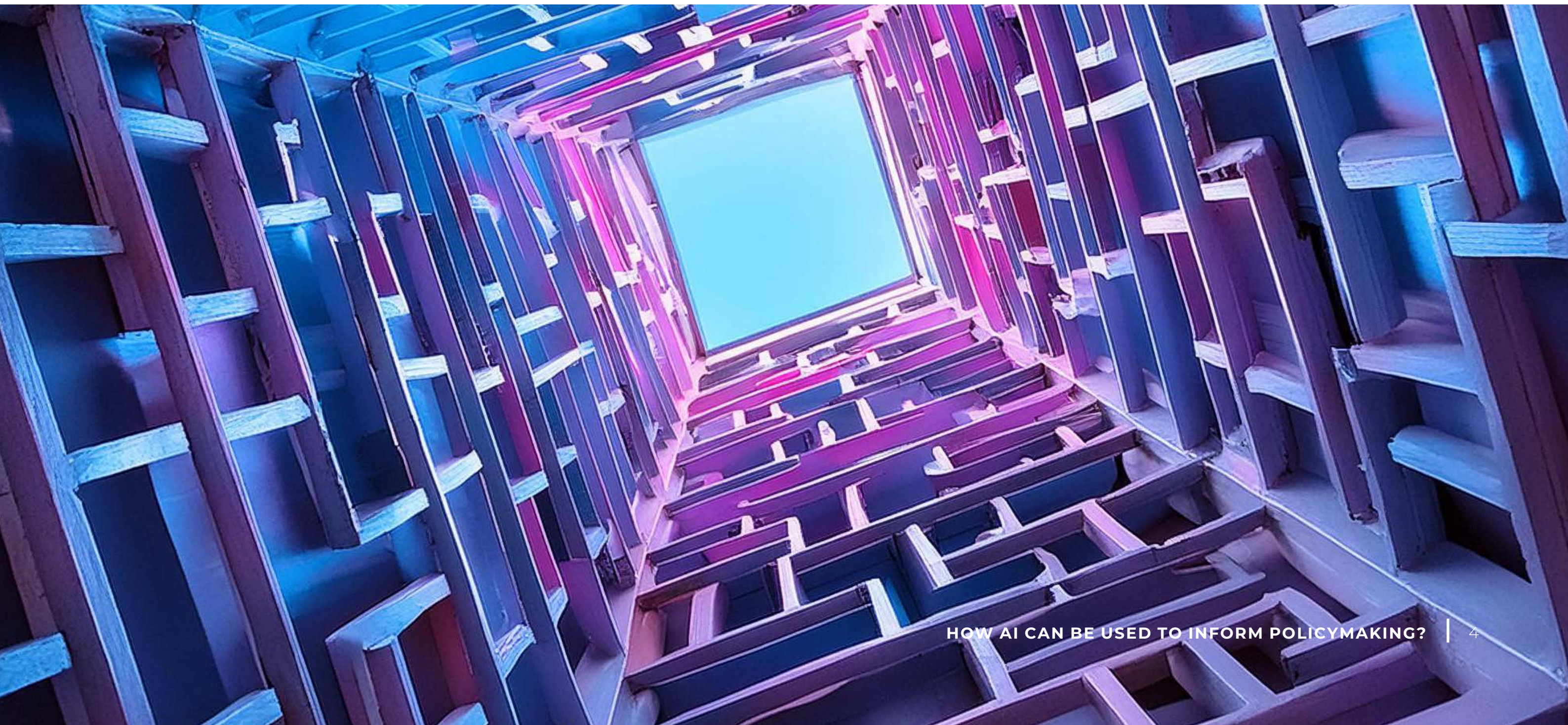
2 Examples include Glen Weyl and Audrey Tang's 2024 book Plurality: The Future of Collaborative Technology and Democracy, and papers such as Democratic Policy Development using Collective Dialogues and AI (Konya et al, 2023)

Existing processes for polling and surveying constituencies must trade off scale for qualitative detail, due to the limits of what teams of human analysts can process: large-scale opinion polls ask specific narrow questions, often with leading framings, while focus groups solicit detailed opinions from their members but are limited in size. In contrast, the NLP advances we see in frontier LLMs present an opportunity for soliciting and analyzing rich qualitative data at scale, identifying broad opinion trends from respondent populations at much greater efficiency than human analyst teams.

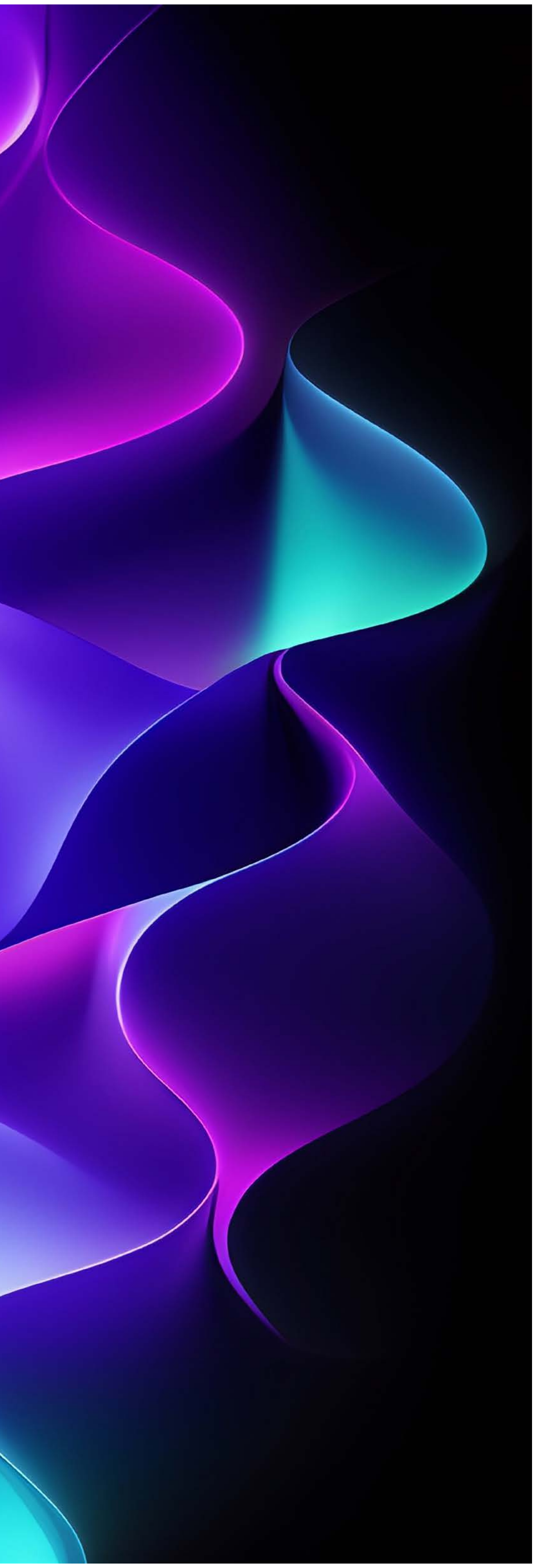
In this paper, we discuss the use of an LLM-based collective decision-making tool, **Talk to the City**, to solicit, analyze, and organize public opinion, drawing on three current applications of the tool at varying scales:

1. Finding shared principles within constituencies through large-scale citizen consultations
2. Compiling shared experiences in community organizing
3. Action-oriented decision making in decentralized governance

We highlight the ways in which current-generation LLM tools can help leaders understand the needs of their constituents, review what measures are necessary to mitigate the flaws in these existing tools, and explore what future progress in foundation models would be most beneficial for the progress of tools like Talk to the City. Our conclusion is that AI-driven technology for democratic decision-making has the potential to support collective agency in ways that systematically feed back into AI governance and AI safety institutions, creating a virtuous circle of improving AI's impact on society.



BACKGROUND



Democratic policy-making requires aggregating the interests and perspectives of the population at hand into a representation of their collective preference. The challenge of collecting and organizing this information has resulted in a variety of approaches, each with benefits and tradeoffs. Direct democratic referendums involve large portions of constituencies in decision-making, but this is a slow process at best, and the majority of the population does not have the time to dwell on every issue that a governing body has to decide. Representative democracy is a strategy for automating some of this information gathering process: by voting for a representative, the population chooses a surrogate decision maker. But because it is logistically infeasible for all those affected to be directly involved in the decision-making process, policymakers are limited in their ability to understand the desires of their constituents.

Addressing this specific problem is the goal of current efforts to create technology for assisting deliberation.[3]

By automating aspects of the data collection and analysis process with contemporary AI, collaborative deliberation systems have the potential to become more useful than they have been in the past, due to this increased ability to survey large populations in depth.

TECHNOLOGY-ASSISTED DELIBERATION

The potential of using information technologies to assist in collaborative deliberation is not a speculative matter. One example, currently in use by US congressional representatives, is **Common Ground for Action** (CGA). This synchronous discourse facilitation tool for video supports a process that allows participants to view the changing opinion landscape of the discourse on their call as they discuss potential responses to a given policy proposal. Given that this system has been able to successfully reveal or generate its target of

3 Further discussion of technological tools for deliberation can be found in Deliberative Technology for Alignment (Konya et al, 2003)

75% consensus among participants, and that such a level of agreement is uncommon compared to elections in the US's two-party system, the ability to design a system that can sum the perspectives of a small set of decision makers can be considered well established.

