Process Card: Talk to the City

Process Overview

Basic information

- **Process developer:** AI Objectives Institute (Team: Colleen McKenzie, Stacey Svetlichnaya, Brandon Perry, Emre Turan, Pawel Sisiak, Deger Turan, Brittney Gallagher)
- **Process design date**: Feb 2023 – present (updated May 2024)
- **Process design version**: v0.2
- **Process card version**: v0.2
- **More information:**
    - Project overview: https://ai.objectives.institute/talk-to-the-city
    - Github repo: https://github.com/AIObjectives/tttc-light-js
- **Maturity:** Alpha-stage tool undergoing rapid iterations after 8 months of pilot studies
- **Tools and methods used:** GPT-4, Claude, firebase, custom visualization
- **Where to send questions**: colleen@objective.is

Intended Use

- **Primary intended use:** Distill aggregate unstructured opinions of a population into clear summaries that allow zooming in for individual nuance
- **Primary intended users**: Leadership of small (e.g. unions) and large (e.g. government agencies, DAOs) constituencies; self-governing organizations
- **Primary intended context**: Complex or loosely bounded discussion contexts, in which the topics of highest importance are unclear or contentious, and in which leaders want to avoid imposing frames by asking specific narrow questions
- **Out-of-scope use cases**: Secure and validated preference aggregation applications, such as identity-validated voting; analysis and visualization of complex *quantitative* data (TttC focus is qualitative)

Structure

Inputs

- CSV dataset representing a **population**'s views on a specific **topic of discussion.**
    - Individual views can include unstructured text, audio/video media, and structured data (e.g. from Polis surveys) in any combination matching the data format spec
    - *Example:* "What should be our top priorities for improving workplace conditions?" asked of members of a labor union
- Revisions to default prompts specifying how an LLM (by default GPT-4) should extract, cluster, merge, and summarize views in the dataset

Outputs

- An automated **report** displaying:
    - Two levels of **clustering** of the population's responses
    - Specific **claims** about each topic, and **quotes** from individual responses that map to those claims
    - **Quantitative analysis** of the population discussing each topic and subtopic, including respondents discussing each topic, and analysis of agreement between respondents
    - Suggested reasons for significant disagreements between respondent clusters
- JSON-formatted output of data displayed in report

Additional impacts (state changes)

What else happens externally or to participants as a result of the process?

- Report creators **understand the broad-strokes opinion trends and the individual perspectives** of the populations they serve
    - Report creators have a **better understanding of the context** for asking further specific questions
- Participants **understand how their stances compare** to their peers'
- Both report creators and participants **understand areas of common ground and controversy,** and potential **routes to resolving disagreement**

Details

Principles & Rationale*

- **Large-scale discussion and opinion analysis is essential to collective coordination** and governance, but current methods can't process rich, qualitative opinion data at scale. LLMs present a new solution to this problem: automating the collection and analysis of content from large populations, allowing analysis to scale not only in the size of the respondent populations analyzed, but in the availability of such analysis to groups for whom human analysts teams would be cost-prohibitive.
- Ideally a collective's understanding of its aggregate perspectives would involve **understanding both high-level opinion trends and the nuance of individual opinions**. Interfaces that help people explore collective discourse at multiple scales will be most helpful in presenting general trends without losing nuance or diversity of opinion.
- If we create **tools for coordination and decision-making that improve apace with advances in AI capabilities**, we can help human collectives become increasingly effective in the face of increasingly complex questions about use and governance of new technologies, creating a **virtuous circle** of improving AI's impact on society. Automated analysis can't replace human discussions—but it can improve the information that collectives and their leadership have when making impactful decisions.

Benefits

What are the reasons to use this process or include it in a larger process? What are difficult challenges that it addresses?

- **Open-ended elicitation**: Report creators can define a nebulous area for discussion to capture the full breadth of what respondents believe is relevant to a particular issue
- **Flexible input data types**: The pipeline works with unstructured survey responses, but also with traditional multi-choice surveys, Pol.is consultations, scraped tweets, and blog post collections (as CSVs), which covers many use cases and allows surveys to be tailored to the needs of specific populations
- **Efficient and inexpensive**: The backend calls out to GPT-4 for data processing at a very low cost per report (usually <$10 for ~100 paragraph-long responses), replacing the expensive teams of human analysts previously necessary for analyzing unstructured data

- **End-user and API accessible**: hosted report UIs let nontechnical users easily view results, while other projects can use TttC as an API in more complex collective governance processes

Current Challenges & Limitations

What are the current challenges and limitations of the process which may be improved in future versions or process runs?

