Crowdsourced Adaptive Surveys

Yamil Ricardo Velez[1]

[1]***Department of Political Science, Columbia University, Email: yrv2004@columbia.edu***

January 20, 2024

Abstract

Public opinion surveys are vital for informing democratic decision-making, but responding to rapidly changing information environments and measuring beliefs within niche communities can be challenging for traditional survey methods. This paper introduces a crowdsourced adaptive survey methodology (CSAS) that unites advances in natural language processing and adaptive algorithms to generate question banks that evolve with user input. The CSAS method converts open-ended text provided by participants into Likert-style items, and applies a multi-armed bandit algorithm to determine user-provided questions that should be prioritized in the survey. The method's adaptive nature allows for the exploration of new survey questions, while imposing minimal costs in survey length. Applications in the domains of Latino information environments and issue importance showcase CSAS's ability to identify claims or issues that might otherwise be difficult to track using standard approaches. I conclude by discussing the method's potential for studying topics where participant-generated content might improve our understanding of public opinion.

This is a working paper. Do not cite without permission.

Survey research plays a critical role in informing political decision-making and measuring fluctuations in public opinion (Page and Shapiro, 2010). However, traditional surveys may struggle to identify nascent issues that gain traction in the public consciousness due to the inherent time lag between design and implementation. These challenges are especially acute when examining social media trends or political events that can rapidly rise and fall in popularity. Furthermore, even when surveys manage to capture salient issues, the phrasing of questions and the determination of pertinent policy dimensions are frequently guided by subjective criteria set by the survey researchers.

In this paper, I propose a crowdsourced adaptive survey methodology (CSAS) that leverages advances in natural language processing and adaptive algorithms to create participant-generated questionnaires that evolve over time. I use open-ended responses from participants to create question banks comprised of potential survey questions, from which questions are prioritized using a multi-armed bandit algorithm. While the survey is in the field, participants contribute to the question bank and rate existing questions, enabling the algorithm to "learn" emerging issues and claims that resonate with the study population. Even in well-trodden settings such as identifying important issues, the CSAS method produces promising items that are worthy of further exploration (see Figure 1 for a summary of the process and representative issues recovered using the method).

I employ this methodology across two challenging domains: gauging the prevalence of misinformation within minority communities and evaluating issue importance in the aggregate. In an application examining the prevalence of negative beliefs about American political parties and candidates among Latinos, I find that CSAS recovers beliefs about events that would likely escape the notice of survey researchers. Participant-submitted items are also instructive, in that they focus more on party stereotypes or politically harmful – but true – claims, rather than discrete pieces of misinformation. I then use CSAS to construct an adaptive battery of nationally salient issues. Despite seeding the algorithm with popular items from Gallup surveys, I find that popular issue topics based on open-ended responses depart from the set of "most important issues," reflecting concerns over healthcare, inflation, political accountability, and crime.

Across the two studies, highly-rated items reflect salient news events or socioeconomic conditions, showcasing the method's ability to quickly adapt to changes in the information environment. Moreover, the exploration of survey questions using an adaptive algorithm such as Gaussian Thompson sampling produces issue and claim-level estimates for a variety of items; a feature that distinguishes the method from traditional approaches using fixed item batteries where only a set number of items can be examined. Specifically, as the algorithm attempts to "settle" on the best-performing question, the consideration of questions along the way renders it possible to compute prevalence estimates for an extensive set of claims.

The advantages of CSAS are two-fold. First, it enables survey researchers to capture trends in public opinion in real-time, reflecting the public's evolving beliefs and concerns. Second,

it democratizes the survey process by allowing respondents to contribute to instruments. These benefits come at little cost in terms of survey length. Researchers can set the number of "dynamic questions" in advance, and select the appropriate algorithm for determining how questions should be prioritized. For example, one can rely on a set of tried-and-true items, while allocating one or two survey slots for dynamic questions.

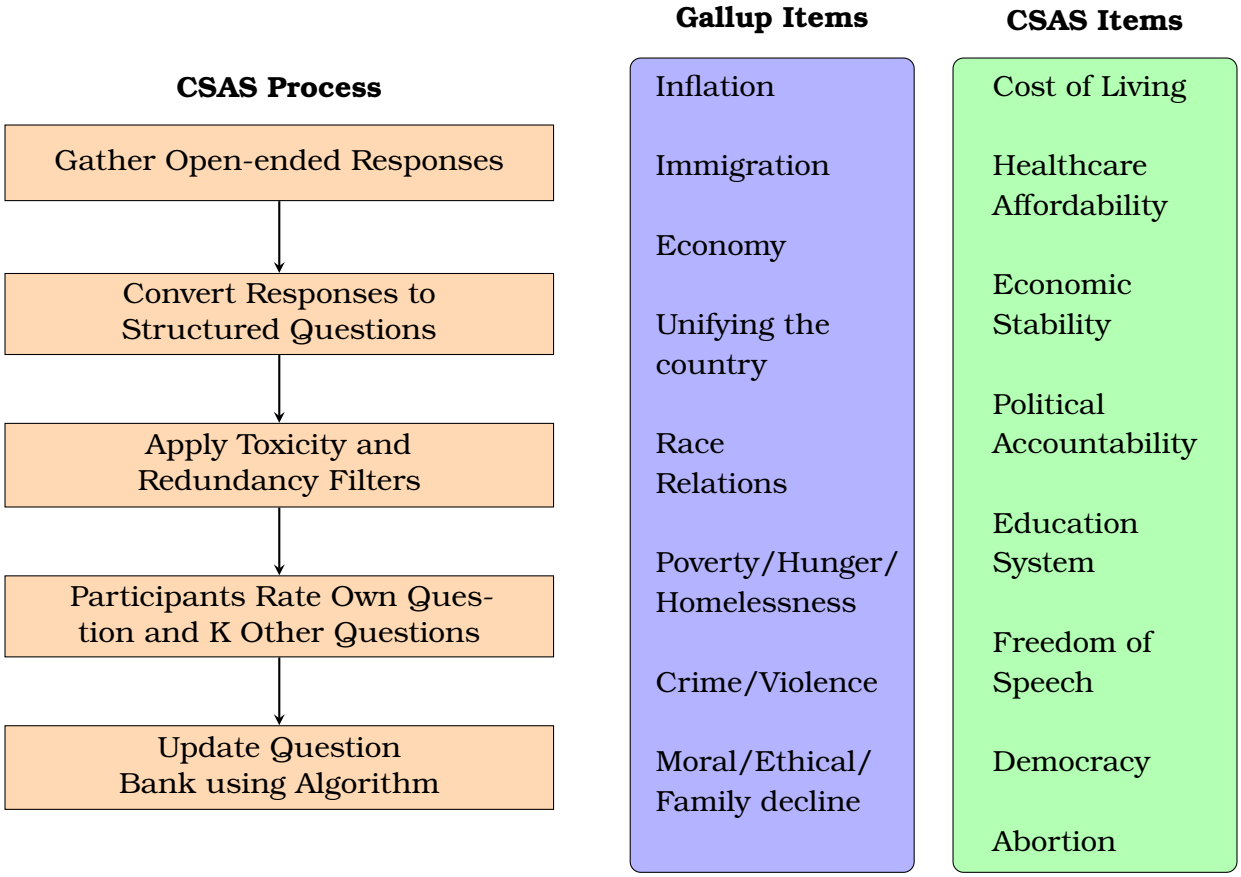

Figure 1: CSAS Process Flowchart and Representative Set of Issues uncovered by CSAS.

I close the paper by discussing how the CSAS method can be used to study other topics where information environments are evolving and scholarly assumptions about constructs might not match how study populations perceive them. Moreover, I highlight potential guardrails that researchers can implement to ensure a "clean" question bank with respect to quality, redundancy, and toxicity. Given rapidly evolving information environments in the age of social media, it is imperative that survey research is capable of adapting to these changes: the CSAS method is one step in that direction.