Asynchronous tools for coordinating large populations have also seen success over the last decade. Among them, **Polis** and **Remesh** stand out in their widespread use and demonstrated efficacy. Both tools rely on underlying algorithmic distribution of topics to participants who vote on them, though neither system incorporates a level of algorithmic complexity equal to that of current-generation LLMs. Polis, a crowdsourced survey platform which visualizes degrees of agreement, was used most prominently in 2016 as part of the **vTaiwan project, to inform policy on ridesharing services**. After gathering and informing relevant stakeholders (including taxi drivers and frequent users of taxis and Uber), Taiwan's Ministry of Digital affairs used Polis to poll these stakeholders, map the breadth of their opinions, and identify common ground that could suggest potential beneficial policy directions.

Remesh also collects responses from a wide variety of participants, but topics are specified by a moderator facilitating the process, who has disproportionate insight into the results of the voting process. Its successful **2020 application to ceasefire negotiations in Libya** demonstrates the tool's aptitude for situations involving complex diplomatic concerns. The UN Special Advisor tasked with this mission used Remesh as a platform for identifying the needs of leaders on both sides of the conflict and of the general public, and insights from this consultation resulted in a ceasefire agreement being reached within a week of deliberation, after decades of conflict. Remesh continues to be used in Libya at key political moments to inform economic and policy decisions.

LLM-ASSISTED DELIBERATION

Current-generation LLMs hold promise for building on these existing tools with powerful NLP technology, and especially for applications generating text. Previous tools have used AI to organize, analyze, and visualize discourse, but until the release of GPT-3, AI text generation was not sufficiently precise and accurate to incorporate into tools with real-world impact. Thus these recent developments hold promise for automating—and thus speeding and democratizing—aspects of the deliberative process that previously required human analysts to document and summarize their findings.

We see two especially promising areas for application of text generation to deliberation processes:

1. Automated distillation of results
2. Interactive, individualized elicitation

AUTOMATED DISTILLATION

While some tools for deliberation (e.g. Polis) rely on structured data, making report generation a straightforward quantitative task, deliberative tools that capture raw, unstructured language from discussions require more advanced tools to create organized reports on their outcomes.

Because LLMs can produce reasonably clear summaries of large volumes of text, they are well suited to performing some portion of the work of distilling the results of deliberation, summarizing takeaways, and discussing the relationships between different points and principles discussed. Automating this work offers two key benefits: speed of summary generation, and democratization of access to analysis.

Because LLMs can summarize large corpora in a small fraction of the time it would take humans to read and understand the same materials, they can be incorporated into the process of deliberation itself: distillations can be generated in parallel with deliberations, and these reports can themselves become resources for ongoing discussion. The democratization of access to this process, as a result of LLMs being much less expensive than teams of human analysts, gives more groups the opportunity to synthesize, summarize, and present the results of deliberative processes—which is especially promising for helping marginalized groups benefit from such processes.

Automated distillation is the feature of AI-assisted deliberation discussed most in this paper, as it is the key development in Talk to the City that has made the tool useful in the use cases discussed.

INTERACTIVE ELICITATION

AI agents can provide a variety of support to participants involved in deliberation. Interactive elicitation processes may help prompt richer, more detailed perspectives than respondents would give without iterated conversations: an LLM can ask follow-up questions, request elaboration, or point out aspects of the situation the respondent has not discussed. LLM bots prompted with details of a particular consultation could also explain details and context of discussion topics to help inform participants about their specific areas of uncertainty, could provide technical and empirical details to help participants refine and calibrate their existing opinions, or could suggest moral and practical considerations a participant has not yet considered. Such agents could be relatively passive and only provide help when explicitly asked, but in some situations it may be appropriate for agents to be more active, such as when participants are sufficiently confused that they don't know what questions to ask. Automating these support functions presents an opportunity for each participant to have in-depth conversations about the topic to organize their thoughts before deliberation begins, at a scale which would be infeasible or financially prohibitive when relying on human facilitators.

This potential avenue of AI-assisted deliberation is not discussed further in this paper, as the Talk to the City features that incorporate it are still in development and were not used in the use cases discussed.

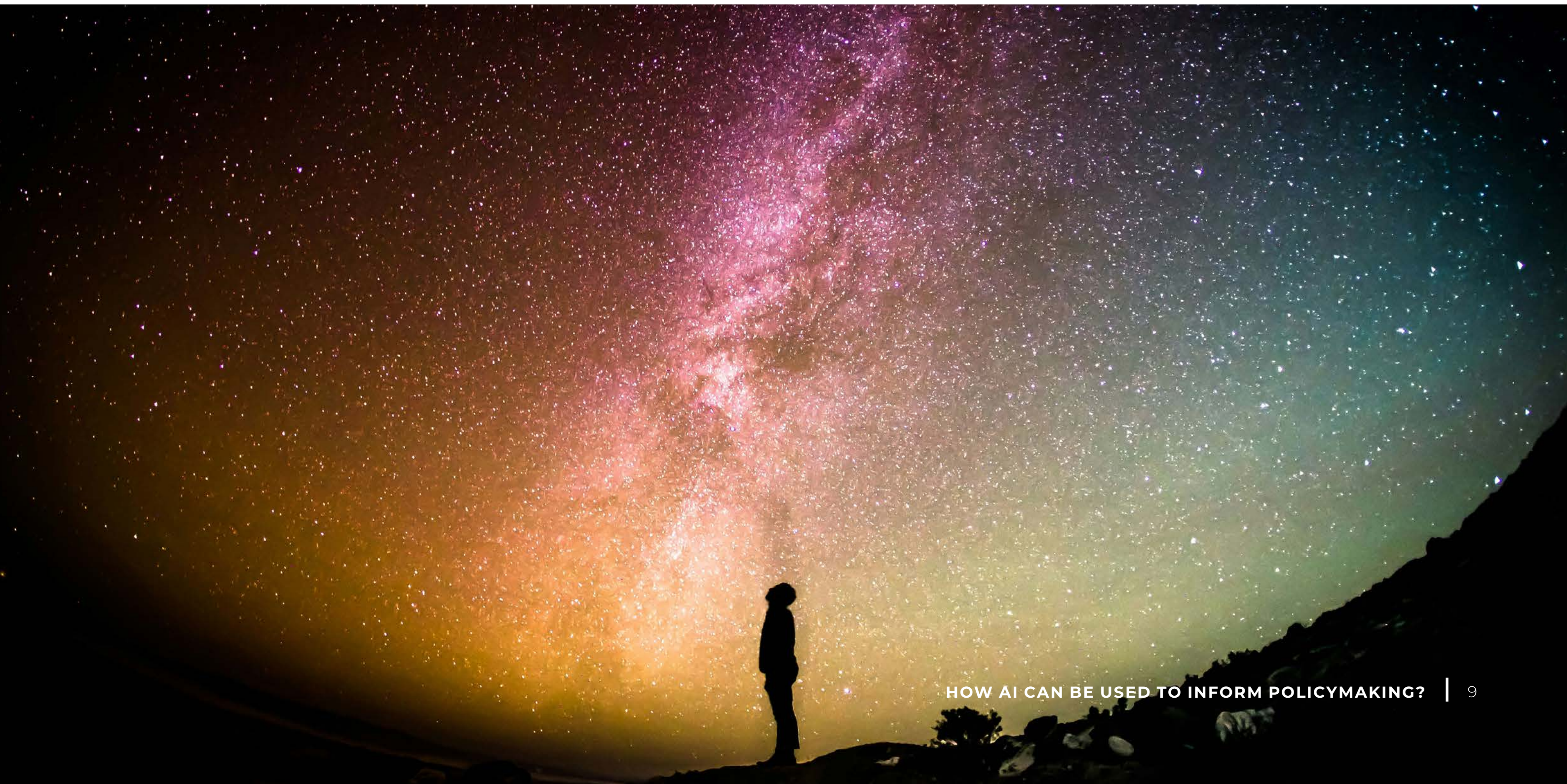
RISKS OF LLMS IN DELIBERATION

There are of course risks in using LLMs in any process with the potential for significant real-world impact. In addition to known risks of LLMs across many domains of application, such as hallucination of inaccurate information, the use of LLMs in representing the subjective information contained in human preferences and opinions—key elements of deliberative discourse—presents additional challenge in the creation of AI-assisted tools for these purposes. To be sufficiently trustworthy for such analysis, these tools must present accurate, high-fidelity representations of the opinions expressed by all participants, even when those opinions are politically incorrect or widely considered immoral.

This means that some of the current approaches to AI safety—such as RLHF, when used to make model outputs less offensive—may in fact be introducing bias that will skew LLMs representations of individuals' stated opinion.