- **Analysis errors and oversights**: Up to 10% of the specific opinions extracted in reports so far are redundant with other opinions in the same topic or are miscategorized (i.e. not relevant to their parent topic), making reports less readable and occasionally misleading
- **Some computer literacy required to run reports**: The interface for report creation requires uploading a CSV and editing an LLM prompt, which some of our less technical users have found intimidating
- **Intuitive UI vs. complex data visualization**: Our most user-friendly interface presents basic graphs about population trends; we expect there are ways to allow more zooming in and out of different levels of detail, but creating intuitive interfaces for doing so across varied use cases is challenging

Intentional Limitations

What are the limitations of the process which are expected by design?

- **Limited opportunities for recursive deliberation**: The current process lacks the feedback loops of participant dialogue that can help populations coalesce around common ground
- **Focus on analysis of qualitative data, not quantitative:** TttC is not a full-service data visualization platform for complex quantitative data, and supports only basic graphs for structured data (e.g. voting)

Assumptions

What assumptions must be true for the process to be applicable and effective?

- Report creators can reliably survey sufficiently representative samples of their populations
- Report creators can define topics of discussion sufficiently clearly to elicit useful responses
- Report results are trusted as accurate but not precise representations of populations' beliefs (to account for LLM errors)

Explanation Overview

Talk to the City is an open-source LLM interface for improving collective deliberation and decision-making by analyzing detailed, qualitative data. The basic process is as follows:

1. A representative of a particular constituency surveys their members on an open-ended area of discussion
2. The representative uploads responses into Talk to the City, and edits LLM prompts as desired
3. The Talk to the City pipeline:
    a. Extracts common opinions and beliefs from the responses
    b. Aggregates opinions and beliefs with a two-level clustering of similar statements
    c. Maps key opinions and beliefs to each cluster and sub-cluster
    d. Analyzes disagreement between respondent groups, and suggests potential cruxes for areas of disagreement
    e. Produces a report with summaries of aggregate-level opinions and beliefs, and specific segments of individual responses that support them
4. The report is made public so that users (both report creators and surveyed populations) can understand high-level opinion trends and drill down to the subclusters they find most interesting. All data used in the report is available as a JSON export.

- Live tool: talktothecity.org
- Open-source github repo

Parameters

What are the things that might change across different runs versus stay the same, and what are the variables that you might toggle for different variants of the process?

We provide report creators with default **LLM prompts** that we've validated to produce good results on previous data sets, but because the tool is intended to process a variety of reports, ideal prompts differ between reports. Separate prompts are used for each step of the pipeline, and each prompt represents many potential parameters that report creators can influence. The parameters we've found most useful have been:

- Number of clusters in each level of clustering
- Suggested themes of top-level clusters
- Length of quotations extracted from each respondent

- Bias towards more or less aggressive deduplication

We give report creators basic guidance for editing prompts to produce the best results, but allow them to explore the full parameter space by editing the prompts themselves.

Evaluation

Results of current evaluations

We evaluated TttC in an initial case study of interviews with formerly incarcerated people in Michigan.

Using GPT-4's larger context window and allowing prompt editing enabled a well-organized report with compelling stories from interviews. GPT-4 pulled out sensible topics and subtopics, as judged by the person conducting the interviews, and was largely accurate in its summarization of specific claims.

From a total of 529 claims extracted from interviews, 4.91% of the claims were flagged as inaccurate, miscategorized, or removed. After correcting these claims, the Revised Error Rate (i.e. the proportion of removed claims) was 2.84%.

Areas for improvement include:

- Highlighting interesting and surprising claims
- Removing duplicate claims
- Avoiding LLM-generated language that goes against community norms (e.g. "convict" in this report)

Suggested evaluations for assessing process runs

We are in the process of conducting more scalable evaluations of the TttC process by comparing its outputs to those from simpler NLP methods, such as topic modeling and clustering. Comparing the results side-by-side will allow us to judge the improvement LLMs offer over these older approaches, beyond the automation of analysts' decisions.