1 Dynamic Survey Methodologies: Existing Approaches

Influential texts on survey design stress the importance of a "tailored" approach to recruitment and stimuli (Dillman et al., 2014). Adapting questionnaires to respondents can enhance measurement and satisfaction. For instance, branching questions can reduce respondent burden and measurement error by eliminating irrelevant sub-questions (Krosnick and Berent,

1993; Dillman and Christian, 2005). Questions measuring recognition of elected officials and voting in subfederal elections can be tailored using location to produce more relevant question content via "piped in" text (Ansolabehere and Schaffner, 2009). These examples showcase how surveys already possess dynamic elements that respond to user input or data.

1.1 Computerized Adaptive Tests

Scholars have recently developed methods for carrying out computerized adaptive tests (CATs) in public opinion surveys (Montgomery and Cutler, 2013; Montgomery and Rossiter, 2020, 2022). CAT algorithms "respond to individuals' prior answers by choosing subsequent questions that will place them on the latent dimension with maximum precision and a minimum number of questions." (p. 173). CATs are typically employed using latent scales, where the goal is to optimize the number of questions. Montgomery and Cutler show that CATs offer a superior approach to traditional static batteries, and these tools can be easily implemented in survey software such as Qualtrics (Montgomery and Rossiter, 2022).

CATs rely on pre-existing measurement scales (e.g., political knowledge, personality batteries). However, in settings where the objective is to capture emerging issues or changes in the information environment, scholars and practitioners may want to learn about the prevalence of discrete beliefs, some of which cannot be known in advance. Thus, while CATs allow us to enhance precision when estimating latent traits, there are settings where the question bank cannot be fixed in advance and describing the nature of discrete items, rather than estimating positions on latent scales, is the primary objective.

One such setting is research on misinformation and rumors, where survey questions are often developed with assistance from fact-checking organizations (Vosoughi et al., 2018) and social media databases (Sanderson et al., 2021). Though these sources can be useful, they are limited in their timeliness and reach. First, fact-checks are lagging indicators, given that news organizations often wait until claims are viral before correcting them (Mena, 2019). Claims may be irrelevant by the time surveys are in the field. Second, identifying public claims through social media databases in real-time is possible, but the rising costs of APIs for platforms such as Twitter have made them less accessible to researchers.[1] Finally, the inclusion of questions can be affected by the institutional resources of fact-checking organizations and their assumptions of the study population. For example, fact-checkers focusing on Latino communities might prioritize immigration-related claims and inadvertently overlook other significant types of misinformation within those communities, such as health-related or economic misinformation.

1.2 Wiki Surveys

The Wiki survey is a collaborative survey methodology that incorporates user input, and thus, may be more flexible than approaches using fixed batteries (Salganik and Levy, 2015). Drawing

[1] https://www.theverge.com/2023/3/30/23662832/twitter-api-tiers-free-bot-novelty-accounts-basic-enterprice-monthly-price

inspiration from online information aggregation portals such as Wikipedia, Wiki surveys balance three principles: greediness, collaborativeness, and adaptivity. Greediness refers to capturing as much information as respondents are willing to provide, collaborativeness refers to allowing respondents to modify instruments (e.g., proposing new items), and adaptivity refers to optimizing instruments to select the "most useful information." While wiki surveys have shown promise in facilitating collective decision making (e.g., allowing users to vote on policies – both pre-determined and user-generated – that should be considered by local governments), existing applications rely on pairwise comparisons between options provided by survey designers and participants. However, pairwise comparisons may not be useful in settings where options can be accorded the same weight, the decision is not zero-sum, and outcomes can be more accurately measured on an ordinal or continuous scale.

In sum, existing adaptive survey methods offer distinct advantages over traditional surveys with respect to dimensions such as efficiency, but both have limitations when applied to certain question formats or research settings. CATs excel in improving precision for scales with fixed items, but may not be appropriate for settings where the question bank cannot be determined in advance. Wiki surveys, owing to their collaborative and adaptive nature, provide a means for creating surveys that evolve with user input but may not always be suitable, especially when attitudes, behaviors, or beliefs are best measured on an ordinal or continuous scale rather than through pairwise comparisons.

2 The Crowdsourced Adaptive Survey Method

Existing principles of Wiki surveys such as greediness, collaborativeness, adaptivity can guide the construction of methodologies achieving similar objectives. Building on the Wiki survey and other attempts to insert dynamic elements into existing surveys (e.g., CAT approaches), I develop a crowdsourced adaptive survey (CSAS) method that enables question banks to evolve based on user input and does not impose constraints on question formats. I leverage generalized pre-trained transformers (GPTs) to convert open-ended text produced by participants into questionnaire-friendly formats (see (Velez and Liu, 2023) for an example) and implement adaptive algorithms (Offer-Westort et al., 2021) to identify best-performing questions from a question bank. First, each respondent answers an open-ended question about a given topic that is cleaned, summarized, and converted into a structured survey question format. Second, respondents rate their own questions, along with k other questions from a question bank generated by previous participants. Finally, ratings for user-submitted questions and k questions drawn from the question bank are updated using a multi-armed bandit algorithm, adjusting the probabilities of presenting these questions to future participants in subsequent surveys.

The three essential features of the proposed method are open-ended questions, a question bank, and a multi-armed bandit algorithm.[2] Table 1 displays the different steps of the

[2]One could also sample from the question bank using a uniform distribution. However, this entails devoting similar resources to both poorly performing and better performing items.

Table 1: Overview of the CSAS Methodology

Step	Model	Details
Collect Open-ended Responses	-	Participants provide responses to open-ended questions.
Process Open-Ended Responses	LLM (e.g., OpenAI's GPT-3.5)	Open-ended responses are processed and converted into structured survey questions using LLMs.
Apply Filters	Document Embeddings (e.g., OpenAI's *text-embedding-ada-002*)	**If flagged as redundant,** do not add to question bank.
	Toxicity Detection (e.g., OpenAI's moderation endpoint)	**If flagged as toxic,** do not add to question bank.
		If not flagged as redundant or toxic, add structured text to question bank.
Participant Ratings	-	Participants rate their own question and k other questions selected using a multi-armed bandit.
Update Question Bank	Multi-Armed Bandit Algorithm	The question bank is updated using ratings, prioritizing the most highly rated questions.

CSAS method: eliciting potential items using open-ended questions; processing and filtering candidate items; and optimizing question selection. I walk through each step in turn.

The open-ended question is used to query participants in a free-form manner about a given topic, issue, or claim. For example, a study examining rumors or misinformation could ask participants to provide a list of claims that present parties or candidates in a negative light. Since these data will be unstructured to some degree, introducing heterogeneity on dimensions such as length, style, and grammar, a response conversion stage is typically necessary. At this stage, researchers can refine the unstructured text to match their objectives. For example, given that Likert scale questions tend to be brief, one can convert the open-ended response into a sentence-long summary using a Generative Pre-Trained Transformer (GPT). GPTs are large-parameter language models that can perform various tasks such as text prediction, summarization, and classification at levels that mirror human performance (Vaswani et al., 2017; Radford et al., 2018). Recent years have seen the development of a diverse range of GPTs, with proprietary models like OpenAI's GPT-3, Anthropic's Claude, and Google's Gemini

demonstrating robust capabilities across various text manipulation, summarization, and generation tasks. In parallel, "open source" and "open weights" models, including Meta's Llama 2 and Mistral AI's Mixtral, have emerged, offering competitive performance in similar domains (see Appendix C for a detailed discussion and direct comparison of these models).

Once open-ended response data are in a usable, structured format, they can be included in a question bank, which is a collection of structured survey questions or statements. Though inclusion of questions can be unrestricted at this stage, such that the question bank includes all responses that have been converted into a structured format, researchers may want to impose additional restrictions to reduce redundant questions and apply filters to ensure that survey questions meet the researcher's objectives. For example, a scholar studying misinformation may want to prioritize the inclusion of verifiable statements or claims versus value judgments. They may also want to minimize exposure to toxic and offensive content that can be produced by antagonistic respondents (Lopez and Hillygus, 2018).