We believe that some of these general downsides can be mitigated if AI tools are situated in process designs that bring transparency and human oversight to the results they produce. Results are most transparent if they are linked directly to underlying source material, and thus there is promise in building tools which allow users to inspect LLM-generated summaries to understand the participant statements they summarize. Deliberative processes can also be designed to incorporate both LLM summaries and cursory human analysis from participants, with the latter serving as a test for how accurate the former may be. Differences in human and LLM summaries may not indicate error—indeed, they may capture the most surprising elements of a deliberation—but having humans reflect on the validity of LLM summaries in such comparisons offers another layer of security against inaccurate automated reporting. Finally, explicit participant confirmation of LLM summaries of their viewpoints can serve as an additional error check, and source of legitimacy, for automated distillation.

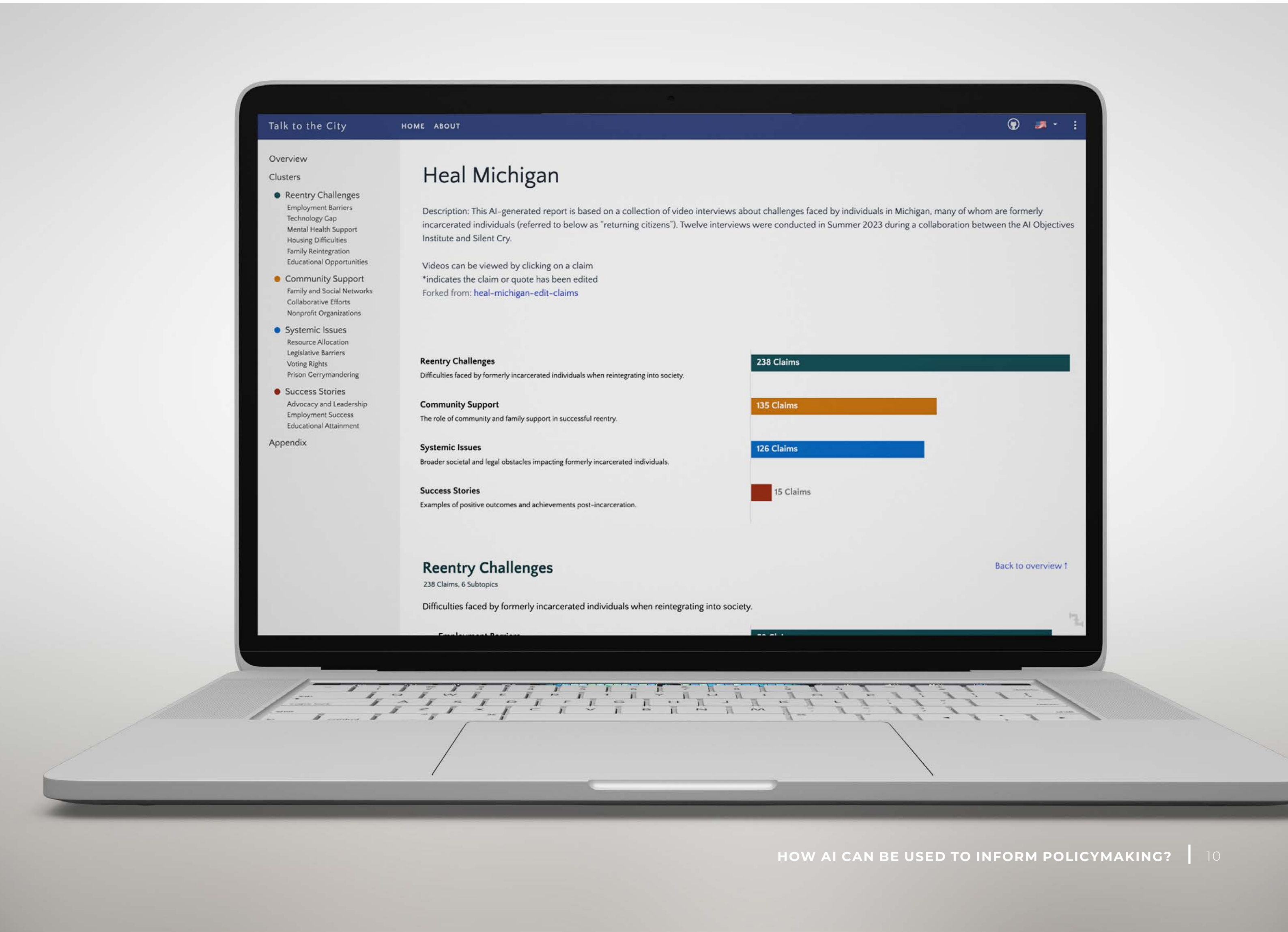
We discuss these and other limitations in more detail in each of our case studies, focusing on the risks and drawbacks most relevant to each application.



TALK TO THE CITY OVERVIEW

The case studies discussed in this paper were conducted using the [Talk to the City tool] (TttC), developed by the [AI Objectives Institute]. TttC is an open-source LLM data aggregation tool for improving collective deliberation and decision-making by analyzing rich qualitative datasets. It aggregates responses and clusters similar ideas, and provides an interactive interface for exploring the diversity of a population's opinions at both individual and group scale—revealing complexity, common ground, and polarization. The tool works with a wide variety of input data, including both structured data (e.g. Polis surveys), and unstructured content (e.g. freeform text, audio/video transcripts of interviews).

TttC is an open-source project, with code available in [GitHub].



PIPELINE WALKTHROUGH

The TttC data processing pipeline starts by processing a variety of data types, then uses an LLM to extract key arguments, and finally arranges similar arguments into clusters and subclusters. For all case studies in this paper, we used GPT-4 and GPT-4-Turbo for our LLM calls. The pipeline steps are explained in detail below, excluding data preprocessing (e.g. generating interview transcripts with the OpenAI Whisper API).

STEP 1

TAXONOMY GENERATION

An LLM is given the full text of all responses, transcripts, or discourse from the consultation, and returns a two-level summary of topics and subtopics discussed.

STEP 2

CLAIM EXTRACTION

Given the two-level list of topics and subtopics previously generated, an LLM extracts key claims from the text of each response, transcript, or individual participant in conversation, and maps each claim (and its underlying raw text) to its corresponding topic and subtopic.

STEP 3

DEDUPLICATION

An LLM checks claims within each topic and subtopic against each other, to merge claims with equivalent content and combine their raw text references.

STEP 4

SORTING AND REPORT GENERATION

Topics are sorted by claim volume, interactive charts are generated, and the results are formatted in an HTML report.

THREE PARADIGMS FOR AI IN POLICY DEVELOPMENT

In this section, we discuss three case paradigms for the application of AI tools for collective decision-making, exploring each with a case study of a Talk to the City deployment:

1.

Finding shared principles within constituencies:

Discourse analysis in Taiwan's AI Assembly workshops

2.

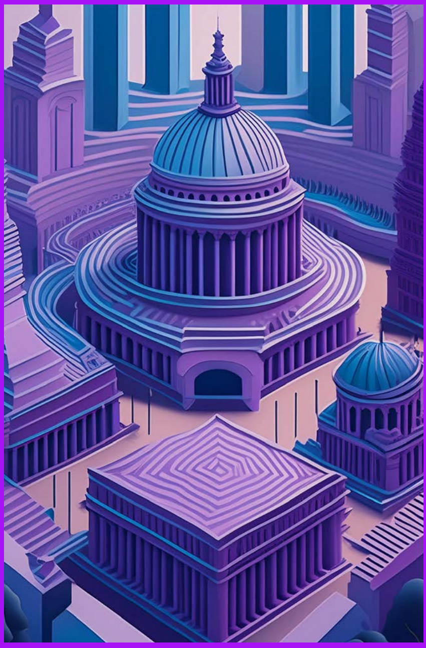
Compiling shared experiences in community organizing:

Interviews with Silent Cry's formerly incarcerated community

3.

Action-oriented decision making in decentralized governance:

Governance infrastructure for Mina Foundation



The **first paradigm** (finding shared principles) applies to policy development in democratic governance, where large populations are impacted by the outcome of decisions. The goal in these cases is to capture a set of governing values, rules, and principles from constituents' input, distilled from large volumes of deliberative discourse. Our case study for this paradigm was a collaboration with Taiwan's Ministry of Digital Affairs and the Taiwan AI Assembly project, applied to deliberations in a series of workshops discussing the future of AI.

The **second paradigm** (compiling shared experiences) applies in cases where communities seek to distill their shared experiences into a collective source of truth, producing an artifact which can then be used to help policymakers and other stakeholders act in the best interests of that community through an understanding of the details and context of the community's needs. The corresponding case study was conducted in partnership with an activist organization, Silent Cry, serving formerly incarcerated people in Michigan.

The **third paradigm**, action-oriented decision making, applies to policy resolutions over narrow sets of options, when time for decision making is limited. The corresponding illustrative case study is a decentralized autonomous organization (DAO) deciding on governance structures based on input from DAO members, where members have influence proportional to volume of tokens owned. Resource allocation problems are often classified under this paradigm, which necessitates expert input for forecasting consequences of policy options and eliciting desirabilities from the general public.

1

Finding shared principles within constituencies:

Discourse analysis in Taiwan's AI Assembly workshops.