Focusing on redunancy first, if two respondents submit responses about Democratic spending priorities with only minor differences in wording, it may be unnecessary to include both questions in the same question bank. Moreover, multi-armed bandit algorithms can struggle to identify the best-performing arm when arms are equally matched (i.e., "no clear winner"). Assuming near-identical questions are rated similarly, this increases the odds of failing to identify the best-performing item (Offer-Westort et al., 2021, 832-833). Given that open-ended responses are unlikely to be *exact* duplicates, filtering using more sophisticated text analysis methods such as document embeddings can be helpful. Document embeddings locate texts on a high-dimensional space and can be used to identify similarities between texts (Rheault and Cochrane, 2020). By applying document embeddings, researchers can quantify the similarity between different questions, even when the wording varies, and retain only questions that surpass a pre-defined threshold of similarity.

Researchers may also want to apply additional filters to ensure that questions meet pre-specified criteria on dimensions such as relevance and toxicity. For example, for a survey of issue importance, a survey researcher may choose to exclude responses referring to the personal characteristics of politicians, and retain only those that refer to policies. Given that this is a classification task, one may opt for a supervised learning model trained on a labeled dataset or a GPT model, among other options. The same holds for identifying and removing toxic responses. Given that a small percentage of respondents resort to "trolling" or toxic behavior, ensuring other participants are not exposed to harmful content is paramount (Lopez and Hillygus, 2018).

The next challenge is in how these questions are presented and selected within the survey. Instead of presenting questions with equal probability, and thus "wasting" responses on questions with low ratings, multi-armed bandit algorithms can be leveraged to identify a set of "best-performing" questions from the pool. Multi-armed bandits have been applied in recent research on adaptive experiments, where the goal is to identify the best-performing

intervention from a set of interventions (Offer-Westort et al., 2021). Rather than assign uniform probabilities to each arm, multi-armed bandit algorithms such as Thompson sampling "learn" the most-effective arm over a set of trials by balancing the trade-off between exploitation (i.e., concentrating on promising arms) and exploration (i.e., experimenting with additional arms to assess their potential). In experimental settings, outcome information is used to determine future treatment allocation, such that arms associated with higher scores on pre-specified metrics are generally prioritized. When measuring the prevalence of a false or contested claim, for example, Thompson sampling can also be employed. Instead of interventions, different survey questions or statements can be considered "arms." By iteratively selecting questions based on their estimated prevalence, Thompson sampling can adaptively allocate more participants to claims that resonate with the sample.

3 An Application to Latino Information Environments

I use the CSAS method to identify rumors, negative political claims, and misinformation reaching Latinos, a group that has received attention among journalists and social scientists due to potential misinformation campaigns targeting the community (Cortina and Rottinghaus, 2022; Velez et al., 2023). I focus on Latinos for two reasons. First, fact-checking is still a relatively new institution within Latino-oriented media (Velez et al., 2023, 790). Existing organizations might overlook important claims that circulate within the community due to resource constraints and the possibility that best practices for verification have not yet been identified. Second, private, encrypted messaging applications used by Latinos such as WhatsApp and Telegram may hinder the detection of false claims (Lee et al., 2023). In contrast to misinformation that is transmitted through social media such as Instagram, Twitter, and Facebook, these private channels can operate largely beyond the reach of fact-checkers and researchers. These two factors create unique challenges for identifying and addressing misinformation within the Latino community.

Implementing this more "bottom-up" approach to misinformation detection, I fielded a survey of 319 self-identified Latinos using the survey platform, CloudResearch Connect, from July 6-7, 2023. First, participants were asked two open-ended question regarding negative claims they had heard about Republicans and Democrats. These claims were then passed to a fine-tuned generative pre-trained transformer model, OpenAI's *ada*, that classified the text as a "verifiable claim." Fine-tuning was necessary to ensure that questions entering the question bank were falsifiable political claims, rather than value judgments (e.g., politicians are evil). To carry out the fine-tuning step, a mixture of researcher-provided examples and participant-provided examples (N=87) were hand-coded to indicate whether a claim was, in principle, falsifiable. Hand-coded classifications were then used to fine-tune the *ada* model.

Given that similar questions can be introduced by more than one respondent, and to guard against exposing participants to toxic content, I also used a similarity and toxicity filter before adding items to the question bank. For each submitted question, I used OpenAI's embedding application programming interface (API) to generate a 1536-dimensional embeddings vector,

calculated the cosine similarity between the new item and other items in the question bank by retrieving the five nearest neighbors, and filtered out questions with a similarity score above .90.[3] I also used OpenAI's moderation endpoint to filter out offensive and toxic claims. Claims that passed these filtering steps were added to an item bank that allowed them to be rated by future participants.

After the submission and cleaning step, participants rated their own submissions, along with four items from the question bank and six items capturing conspiracy beliefs and more common misinformation items (e.g., Covid-19 vaccines modify your DNA). All of the questions were presented in a matrix format, with a four-point accuracy scale ranging from "not at all accurate" to "very accurate." Since the first set of participants did not have user-submitted claims to rate, the question bank was seeded with an initial set of claims. Four claims were taken from the front pages of Latino-oriented fact-checking websites (i.e., Telemundo's T-Verifica, Univision's El Detector) to create the initial question bank.

Gaussian Thompson Sampling was then used to determine which questions to present to subsequent participants. Though traditional Thompson Sampling requires binary outcomes due to its use of a Beta distribution (Offer-Westort et al., 2021), Gaussian Thompson Sampling can be leveraged to accomodate continuous outcomes (Agrawal and Goyal, 2017).[4] Gaussian Thompson sampling was implemented in real-time using an API created by the author. In contrast to previous applications of adaptive experiments in political science that have leveraged batch-wise Thompson sampling, probabilities were adjusted at the respondent level.

4 Results

Figure 2 displays mean accuracy estimates for claims receiving more than ten ratings.[5] The highest-rated claims covered information about both Republicans and Democrats. Party stereotypes about Republican positions on moral issues were rated as highly accurate ($\bar{x}$ = 3.54; SE = .07), as were claims that Trump used the presidency to enrich family and friends ($\bar{x}$ = 3.54; SE = .159) and possessed classified documents in his vacation homes ($\bar{x}$ = 3.43; SE = .121). Other highly rated claims focused on Republicans such as Ted Cruz ($\bar{x}$ = 3.40; SE = .049), extreme policy positions (e.g., a claim that Trump plans to deport all undocumented people ($\bar{x}$ = 3.32; SE = .313), and President Biden's son, Hunter Biden ($\bar{x}$ = 3.24; SE = .201). The lowest-rated claims typically involved false statements, allegations, or extreme descriptions of issue positions. Claims that scored especially low on perceived accuracy included "This year, there was a major explosion close to the White House" ($\mu = 1.50$; SE = .15), "There were no wars during the Trump presidency" ($\bar{x} = 1.74$; SE = .17), "Republicans support going back to the 1776 constitution." ($\bar{x} = 1.80$; SE = .33), "Trump will deport **all**

[3] In initial tests before data collection, lower similarity thresholds such as .80 were found to produce false positives (e.g., classifying "Biden is a tax cheat" and "Trump is a tax cheat" as sufficiently similar) and higher similarity thresholds such as .95 produced false negatives.

[4] As in other research (Offer-Westort et al., 2021), a probability floor of .01 was employed to guarantee that every item in the item bank retained a non-zero chance of being presented to participants.

[5] Mean estimates are weighted by the inverse probability of selection, as in Offer-Westort et al. (2021).

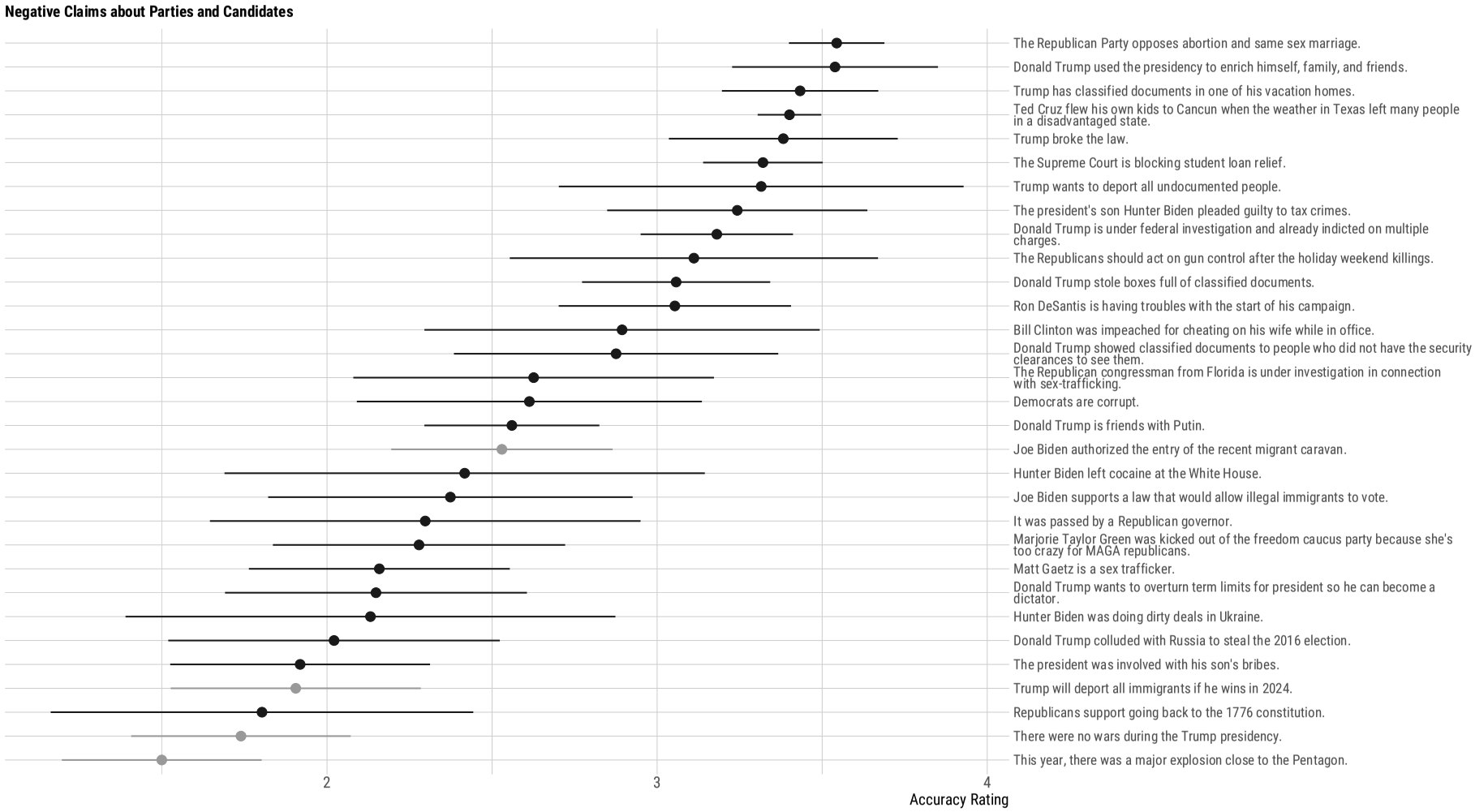

Figure 2: IPW-weighted estimates of survey questions measuring negative political claims with corresponding 95% confidence intervals. Items in gray were initial seed items based on fact-checked claims produced by Latino-focused fact-checking organizations. Black items are participant-generated items.

immigrants if he wins in 2024" ($\bar{x} = 1.91$; SE = .19), and "The president was involved with his son's bribes" ($\bar{x} = 1.92$; SE = .20).

Whereas highly rated claims mostly reflected actual events or generalizations of party positions, claims with lower accuracy ratings typically involved verifiably false information or oversimplifications. These findings are instructive in that they reveal a level of discernment in the aggregate. Objectively false claims are generally seen as less credible by participants. Instead, higher accuracy ratings are observed when participants judge claims that are widely reported in the news (e.g., Ted Cruz's cancun trip) or that reflect commonly-held perceptions of party positions (e.g., Republicans opposing aboriton and same sex marriage). Though the initial seed claims based on fact-checks were small in number, the analysis revealed a surprising disparity: the most readily believed claims were not identified by the fact-checking organizations, but rather originated from other participants. This reinforces the utility of the method, as it highlights the potential for crowdsourcing to identify and address misinformation that can complement traditional fact-checking efforts.

Moving beyond the aggregate estimates, one can further segment the sample into subgroups. For example, one can identify whether highly rated claims are uniformly perceived as accurate or whether subgroups that may be more prone to seeing misinformation might also differ in their accuracy ratings. Previous research, as well as popular discussions of Latino-targeted misinformation, have highlighted WhatsApp as a potential vector for misinformation. In the survey, participants were asked to rate the credibility of information on the platform on a five-point ordinal scale ranging from "Trust" to "Distrust."

Comparing participants below and above the median of WhatsApp trust, several patterns emerge. For the lowest rated claims, there is a general consensus regarding accuracy. However, for the more highly rated claims involving party stereotypes, newsworthy scandals, and salient allegations, differences between the two groups can be seen. For example, those who trust WhatsApp were less likely to have rated true claims about the Supreme Court blocking student loan relief, Ted Cruz visiting Cancun with his family during widespread blackouts in Texas, and Donald Trump being under investigation as accurate relative to other users. These gaps, of course, could reflect other social or demographic causes, but they point to the possibility that observed differences between WhatsApp users and non-users in factual accuracy can be derived from gaps in knowledge of factual political information, rather than a wholesale adoption of misinformation. The relatively flexible nature of the approach, which allows for the inclusion of blatantly false misinformation as well as misperceptions or overgeneralizations allows us to arrive at a more nuanced description of information environments than if we solely measured accuracy ratings for factually verified content.

Our results underscore the complexity of the information environment among Latinos. They engage with a variety of narratives, some of which portray different parties negatively, but also reflect actual events. Understanding this diversity is essential for creating informed strategies to address misinformation and enhance democratic discourse. However, although