In the last decade, Taiwan has been at the forefront of digital tools for democracy. The vTaiwan project, developed in 2014 to increase large-scale deliberation on Taiwanese policy, rose to international awareness after its use in a 2015 mass deliberation on regulation of rideshare services, and this consultation remains a key example of modern tools applied to the governance of modern technology. The original vTaiwan project relied largely on Polis for soliciting responses from large populations, but Taiwan's Ministry of Digital Affairs (moda) continues to explore other tools to integrate into their democratic processes.

Now moda has incorporated TttC into this effort. TttC has been used in large-scale policy consultations, such as one on same-sex marriage in response to recent policy proposals,[4] and to analyze the political platforms of leading candidates in the last presidential election based on their past interviews.[5] In this paper, we will focus on the deployment of TttC to support Taiwan's AI Assembly in late 2023, where it was used to augment existing infrastructure with TttC's analysis of respondents' opinions in the form of discussion transcripts. What we find is that TttC provides a novel opportunity for analysis of consultations to surprise us: in analyzing the large volume of transcript data, TttC pulled out topics of discussion not captured by the workshops' formal programming, and offered insights into what considerations were present across multiple Assemblies' topics of focus.

4 For example, this report on views about same-sex marriage: https://talktothecity.org/report/taiwan-zh

5 Resulting reports of candidates views are available for the Kuomintang (https://talktothecity.org/report/taiwan-2024-kmt-en-us) and Democratic Progressive Party (https://talktothecity.org/report/taiwan-2024-dpp-ZH) candidates

TALK TO THE CITY IN AI ASSEMBLIES

TALK TO THE CITY AI ASSEMBLY REPORT

Outside of projects like vTaiwan, when communicating policies to the public in Taiwan, the government has traditionally held public hearings and convened focus groups (called "listening sessions") to collect and understand citizens' opinions. Public hearings have a lower barrier for public participation, but they tend to have minimal impact on the policy itself. Only invited citizens and stakeholders participate in listening sessions, so despite records of these hearings being publicly available, they remain a relatively inaccessible route for the general public to participate in dialogue, creating difficulties for citizens to engage in policy-making.

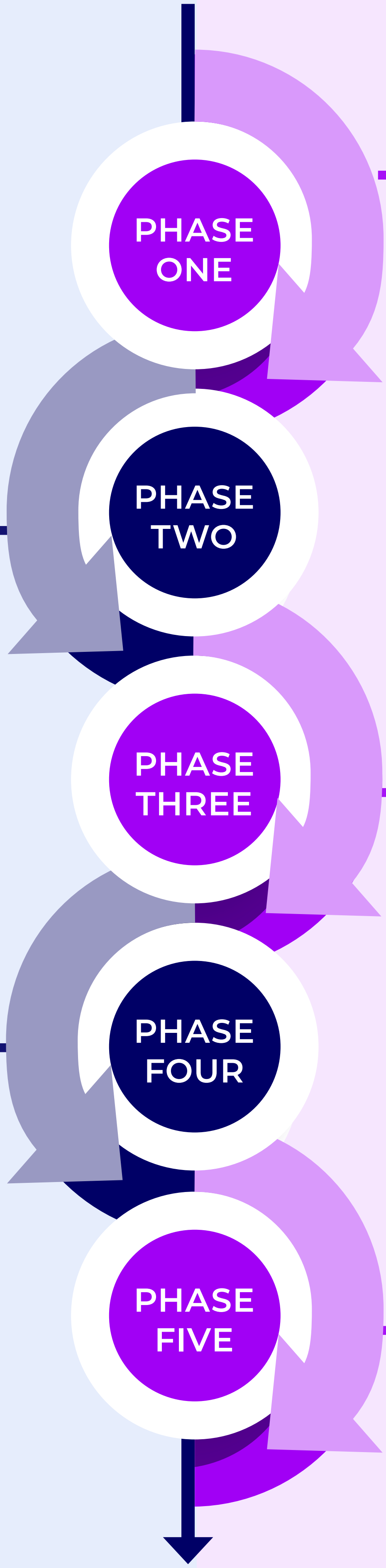
In response, Taiwan's AI Assembly[6] project pioneered a cycle in 2023 that included five AI Deliberative Workshops. After anonymizing the speakers' contributions, over 2,000 opinions were collected, with more than 400 participants involved. AI Assembly is an effort led by the Taiwan AI Academy, inspired by previous work in AI Impact Workshop,[7] a two-day event held by the Taiwan AI Academy discussing the impacts AI has brought to local industries, with the goal of helping Taiwanese citizens understand the changes and challenges ahead, encouraging informed responses and actions. The Assemblies initiative sprung from a shared mission and civic spirit among speakers and participants: to spread knowledge about the logic, use, limits, and impacts of technology.

Recognizing the need for a democratic and diverse approach, the AI Assemblies aim to make workshops a space for dialogue about assessing available resources, choosing the right paths and objectives, and promoting inclusive values through AI—advocating for its use in serving society, and avoiding exploitation. The AI Assembly team saw significant potential in using TttC to capture more of the discourse present in these workshops, and structure that discourse into a report that captures more of detailed deliberations than previously used tools allowed, as those other tools relied on participant responses to structured data.

6 https://ai-assembly.tw

7 https://aigc2023.aiacademy.tw

STRUCTURE OF THE AI DELIBERATIVE WORKSHOPS



1. Expert Interviews

The process began with deep-dive interviews with experts on specific discussion topics. These sessions helped to define the crucial questions that workshop participants would explore in subsequent phases, ensuring debates were grounded in informed perspectives.

2. Pre-Workshop Preparation

To align participants' understandings before diving into discussions, the AI Assembly team provided materials on the chosen topics for pre-reading, to enhance common knowledge among participants.

3. Conducting the Workshops

AI Assemblies experimented with different workshop formats, since different groups of participants require different discussion formats. For the workshop on the development of LLMs in Taiwan, AI Assembly chose small group discussions; while the workshop on the intersection of web3 and AI had a "fishbowl" discussion format, in which participants decide ad hoc whether to join a small panel of discussants or remain in the audience.

4. Data Digitization and Analysis

Using the latest in AI technology, every spoken word was transcribed and summarized using OpenAI's gpt-4-1106-preview model. This process involved an initial categorization based on expert interviews, with the team categorizing opinions into the 5 questions below. This categorization process was followed by a second-tier classification using GPT-4 that distilled the 2,000 opinions for further analysis.

5. Online Opinion Collection

After the workshop's conclusion, the AI Assembly team summarized these 2,000+ opinions using various techniques, including both manual review and online tools like Polis & Talk to the City for interactive online summaries.

RESULTS

TttC was used in Phase 5 of the workshop process to analyze and summarize the opinions participants discussed.

Talk to the City AI Assembly Report

This case study had the largest overall volume of text data processed, which resulted in a rich report covering many topics supported by many distinct claims. This depth stands in contrast to other applications of TttC in analysis of short-form text responses, in which the resulting reports reflect fewer details of the topics discussed, and less of the reasoning behind participants' beliefs.

The resulting report offered two key benefits:

A

It summarized information exchanged among participants, to automatically generate an overview of the understanding arrived at in the workshops and provide readers with more context for the opinions discussed.

B

It highlighted opinion trends across multiple Assemblies beyond the topics of the assemblies themselves.

A CONTEXT SUMMARIES

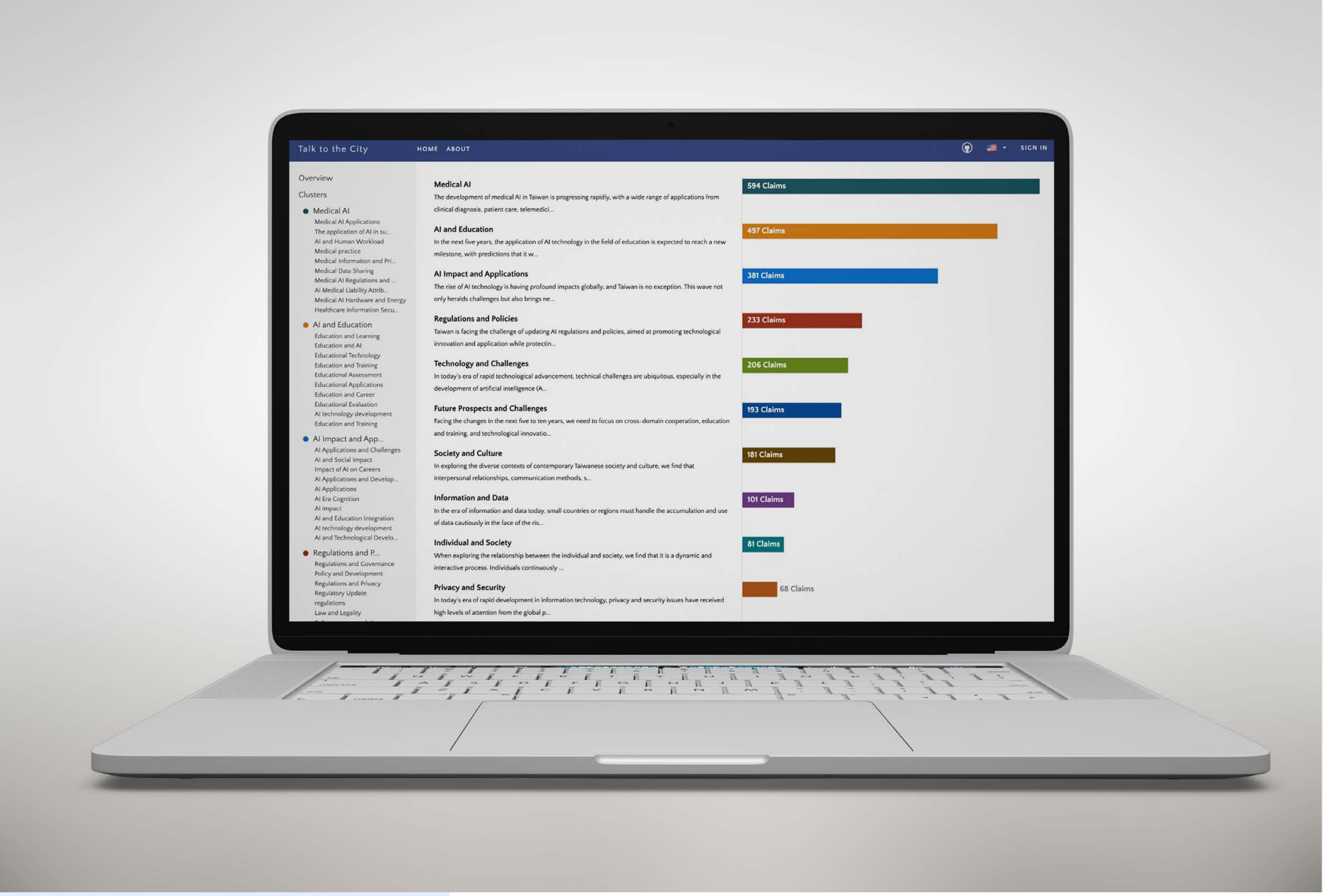
After generating the report, collaborators from moda and AI Objectives conducted interviews with several users who had experience with other online deliberation tools. The consensus among them was that TttC's key contribution was a report that helped readers understand the context and nuance of the issues discussed better than previous tools had done. One participant from Taiwan's civic tech community mentioned that TttC could help participants understand the overall issue of a conversation in more philosophical depth, and compared this to Polis, which is more like a traditional polling tool, reflecting the specific preferences of groups of people. Other participants suggested that TttC allowed deliberations to be summarized with relatively less bias than an opinionated person might introduce, and that the reports themselves were more accessible than previous versions: even individuals who didn't participate in the deliberations, and people without technical backgrounds, could arrive at a deeper understanding of the discussions concerning the development of LLM in Taiwan.

B OPINION TRENDS ACROSS ASSEMBLIES

The five Assemblies each focused on one of the following topics:

- Development and governance of LLMs in Taiwan
- Medical data management
- Data sharing and model authorization
- Web3
- Education

The resulting TttC report featured top-level categories for each of these topics, but also found other topics discussed across assemblies, in sufficient depth to warrant their own categories: among them, concerns about "Society and Culture" and "Individuals and Society" which were relevant to the topics of multiple assemblies.



These results suggest that TttC provides an opportunity for discovering unexpected insights from deliberation: while organizers, participants, and stakeholders in deliberative processes often have expectations of what topics will be relevant to a given deliberation, contingent on their own beliefs or their beliefs about others' perspectives. To some extent, these expectations will guide what topics are discussed, but over the course of contact with others' perspectives and experiences, new materials relevant to discussion are likely to come up.

In the Assemblies, each of which focused on a specific topic, we expect that a human-generated report might have relied largely on the headline topics of each Assembly to structure a report about their outcomes. Using TttC to study the full discussion at each Assembly, the result was instead a report that highlighted themes beyond the specified topics of each discussion.

OBSTACLES AND DRAWBACKS

The key obstacle we found in this case study was in LLMs' ability to **create accurate abstractions over large text corpora.** The AI Assemblies generated the largest volume of unstructured text data of any of these three applications, due to the large volume of text produced by the discussion transcripts. As a result, this case constitutes a stress-test of LLMs' ability to produce clear semantic results from large text corpora.

We found the resulting TttC report to be thorough at the expense of brevity. Overarching topics contain many closely related subtopics, such as the "Workshop and Discussion" and "Workshop Experience" subtopics under the "Future Prospects and Challenges" topic. While upon close inspection the subtopics are semantically distinct—with the former focused on the potential impacts of participating in the workshop, and the latter focused on how it felt to participate—preserving this nuance throughout the report resulted in the LLM generating the maximum number of subtopics we allowed per topic (10), for an already lengthy topic list. Such reports are not as straightforward to skim as those with fewer overall topics and subtopics,

For less semantically complex datasets, it's possible to avoid this problem by prompting LLMs with narrower limits on the number of topics and subtopics generated—but we see a direct tradeoff between the overall number of topics and the semantic clarity of each topic discussed. For datasets that cover a wide variety of discussion topics, producing a brief set of clear abstractions is difficult even for human analysts; and for LLMs, it often results in the conflation of ideas that have only partial semantic overlap. For example, in one version of this report, for example, in a discussion of the automation of medical care, claims about how this automation might impact the frequency of patient visits to hospitals are conflated with claims about the impact on the frequency of social visits these patients might receive.[8]

Ultimately, because the viewers of this report did not require a brief summary of the material, we chose to keep the long topic and subtopic lists to preserve the clarity within each set of ideas discussed, and to maintain a single report covering all four AI Assemblies. In other contexts, it may have made sense to generate separate reports for a few overarching themes (one per Assembly, for example), to decrease the variety of the discussion content analyzed in each report.

8 LLM-summarized claim: "Technological advancements may reduce the number of times caregivers visit the hospital." Example claims include:
a. "If there are smart wheelchairs, will they reduce the number of times patients come to the hospital?"
b. "Technology can replace some... will that reduce the number of people accompanying patients to the hospital?"

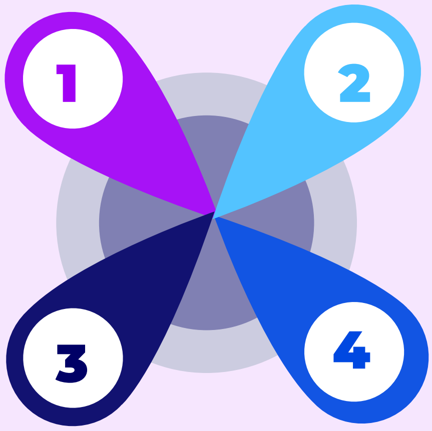
TAKEAWAYS

This case study illustrates a shared principle between constituencies: the desire for more inclusive and effective participation in democratic processes. By lowering barriers to engagement and providing clear, unbiased representations of public sentiment, TttC both improved understanding and ensured that a broader spectrum of voices was heard in the deliberation process. This alignment between government efforts and citizen needs fosters a more democratic, responsive, and collaborative governance environment.

OUR KEY LEARNINGS FROM THIS CASE STUDY ARE AS FOLLOWS:

Large text corpora of discussion transcripts produce **more detailed reports reflecting more of the thinking behind participants' opinions.**

Compared to other tools for deliberation, LLM summaries of discussion transcripts help viewers **quickly understand the context for, and nuances of, the topics discussed.**



LLMs analysis can pull out **unexpected trends and insights** from discussions with set topics.

The inherent tradeoff between **brevity and semantic clarity** in discourse analysis can cause LLMs to conflate ideas with partial semantic overlap.

2

Compiling shared experiences in community organizing

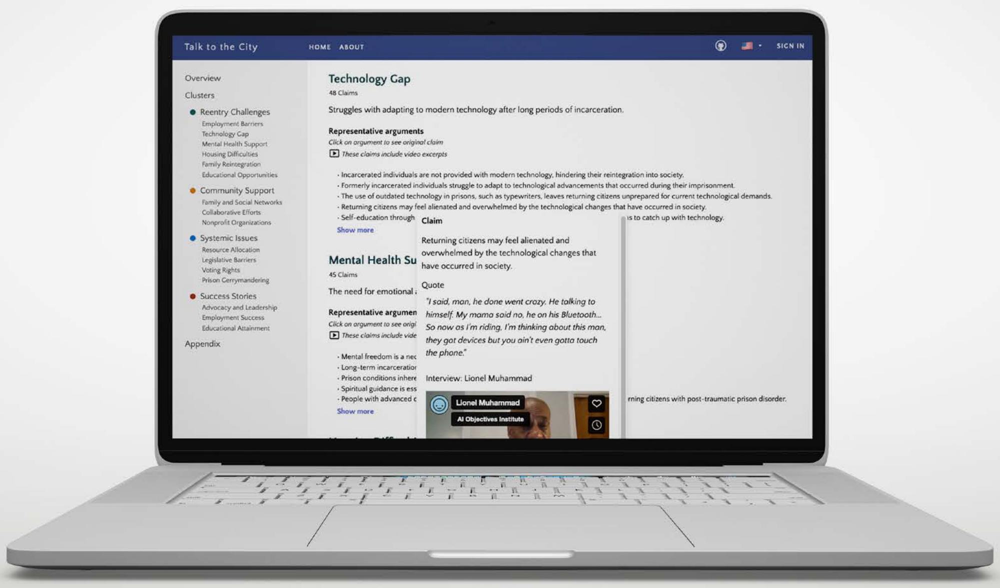
Interviews with Silent Cry's formerly incarcerated community

Many individuals and communities feel unable to influence the governance processes that directly affect them, even at local scales. This feeling reflects institutions' increasing reliance on large-scale aggregate data to inform decisions that affect large populations, even though these aggregates often fail to capture the nuance of individual perspectives and can miss under-resourced groups.