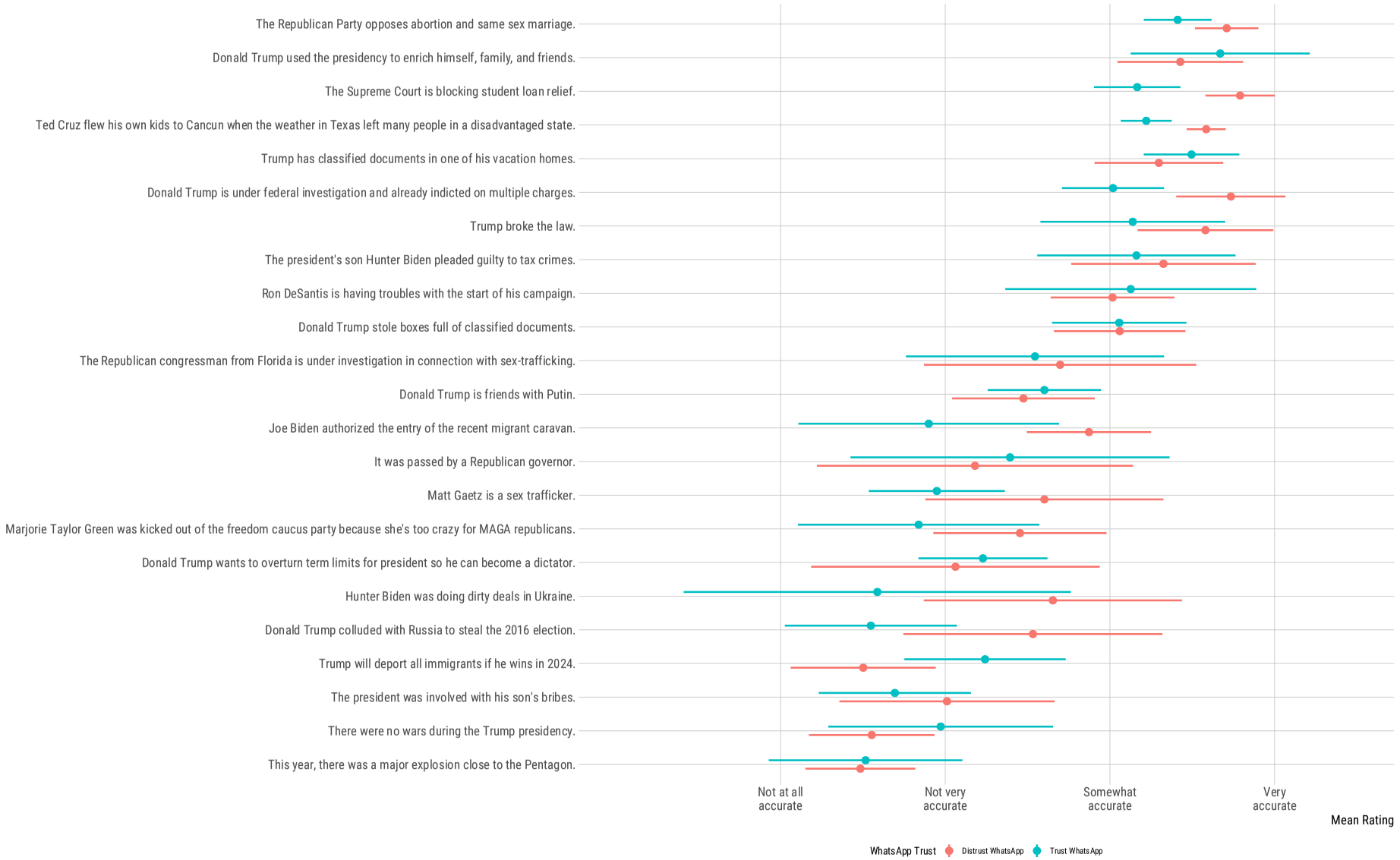

Figure 3: IPW-weighted estimates of survey questions measuring negative political claims across levels of WhatsApp trust with corresponding 95% confidence intervals. Presented claims were rated by at least 10 participants in each category.

the CSAS approach can recover useful questions, filtering can be imperfect. Claims like "The Republican party opposes abortion and same sex marriage" departs from more salacious claims about politician behavior that tend to receive attention from fact-checkers. Moreover, there are low-quality items that were added to the question bank despite the filtering. One such example was "It was passed by a Republican governor." Still, this item received low ratings, and thus, was not seen by many participants. This further emphasizes the benefits of using an adaptive algorithm, given that items with lower ratings will be prioritized less.

5 An Application to Issue Importance

Though the CSAS method can be helpful in crowdsourcing questions that resonate within marginalized communities, identifying salient issues is another possible application. That is, rather than fixing a set of issues for participants to rate, as is often done in research on issue attitudes and importance (Ryan and Ehlinger, 2023), one can use the method to provide an issue bank for participants, much like the question bank comprised of claims in the previous study. This can be helpful in estimating support for idiosyncratic issues that may not appear on the national agenda, but still inspire strong reactions among "issue publics" (Elliott, 2020) or serve as issues that could be mobilized in future elections, corresponding to the elusive notion of "latent opinion" described in Key (1961).

Since 1935, Gallup's Most Important Problem question has been used to approximate the issue priorities of the American public. Using an open-ended format, participants are asked "What do you think is the most important problem facing this country today," with responses being classified by hand into a set of categories corresponding to broad issue or policy areas. Though the question has undergone several changes over the years, its consistent use across surveys has enabled scholars to construct time series of a host of issues (McCombs and Zhu, 1995), and has served as the empirical foundation for numerous studies using national issue salience as an outcome (Dunaway et al., 2010) and independent variable (Ostrom Jr et al., 2018).

Despite its adoption in public opinion research, the measure has been criticized for being an imperfect proxy of national issue salience. As Wlezien (2005) argues, the question asks respondents to provide information on two distinct concepts — importance and "problem status." While some respondents may interpret the question as one where they can offer a personally relevant issue, others may interpret it as an opportunity to highlight a problem affecting the nation as a whole. Focusing on defense spending, Wlezien (2005) finds that increases in MIP estimates regarding foreign policy are predicted by indicators of "problem status" (e.g., dislike of foreign adversaries), but fail to explain defense spending preferences once indicators of "problem status" are included. Though these analyses suggest that revisions of question wording could yield more accurate responses, Jennings and Wlezien (2011) find that a most important issue (MII) question that eliminates the use of "problem" language performs similarly over time, and thus, "an effective measure of issue importance remains elusive" (p. 546).

More recently, Ryan and Ehlinger (2023) make a case for moving beyond fixed questions about a slate of national issues and hand-coded classifications based on open-ended questions. Like the MIP and MII, Ryan and Ehlinger (2023) use open-ended questions to elicit issues from participants. However, in contrast to these approaches, they convert unstructured text by hand into rating scales, and present these scales to participants in a follow-up wave. This approach provides a richer amount of information about the *degree* to which an elicited issue may hold importance to an individual. Recent studies in this vein have recovered high levels of stability (Velez and Liu, 2023) and sizable causal effects for this "core issue" in conjoint settings (Ryan and Ehlinger, 2023; Velez, 2023), suggesting that such a measure might serve as a candidate for the elusive concept of issue importance.

Applying the CSAS method in this context is straightforward. Participants report personally relevant issues using open-ended responses; open-ended responses are transformed into topics for rating and included in a question bank; and participants rate the different issue topics with Gaussian Thompson sampling used to optimize question selection. In contrast to the misinformation setting, where it may be more difficult to select seed questions because of a lack of high-quality data on Latinos, the top-performing MIP items produced by Gallup can be used as a baseline for comparison. Specifically, we can seed the question bank with a set of popular issues and assess whether the crowdsourced issue topics receive higher importance ratings. The issues were the following: 'Immigration', 'Economy', 'Race Relations', 'Poverty', 'Crime', 'Ethics, Moral, and Family Decline', 'Unifying the country', and 'Inflation.'

From September 11 until September 13, 2023, I collected data from a national quota sample balanced on age, race, and gender using CloudResearch Connect (N=820). The question bank was seeded with eight popular issue areas taken from Gallup. In this study, a more capable LLM, OpenAI's GPT-4, was used to convert unstructured text into issue topics and filter out redundant issues in a single operation.[6] Recent studies have shown that in the task of classifying open-ended text to identify "most important issues," the efficacy of Large Language Models (LLMs) like GPT-4 is on par with classification algorithms trained on thousands of examples, achieving performance levels marginally below that of human evaluators (Mellon et al., 2022). User-submitted items spanned salient issues such as abortion and climate change, and less salient issues such as universal basic income and deficit spending. Each participant rated their own issue, along with eight others that were determined using Gaussian Thompson sampling, as in the first study.

5.1 Results

Figure 4 displays mean estimates for issue topics receiving 50 or more ratings. As shown in the figure, the highest-rated issues were focused on the economy and health care, with issues such as "Cost of Living," "Healthcare Affordability," "Healthcare Costs," "Economic

[6] In the previous study, similar items from the question bank were identified using cosine similarity, specifically by retrieving the five nearest neighbors. However, instead of filtering based on a set value, these items were directly included in the prompt and GPT-4's instructions were to avoid including similar issues that had already been mentioned.

Stability," and "Universal Healthcare." Issues rated lower on importance include more social and culture issues such as immigration and voting rights. The MIP issue topics of "Race Relations" and "Ethics, Moral, and Family Decline" appeared among the lower issue ratings, with topics related to immigration (i.e., "Illegal Immigration," "Border Security") being given lower ratings. In the list of highly rated topics, we see issues that would likely not appear in traditional issue importance barries such as "Mental Health Access," "Privacy Protections," and "Candidate Transparency." Moreover, the frequent mention of various economic and healthcare dimensions is instructive, revealing a trend where socioeconomic issues like healthcare and cost of living are taking center stage.

As in the previous study, I present issue importance ratings across relevant subgroups. Figure 5 presents mean estimates of issue importance among Republicans and Democrats. Estimates are ordered by the absolute size of the partisan gap. As seen on the bottom of the figure, there some areas of issue consensus across the two parties, namely those related to healthcare costs, political polarization, and economic issues (i.e., cost of living, economic stability). However, by and large, issue importance ratings exhibit considerable gaps across partisans. For example, Democrats generally rate inequality, progressive policies (e.g., universal healthcare), gun control, and social issues (e.g., women's rights, race relations) as more important than Republicans, whereas Republicans rate issues involving morals, government spending, and immigration higher than Democrats. While partisan differences in issue positions are well known and expected, the dynamic nature of adaptive algorithms allows us to delve into a broader range of "issue gaps" than traditional surveys with fixed sets of questions.

To what extent does our method capture personally meaningful issue positions? As mentioned earlier, recent studies suggest that issues elicited in an open-ended format score high on dimensions such as attitude strength, certainty, and reliability over time. Still, we can use the data from the existing study to verify whether this is the case. First, we find that on a 1-5 scale ranging from "Not at all important" to "Very important," the mean importance score for the personal issue is 4.7 (s = .72). These ratings are generally higher than importance ratings for other issues in the question bank ($\bar{x}$ = 3.43; s = 1.24). Moreover, the mean ranking for the "personal issue" is 1.43 in the list of nine rated issues.

Overall, the findings highlight the ability for CSAS to identify idiosyncratic issues that may be deeply important to segments of the population. For example, universal healthcare and other health-related issue topics received some of the highest importance ratings. Though healthcare policy is not currently a salient issue in American politics, the CSAS method reveals that it remains a deeply important concern for individuals in our sample. The findings offer a valuable glimpse into less salient issue priorities and suggest the potential for the CSAS method to complement traditional measures of public opinion.

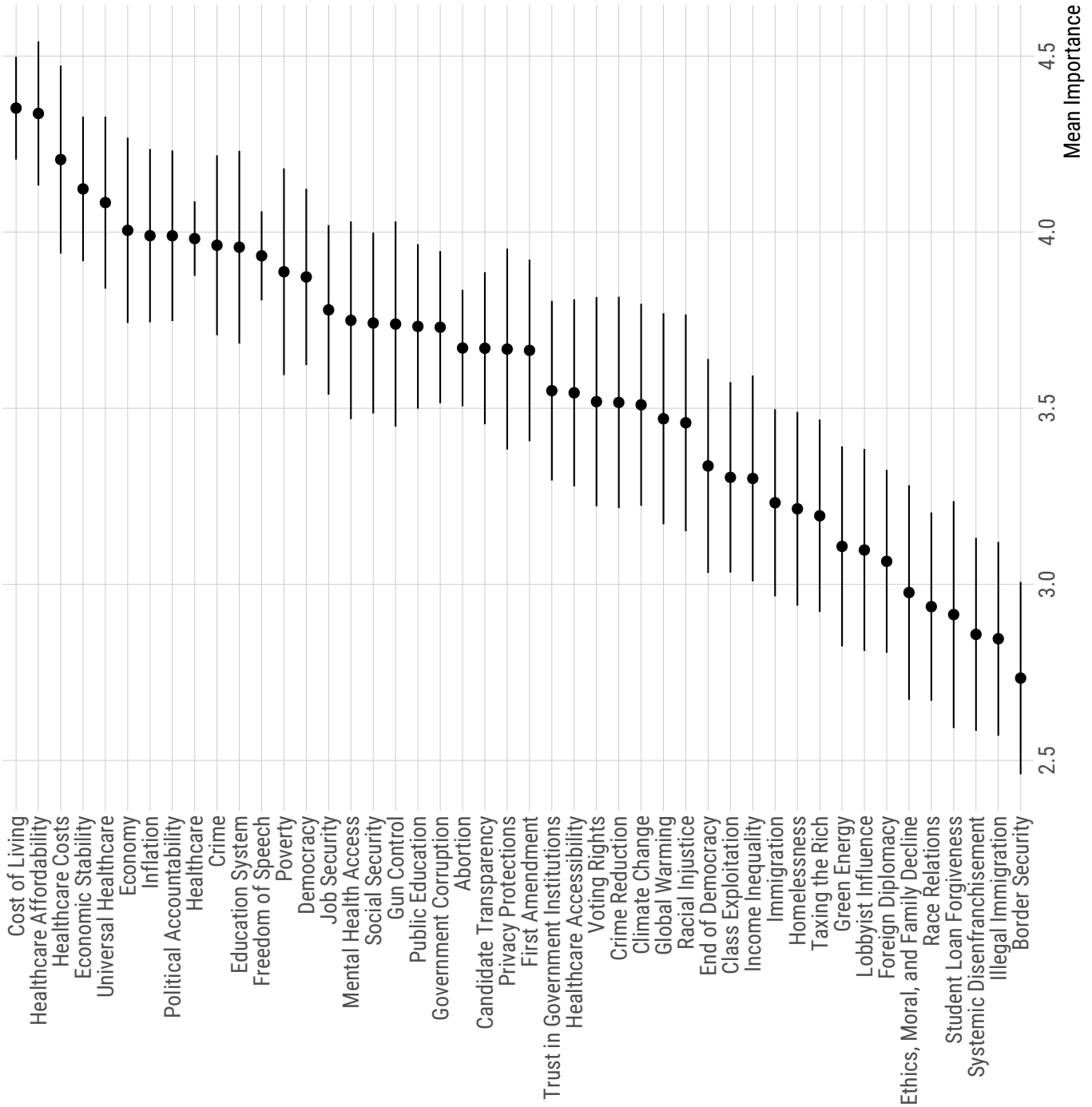

Figure 4: IPW-weighted estimates of survey questions measuring issue importance with corresponding 95% confidence intervals.