Silent Cry is a non-profit serving formerly incarcerated people and their families, working on holistic approaches to caring for people affected by mass incarceration, gun violence, and trauma. Silent Cry's executive director, Shawanna Vaughn, sought a way to tell the story of these individuals and their families in Michigan in more depth than traditional reporting methods. Inspired by the Equal Justice Initiative's exhibit at **The Legacy Museum in Montgomery Alabama**, Silent Cry and TttC built an interactive artifact, or a "collective source of truth," that aggregates the stories of formerly incarcerated people and their families. Our goal for the resulting project was to produce a shareable artifact that represents the priorities and views of the collective, involves every participant's full consent, and allows viewers to hear about people's lived experience from their own voices, ideally visually. By including video interview material, emphasizing the people behind the stories, the project aims to build awareness and empathy for this community and its experiences, and present its aggregate views to other activists, local lawmakers, and potential allies.

In the words of Silent Cry's founder and director, Shawanna Vaughn: "Talk to the City's Heal Michigan report is where advocacy meets legislation to create change."

→ **Report**
→ Video walkthrough: **Heal Michigan Demo**



METHODS

Heal Michigan represents a collection of video interviews of 12 participants (about 8 hours of video), conducted in the summer of 2023. Each participant was asked about challenges their community were facing and if they have ideas for solutions (e.g. policy measures) to those problems. All participants were based in Michigan, and 10 were formerly incarcerated individuals, many of who served terms longer than ten years.

As this was one of the early demonstrations of TttC and the participants were members of a marginalized community, it was decided to have the participants review all claims generated by the model. Once the report was generated, each participant reviewed and flagged any claims that were inaccurate, miscategorized, or did not belong in the dataset. After an internal review of the flagged claims, they were either edited (and approved again) or removed from the final report.

Claims were flagged by participants in three categories: inaccurate (faulty summarization), miscategorized (claim was correct but under the wrong topic), and removed (duplicates and other removal requests). Results of the participant review are listed in Table 1. Initially, 4.91% of the claims were flagged as inaccurate, miscategorized, or removed. The error rate was calculated by summing the removed, inaccurate, and miscategorized over the number of claims for that participant. After correcting these claims, the Revised Error Rate (i.e. the proportion of removed claims) was 2.84%.

Table 1.
Results of participant review of the final report

	Claims	Removed	Inaccurate	Miscategorized	Error Rate	Corrected	Revised Error Rate
1	76	4	2	0	7.89%	2	5.26%
2	35	0	0	0	0.00%	0	0.00%
3	37	5	0	0	13.51%	0	13.51%
4	34	0	0	3	8.82%	3	0.00%
5	38	0	1	0	2.63%	1	0.00%
6	40	3	0	0	7.50%	0	7.50%
7	79	3	4	1	10.13%	5	3.80%
8	35	0	0	0	0.00%	0	0.00%
9	51	0	0	0	0.00%	0	0.00%
10	42	0	0	0	0.00%	0	0.00%
11	32	0	0	0	0.00%	0	0.00%
12	30	0	0	0	0.00%	0	0.00%
	529	**15**	**7**	**4**	**4.91%**	**11**	**2.84%**

RESULTS

Using GPT-4's larger context window and allowing prompt editing enabled a well-organized report with compelling stories from interviews.

We did a demo of an early version of the report at an event hosted by Silent Cry celebrating Black August. Many guests at the event were exposed to a report of this nature for the first time and found interacting with it both novel and compelling. One of the participants enjoyed seeing themselves on screen and being able to explore a report and see the person behind the claim:

> "My involvement with the Heal Michigan project has been profoundly insightful ... This innovative approach not only shifts the narrative but also broadens and deepens the conversation around the experiences of those who have been formerly incarcerated. By focusing on providing nuanced insights into their journeys, the Heal Michigan project endeavors to enact positive change, influencing both individual experiences and societal perceptions. I am truly grateful for the opportunity to contribute to this initiative, which seeks to understand and share the intricate stories of these individuals' lives, fostering a greater understanding and driving meaningful reform."
>
> – Cozine Welch, *program coordinator for Michigan Collaborative To End Mass Incarceration and statewide organizer for the Michigan Criminal Justice Program at AFSC*

OBSTACLES AND DRAWBACKS

Interesting vs. Obvious

Out of the box LLMs (GPT-4) struggled to pull out what is interesting, especially stories that humanize an issue. In this dataset, it did a fairly good job at identifying overarching themes, but some of the stories' detailed content was lost in early iterations. For example, for stories about access to housing for formerly incarcerated individuals, the humanizing aspect of an individuals' experience is lost when the argument is reduced to "there needs to be better access to housing." One participant discusses being released from prison and being unable to find housing after doing all the right things. They chose to live in an abandoned house rather than be on the streets, because that's where they felt safest; over the years, they fixed up the house and eventually the owner signed the house over to her. In all versions, this story was not included in the list of quotes supporting the importance of access to better housing. While the definition of "interesting" will vary between different projects, our long-term plan is to compile a good list of prompt templates allowing us to capture the most appropriate types of stories.

Explaining high-context references

Many of the ideas discussed may not be clear to individuals outside of that community. In this report, some readers may not have context on things like "prison gerrymandering" or "good time credits." Having the ability within the report to explain these references would be beneficial. Additionally, it would be helpful to have links to the numerous organizations and pending legislation mentioned by the participants. We have experimented with various AI services, including OpenAI's GPT-4 (with the ability to search the web), and it seems that the aforementioned improvements could be automated in the near future, although we would prefer for such AI services to become more reliable than there are today, to make sure that we only add factually accurate context.

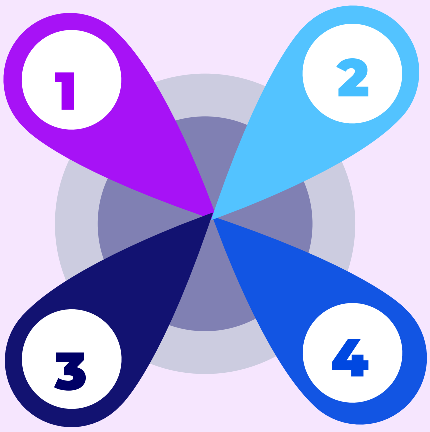
TAKEAWAYS

The Heal Michigan case study demonstrates the potential of leveraging AI technologies to amplify the voices of under-resourced communities. Even with a small dataset of 12 participants, struggles around digital literacy, access to jobs, housing discrimination and other barriers to reentry emerged as common ground among participants. This report constitutes a particularly clear example of how AI can produce artifacts that put human narratives at the forefront, and as language models and other AI capabilities continue to advance, tools like Talk to the City could become powerful platforms for advocacy—allowing under-resourced communities to raise awareness, find common ground, and directly influence decisions that impact them.

This report was the only case study run on a dataset consisting entirely of video interviews, and we believe that this format—in addition to the underlying material—is what made the report so impactful for viewers, both participants and people external to the community. One interview participant told us that even seeing his own video interview as part of the report was an impactful emotional experience of being recognized; and many viewers unfamiliar with the community reported that watching the video clips associated with each claim helped them understand the experiences of community members better than just reading the same excerpts from interview transcripts.

OUR KEY TAKEAWAYS FROM THIS PROJECT INCLUDE:

With only a 5% error rate, **LLM analysis exceeded our expectations of accuracy in** excerpting and categorizing of text, as judged by participant review of their summarized statements.



LLMs are **better at identifying overarching themes** than at highlighting specific, interesting details.

Reports on topics with **substantial latent context** may be confusing for viewers without that context, since LLMs don't automatically clarify that context.

Reports including **video interviews are particularly impactful** for helping viewers understand the ideas discussed.

3

Action-oriented decision making in decentralized governance

Governance infrastructure for Mina Foundation

Over the last 15 years, as blockchain technology has evolved and gained prominence, **decentralized autonomous organizations** (DAOs) have increased in popularity as governance structures for self-organizing collectives. While such institutions have partial overlap with traditional corporate structures, they differ both in their encoding of this structure—in cryptographic blockchain technology,[9] not legal documents—and in the values that drive their collective decision-making. Compared to traditional corporate structures, DAOs tend to be hierarchically flat organizations,[10] with relative influence derived from fungible stores of transferable value (blockchain-based tokens), not by titles.