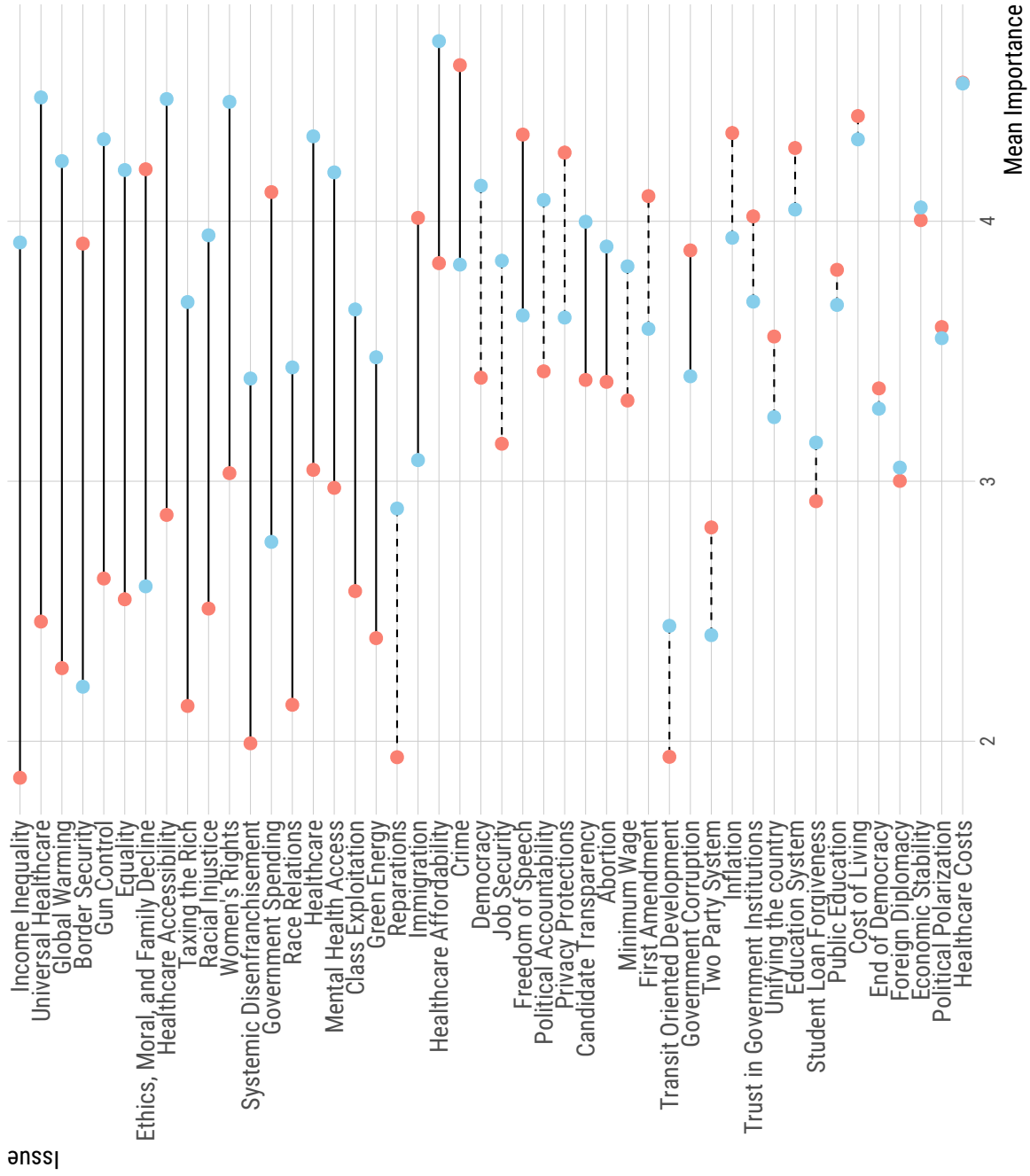

Figure 5: IPW-weighted estimates of survey questions measuring issue importance across Republicans and Democrats with corresponding 95% confidence intervals. Statistically significant (insignificant) differences at conventional levels of significance (α = .05; two-tailed) are depicted using solid (dashed) lines.

6 Concerns and Caveats

6.1 Is the CSAS method compatible with traditional survey design?

Despite the limitations of "static surveys" in identifying changing information environments or measuring responses within niche communities, the two approaches are not at odds. Researchers can decide the number of adaptive questions, and include these questions in standard batteries. For example, in the Latino survey, participants rated a pre-existing set of false claims and conspiracies, along with an adaptive set, in a question matrix. Before using this method, scholars should determine whether the marginal benefit of having a designated slot for exploratory questions is worth the survey time and cost. A distinct advantage of multi-armed bandits is that several items can be explored despite having a smaller set of "dynamic survey slots." The two approaches can also work in tandem when there are multiple phases of data collection. An initial wave (or pilot) could use CSAS to develop a fixed battery of questions for future waves, functioning much like pilot studies that gather open-ended data to inform scale construction (Weller, 1998). With CSAS, however, future surveys can be designed not only with open-ended content in hand, but question ratings and posterior probabilities that a given question is the "best-performing question." This approach may be optimal if researchers prefer to split their research process into exploratory and confirmatory stages, as is recommended in Egami et al. (2018). Pre-registration across these different stages could lend more credibility to conclusions derived using this method (see Offer-Westort et al. (2022) for an example).

6.2 Prompting participants

Leaving aside potential challenges with prompting large language models, prompting participants is also an important step in the process. For example, in the case of misinformation, asking about false claims directly would likely fail to elicit meaningful items, given that participants may not know the veracity of statements or may be inclined to disregard validity when thinking about congenial pieces of information. Instead, asking about claims that possess common features of misinformation, such as negativity, appear to be more fruitful (Carrasco-Farré, 2022). In extending this approach to construct other kinds of question banks, researchers should carefully consider whether the questions they create may diverge significantly from how respondents conceptualize them. Though best practices involving closed-ended questions are well-developed, the use of LLMs and other text analysis methods to extract meaningful information from text suggests that more thought and care may need to go into identifying optimal open-ended formats. Here, lessons from scholars using semi-structured and unstructured interview modes could prove useful (Leech, 2002).

6.3 Late arrivals

Since later submissions to the question bank are reviewed by fewer respondents, there's a risk of not identifying the most effective question. In experimental settings where the determination of optimal treatment allocation may be vital, such as the testing of pharmaceutical

interventions, this risk would represent a significant limitation for the approach. However, it is worth highlighting the exploratory nature of the proposed method, and how it performs against the alternative of pre-determined fixed question batteries. In the absence of an adaptive survey design, the selected questions are assumed, by design, to be the best-performing set. If one compares a fixed battery to a dynamic battery comprised of the same items, the dynamic battery, with its ability to adapt and explore new questions, can potentially uncover additional questions that a static design may ignore. Therefore, the risk of failing to identify useful survey items may be even higher with a static design. Indeed, though "seed questions" included in the Latino survey were drawn from Latino-oriented fact-checking sources (i.e., Telemundo's T-Verifica, Univision's El Detector) and drawn from Gallup in the issue importance study, user-submitted questions were generally rated as more accurate and important. This further highlights the inductive advantages of the method. Researchers can include questions that they possess strong priors about, and allow those questions to be tested alongside new submissions. This creates an environment where established, theoretically grounded items are examined in tandem with potentially novel items. Items introduced toward the end of data collection might also be incorporated into future waves, leveraging the information gleaned from existing ratings. This being said, the method seems well-suited to cases where items cannot be easily constructed due to a lack of relevant information sources, as in the case of Latino-targeted misinformation, or where opinion is in flux, but may not be as germane when validated items exist or use of consistent items over time is the objective (e.g., partisan identification).

6.4 Toxicity and Quality of Responses

The CSAS method thrives on user-generated questions, but one might contend that this also poses its own set of problems. For example, toxic user-generated content, such as hate speech, and low-quality content, such as gibberish, could pollute the question bank and compromise the validity of the approach. As mentioned above, there are several mitigation strategies that can and should be implemented. First, by clearly defining what counts as an issue or a claim, one can simultaneously filter out expletives, racial slurs, or gibberish. These guidelines can be implemented by fine-tuning models or through the use of in-context learning, where examples are directly used to guide completions by including them in the prompt (see prompt examples in Appendix B). Relying on moderation endpoints is another mitigation strategy. OpenAI offers moderation models that are designed to detect and filter out undesirable content when using their API.[7] In many settings, proceeding with automated content filters may be sufficient. At least in the case of misinformation and issue topics, zero items including slurs, offensive, or violent content were included in either of the two question banks despite no manual review of items. Moreover, in the misinformation study, some items

[7] Proprietary models such as Google's Bard and Anthropic's Claude, as well as OpenAI's GPT-4, commonly refuse to produce toxic content even without the use of moderation endpoints. There are also fine-tuned chat versions of open-source models such as Llama 2 that are trained to produce less harmful content.

were not added to the question bank because they mentioned abuse or violence (e.g., "Ted Cruz is the Zodiac killer"). Though these "false negatives" were rare (12% of submissions), they highlight the necessity of a balanced approach in content moderation (see Appendix D). Still, in more sensitive contexts, combining automated filtering with human moderation might be necessary to ensure accuracy and appropriateness.[8] Incorporating "human-in-the-loop" elements could help ensure that open-ended responses are legible, not harmful, and provide relevant information.