This organizational structure makes DAOs an ideal context for experimentation with direct democratic processes, and indeed, such processes are essential to the collective self-governance of such institutions. While other democratic systems (such as nation-states' voting infrastructure, and intra-corporate consultations) employ specific stakeholders to make final decisions, DAO governance is inherently collective, with no guarantee of disproportionate influence by any single stakeholder. The incipient state of DAO structures means that these organizations are not yet well served by existing coordination technology, much of which relies on a hierarchical paradigm in which certain administrator users have more power than other group members.

In response to this lack of tooling, many DAOs have begun to create their own platforms, tailored to their specific needs, to support their distributed governance processes. The **Mina Foundation**, which governs the Mina Protocol ZK blockchain project, has partnered with TttC in creating one such platform for collective decision-making about protocol design and updates. In this case study, we discuss early applications of this developing platform, and our learnings about the aptitude of LLM discourse tools in DAO governance applications.

9 See the Ethereum whitepaper for an in-depth discussion of the technical details of blockchain-based DAO structure

10 See Decentralized Autonomous Organizations – DAOs: the Convergence of Technology, Law, Governance, and Behavioral Economics (Cardoso, 2023) for more detail on typical organizational structures of DAOs

TALK TO THE CITY IN MINA GOVERNANCE

TttC is one component of Mina's overall governance process, still in development, which will run continually in response to member-submitted proposals. To test the use of TttC in this pipeline, we collaborated on an initial study using TttC to evaluate Mina Foundation governance suggestions proposed by Foundation board members, but we include the full governance process proposed to provide context for the inclusion of TttC in the soon-to-be-launched full process.

PROPOSAL GENERATION

Any member of Mina can submit a Mina Improvement Proposal (MIP) for consideration by submitting a PR to the MIP GitHub repository.

EXPERT COMMITTEE EVALUATION

Mina member profiles contain metadata on each member's expertise. For each proposal, a small committee of members with a variety of expert backgrounds is randomly selected from the expert population. Expert committees deliberate on the proposal, and arrive at one of two outcomes:

1. **Expert revision:** If the expert committee believes the proposal is promising but would benefit from changes, the committee revises the proposal accordingly
2. **Acceptance as-is:** if the expert committee believes the proposal is promising as is, it accepts the proposal without revisions

COLLECTIVE EVALUATION USING TTTC

All voting members of Mina are sent the original proposal, and if applicable the revised proposal, for evaluation. Members submit responses that include a vote on whether one or both proposals are acceptable, and a discussion of their perspective on benefits, risks, considerations, and relevant context for their vote. Member responses are collected and displayed using TttC. Depending on the outcome of this voting process, the pipeline branches to one of the following steps:

1. **Near-consensus yes votes:** the proposal is accepted as-is for future implementation
2. **Near-consensus no votes:** the proposal is rejected outright, with no further review
3. **Mixed votes:** the proposal proceeds to another round of committee revision and voting (steps 2 & 3). If the threshold for consensus is not reached after this second round, the proposal is rejected.

TttC is employed here to capture the data generated by collective proposal evaluation, in which all Mina Foundation members are solicited for input through a Discord bot. The current pipeline uses a Google Forms integration to support ongoing report generation as Mina members contribute to the discussion. This use case also features TttC features that support quantitative analysis, in addition to the qualitative analysis demonstrated in previous sections: the report shows a breakdown of members' votes, in addition to visualizations of the relative importance of topics discussed.

In the prototype study we conducted,[11] the resulting TttC report segmented positive and negative sentiment on proposed governance, followed by six areas of key contextual relevance for proposal evaluation. Categories were clearly distinct, supported by clear claims for each topic and subtopic. But we found room for improvement in the length of verbatim participant responses attached to each claim: in this context of complex technical, economic, and cultural considerations, the quoted text was often insufficient to represent the specific reasoning behind individual claims.

11 We have not included the full TttC report Mina Foundation as part of this report to preserve the confidentiality of Mina governance discussions.

WHY TTTC FOR DAOS?

In our discussions and applications of TttC to Mina Foundation governance, we identified several specific benefits of LLM-assisted discourse analysis that make this approach particularly apt for application to DAO governance. Our conclusion is that the use of LLMs as third parties in discourse analysis preserves the flat hierarchy core to DAO structure, and the efficiency of LLM text processing is well suited to large constituencies making rapid decisions about a possibly high volume of proposals.

EQUAL VISIBILITY AND INFLUENCE

DAO governance protocols are encoded to ensure proportionate influence by token holders, preventing any single entity or group from dominating decision-making processes. While trustworthy individuals might be empowered to conduct the analysis of feedback on proposals like Mina's, entrusting a subset of DAO members with this task would be a divergence from the strict proportionality of DAO governance norms:

> such individuals would have disproportionate ability to influence the distillation and presentation of results, and disproportionate insight into the details of members' opinions. This increases the potential for nefarious actors to interfere with results, and even individual analysts acting in good faith will necessarily discuss results from their own frame of reference, informed by their own perspectives.

LLMs offer a relatively objective and uninvolved alternative to individual analysts. Much like impartial third-party human analysts, LLM processing of results will not alter the balance of influence or understanding among members of a DAO—but unlike uninvolved people, who may decide to join a DAO after gaining insight into its domain and processes, current-generation LLMs are guaranteed to be continually uninvolved in DAO governance outside their use in discourse analysis. While LLMs do not constitute an objective frame of reference, the fact that they are guaranteed not to benefit from specific proposal outcomes ensures trust that the summarization process is not influenced by perverse incentives.

EFFICIENCY AND FLEXIBILITY OF CORPUS ANALYSIS

As discussed in the **Background section**, LLMs offer the benefit of fast, inexpensive analysis of large corpora of freeform text. The speed of LLM analysis is of particular relevance to collective decision-making in DAO governance, especially for DAOs with low barriers to submission for new proposals (including Mina Foundation). The parallel processing of these proposals is already bottlenecked on the attentional bandwidth of foundation members; the additional bottleneck of human analyst capacity would further slow down this parallel governance process. Because many governance proposals will have interrelated concerns—particularly those that require substantial technical efforts to encode—and because many DAOs lack a central authority to make quick decisions, increasing the speed of collective decision-making for such organizations is essential to their functioning.

An additional, related benefit is the flexibility of applications of LLMs to corpora of changing size, without complex reallocation of human resources. This flexibility is of particular relevance to DAO governance, as these organizations can grow rapidly in response to viral publicity,[12] and members have variable and unpredictable levels of engagement with governance processes. LLMs provide on-demand capacity for distillation of member responses, without concerns for insufficient analysis bandwidth, or funds wasted on analysis capacity that goes unused.

OBSTACLES AND DRAWBACKS

This case study represented the most diffuse set of participants, with the lowest intrinsic incentives for submitting longform responses to the questions we presented. In contrast with the AI Assemblies, where participation was mostly verbal and the social elements of workshop participation elicited in-depth opinions, and with the Heal Michigan project, in which participants knew the resulting report could be used to improve policies with significant impact on their daily lives, participants in the Mina governance process had fewer incentives to respond with long, detailed descriptions of their opinions and beliefs. Respondents represented less than half of Mina members, and many respondents placed votes without offering the reasoning behind their choices. The resulting reports were sparse, with fewer overall topics and subtopics, and substantially fewer claims supporting each subtopic than in other reports with similar respondent numbers.

12 For discussion of one such rapid growth, see https://www.coindesk.com/learn/understanding-the-dao-attack

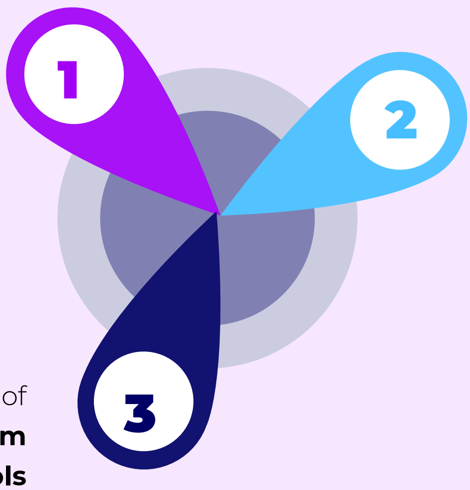
The limited report output does not suggest a failure of LLM analysis—in contrast, we believe it constitutes an opportunity for application of LLMs to the elicitation process, as previously discussed. Ultimately, participation may rely on respondents' underlying motivations for the outcomes of governance decisions, but it's possible that automated elicitation tools could attempt to make the voting process more engaging through conversation. The current Discord survey bot is not designed to involve members in deep interactive discourse, but tools like Futr and **Rival** offer such features for business use cases, and AI Objectives Institute is developing a similar tool for use with TttC.

TAKEAWAYS

This case represents the largest number of participants surveyed for the resulting report, and the most complex technological integration we've explored with TttC. The incorporation of our core functionality confirmed that TttC is a viable open-source tool for developers to integrate into more complex governance processes, which holds promise for the growing ecosystem of interoperable open-source civic technology.

OUR KEY TAKEAWAYS FROM THIS PROJECT INCLUDE:

LLM analysis is particularly compatible with DAO governance structures, due to the speed of processing and the preservation of flat hierarchies



Processes that **pair LLM analysis with validated quantitative analysis of structured data** (votes) can offer explanations for the majority voters' choices

Eliciting longform explanation of voters' rationales could **benefit from LLM elicitation tools**

COLLECTED BEST PRACTICES FOR LLM-ASSISTED DELIBERATION

The three case studies discussed represent a variety of applications of LLMs to collective discourse, and presented us with a similarly wide variety of results to study. In this section, we summarize our strongest conclusions about best practices for the use of such tools, based on the key insights and obstacles we encountered in each application.

In our appraisal, LLM analysis works best for:

1. Summarizing large-scale or longform discussions with reasonable accuracy (5% error rate)
2. Identifying overarching themes from broad discussions, including themes that may not have been expected given the stated topic of discussion.
3. Producing reports that efficiently communicate the context for, and nuances of, the topics discussed, including to viewers without prior exposure to the discussion.
4. Pairing with structured voting data, analyzed separately, to offer context and rationale for voters' decisions

LLMs are less well-suited in the following contexts:

1. Discussions that include jargon, terms of art, or ideas that require substantial latent context to parse
2. Sparse datasets, where the efficiency of summarization does not constitute significant additional value
3. Producing reports that highlight specific interesting anecdotes, as opposed to overall trends

In light of these opportunities and constraints, we suggest the following as best practices:

1 Seek rich, substantial datasets, but consider semantic complexity

In general, report quality has varied directly with the size of the datasets analyzed: the more text the LLM has to distill, the more interesting the condensed summary will be. Separate from the overall amount of data, we have noticed tradeoffs between the volume of ideas discussed in each dataset and the brevity of the resulting reports. We suggest users of LLM tools for deliberation consider these tradeoffs when collecting data, and possibly conduct separate rounds of elicitation for discussion of multiple complex topics.

2

For ideal topic hierarchies, iterate on prompts used to generate reports

Depending on the underlying structure of the data, the first hierarchy generated by an LLM analysis may or may not pull out the clearest description of overall themes. In some reports, many topics have semantic overlap—for example, most topics in initial versions of the AI Assembly report included the term "AI"—and the resulting repetitive titles will not be the clearest representation of the underlying information. We found that generating an initial topic hierarchy, and then amending the prompt with directions to merge or rephrase certain topics, consistently generated clear reports with representative topic and subtopic titles.

3

Inspect results for conflation, not hallucination

While LLM hallucinations are a well-known risk of this new technology, they did not represent a significant obstacle for report generation: in our most in-depth review (in the Heal Michigan study), we found no hallucinated claims from participants. We did, however, encounter problems of LLMs conflating similar ideas. The automated analysis occasionally combined separate but related claims, and miscategorized claims with only partial relevance to a particular topic.

4

Supplement LLM analysis with materials clarifying latent context

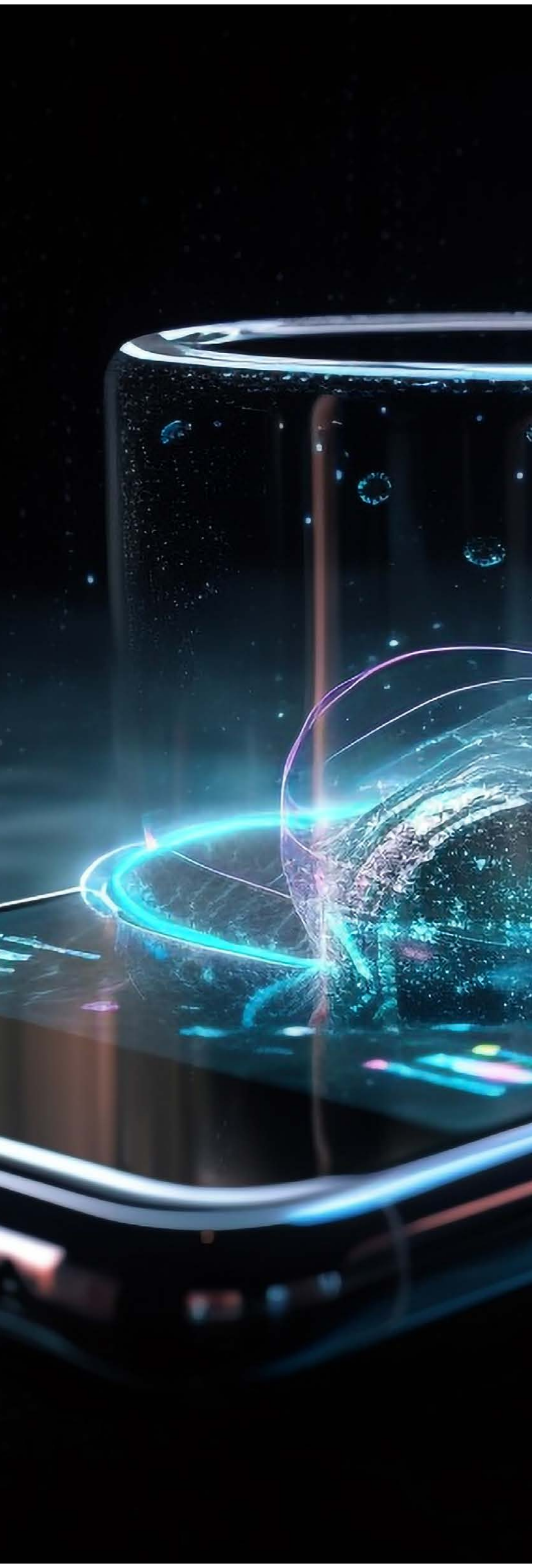
While a human analyst might consider how to frame content for unfamiliar readers, when analyzing a topic that requires significant background knowledge for interpretation, LLMs summarize only the information provided, without offering additional context. Jargon and terms of art are not defined in clearer language, and no additional information outside the discussion is incorporated into the report. For some reports, this problem can be easily solved by amending prompts to include directions to substitute simpler language for jargon; but for reports like Heal Michigan, where the context of ongoing societal processes (e.g. "prison gerrymandering") is relevant, reports would benefit from supplementary information curated by people familiar with the context of discussion

5

Use audio and video content when possible

This suggestion is specific to TttC and other tools that take multiple data formats as input. While we believe LLM summaries are valuable to a variety of use cases, the impact of the video interviews on unfamiliar viewers is unparalleled by analysis of pure text. Few other automated platforms offer a combination of overarching summary and individual video clips—a format more representative of in-depth journalism than of technology for collective discourse—and in our experience, bringing this human element into mediated communication helps readers better understand the other perspectives they encounter.

CONCLUSION



It is not surprising that good policy making requires a plurality of processes and tools, given the diversity of expectations and utilities required for guaranteeing successful policies that achieve constituents' desired outcomes. AI tools, and in particular those based on LLMs, will be most useful if they can complement the current modular context of tools for policy development—including consultations and desirability elicitations, expert opinion solicitations, regulatory feedback on drafted policies, and ranked choice voting systems. While the results from these Talk to the City pilot projects are promising, much more experimentation and assessment is still needed to fully understand the feasibility and extent of the application space of AI use within policy development.

The three forms of policymaking explored in the paper all benefit from the same shared toolkit of Talk to the City, though the role and responsibilities of the toolkit is different in each case. We expect similar case-specificity in applications of AI tools in other contexts, making it critical for multiple organizations to experiment further with different policy creation processes. One of the shared drives that has yielded successful case studies has been shared intent: if a large constituency has a shared goal or a mission, that collective and its leadership can iterate substantially faster than collectives without such incentives. Activist communities trying to increase visibility, or DAOs competing for participatory attention, are good examples of governance and policy-making outside the governmental scope. Labor unions represent a particularly compelling next case study, which we plan to explore in the coming months.

Different policy creation processes will likely require distinct tools and capabilities, underscoring the importance of a modular, open-source approach that enables collective development tailored to diverse needs. Navigating the sociotechnical complexity of AI governance will necessitate an interdisciplinary, multi-stakeholder effort founded on continued empirical pilots and shared learnings. We present here an early foray into LLM tools for such contexts, and conclusions from our experiments about the most promising directions and situations in which to apply these new tools. Only through such collaborative inquiry can we responsibly harness the potential of AI to augment rather than automate the policymaking process.

UNIVERSITY

CENTER FOR THE GOVERNANCE OF CHANGE

Written by:
Deger Turan and Colleen McKenzie.

This paper is part of a series of four papers within AI4Democracy, a global research and outreach initiative led by the Center for the Governance of Change at IE University, with Microsoft as strategic supporter. AI4Democracy seeks to harness AI to defend and strengthen democracy through coalition-building, advocacy, and intellectual leadership.

Suggested citation:
Turan, D., & McKenzie, C. (2024). *How can AI be used to inform policymaking*, AI4Democracy, IE Center for the Governance of Change.

Images: All images were generated by different AI tools.

Design: epoqstudio.com

FOR MORE INFORMATION ON THE AI4DEMOCRACY INITIATIVE, VISIT:

IE.EDU/CGC/RESEARCH/AI4DEMOCRACY

cgc.ie.edu