6.5 Costs

Though serving large language models to several users at once on personal computers is still largely cost prohibitive, various APIs have emerged that allow scholars to interact with proprietary models such as OpenAI's GPT-3.5 Turbo and GPT-4, as well as open-source alternatives such as Meta's Llama 2 and MistralAI's Mixtral. At present, these models are cost effective for use in survey settings. Two studies with 100 CloudResearch Connect participants each were carried out to precisely estimate the inference costs of using CSAS with OpenAI's GPT-4 and MistralAI's Mixtral, a high-performing open-source model, at the time of the study. The total cost of inference on AnyScale for processing 100 open-ended responses was 5.4 cents, with a cost per participant of 0.005 cents. In comparison, the total cost of inference on OpenAI amounted to 14 cents, with a cost of 0.01 cents per participant.

6.6 Implementation

The method requires interacting with APIs (Application Programming Interfaces) and managing a personal database, which can be daunting for those without a background in programming or database management. Though the implementation used in the first two studies relies on PHP and MySQL, more flexible frameworks exist that can be deployed across a variety of providers. To facilitate adoption of the method, a Django application with a graphical user interface to save user settings and deploy the model on popular survey software has been developed (see Appendix E). Users can use the GUI to submit an initial set of seed items, select an API provider, modify prompts depending on objectives (e.g., summarizing claims versus issues), and various other parameters. The quality of completions can also be assessed before deploying the model in a survey context, such that one can try different user inputs and evaluate the quality of text completions. In the Appendix, instructions for hosting a CSAS survey are described. Given that the application is programmed using Python's Django framework, a wide range of hosting providers are available for deploying the CSAS survey.

[8]Though the current applications use a real-time algorithm, one could also carry out Thompson sampling in batches. For example, submissions in earlier batches could be approved before being rated in subsequent batches and used to update assignment probabilities. This approach is still adaptive in that the probability of presenting items is not uniform. However, it comes at the cost of slower learning. Using this batch-wise method, the delay in processing and approving submissions could reduce the flexiblity of the approach to update item probabilities on the fly as contexts change.

7 Conclusion

This paper introduced the Crowdsourced Adaptive Survey (CSAS) method, an approach to developing evolving question banks in surveys. Leveraging the capabilities of large language models, open-ended text was converted into structured survey questions, while adaptive algorithms identified best-performing questions from the question bank. The method's efficacy was demonstrated through an application to Latino information environments, uncovering both factual and partisan narratives. Moreover, the method was applied to issue importance, highlighting the ability for the method to recover niche issue positions that would otherwise be missed by traditional survey methods.

In a study of negative beliefs about parties and candidates held by Latinos, I found that the most highly rated claims were either popular stereotypes of the parties, true claims, or widely reported allegations. In contrast, claims that were rated lower on accuracy were often objectively false and reflected more blatant misinformation. An additional analysis of Latino subgroups defined by WhatsApp usage revealed similar accuracy ratings across a variety of claims. The most consistent difference between the two groups emerged for true, but politically incongenial, claims, such that participants who trusted WhatsApp less were more likely to rate these claims as accurate relative to those who trusted WhatsApp for information. These descriptive patterns suggest that information gaps across levels of engagement with WhatsApp may be more pertinent with respect to more mainstream claims, rather than outright misinformation. However, research is needed to assess whether this holds more broadly.

Moving from the identification of contested claims within marginalized communities to the topic of issue importance, the CSAS method was used to develop a question bank comprised of issues provided by participants. While economic factors, healthcare, crime, and education were among the most highly-rated issues, more niche issues such as "candidate transparency" and "privacy protections" also received high ratings from the sample. A follow-up analysis comparing issue ratings across partisan subgroups revealed substantial differences in importance ratings, but also some areas of agreement on topics such as healthcare costs, public education funding, and inflation. A distinct advantage of the adaptive design is that the exploration of new items can yield a larger number of items than one would measure in a traditional static survey. By dynamically adjusting the question bank, CSAS can delve deeper into emerging topics and unexpected areas of agreement across partisan or ideological subgroups. Given heterogeneity in ratings across individuals, applying more sophisticated adaptive designs such as contextual adaptive experiments could improve estimates of politically contentious claims or issues (Offer-Westort et al., 2022).

Despite the advantages of the CSAS method, scholars should be aware of potential limitations or drawbacks of using LLMs. Simpler classification tasks, like those demonstrated in the work on important issue identification (Mellon et al., 2022), may not require the additional investment of fine-tuning pre-trained LLMs, as their existing capabilities may suffice. However,

capturing more abstract concepts such as misinformation may require additional steps such as fine-tuning and few-shot learning (i.e., including positive and negative examples in an LLM prompt). For more sensitive topics, it is strongly recommended to implement toxicity filters and carefully verify whether these filters are working as expected, as participants may potentially submit harmful content. Inverse-probability weighting is used to capture uncertainty in estimates, as in other work (Offer-Westort et al., 2021); however, this is an active area of research and future studies could the assess statistical properties of IPW in dynamic adaptive settings, and the requisite sample size for CSAS designs (Hadad et al., 2021).

Future research employing CSAS has the potential to explore a variety of topics where participant-generated content is particularly valuable. For instance, in candidate choice, voters not only consider issue positions or party affiliation, but also personal attributes such as honesty and competence. CSAS could be useful in uncovering additional elements that influence voting decisions. Furthermore, when identifying norms, core beliefs, or key sources of identity within hard-to-reach communities, CSAS could provide significant insights, reducing the misalignment of researcher-defined concepts with respondents' actual perceptions. CSAS could also be used to develop measurement scales that more accurately reflect the diverse considerations of participants. A great number of constructs in the social sciences are latent (e.g., democracy, gentrification, identity), and the CSAS method could serve as a useful tool for extracting lay definitions or perceptions of these various concepts. This kind of research, which prioritizes the perspectives and realities of study populations, not only bridges the gap between researchers and respondents but could also produce significant advances in our understanding of public opinion and political behavior, more broadly.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Agrawal, S. and Goyal, N. (2017). Near-optimal regret bounds for thompson sampling. *Journal of the ACM (JACM)*, 64(5):1–24.

Ansolabehere, S. and Schaffner, B. (2009). Guide to the 2008 cooperative congressional election survey. *Harvard University, draft of February*, 9.

Beeching, E., Fourrier, C., Habib, N., Han, S., Lambert, N., Rajani, N., Sanseviero, O., Tunstall, L., and Wolf, T. (2023). Open llm leaderboard.

Carrasco-Farré, C. (2022). The fingerprints of misinformation: how deceptive content differs from reliable sources in terms of cognitive effort and appeal to emotions. *Humanities and Social Sciences Communications*, 9(1):1–18.

Cortina, J. and Rottinghaus, B. (2022). Conspiratorial thinking in the latino community on the 2020 election. *Research & Politics*, 9(1):20531680221083535.

Dillman, D. A. and Christian, L. M. (2005). Survey mode as a source of instability in responses across surveys. *Field methods*, 17(1):30–52.

Dillman, D. A., Smyth, J. D., and Christian, L. M. (2014). *Internet, phone, mail, and mixed-mode surveys: The tailored design method*. John Wiley & Sons.

Dunaway, J., Branton, R. P., and Abrajano, M. A. (2010). Agenda setting, public opinion, and the issue of immigration reform. *Social Science Quarterly*, 91(2):359–378.

Egami, N., Fong, C. J., Grimmer, J., Roberts, M. E., and Stewart, B. M. (2018). How to make causal inferences using texts. *arXiv preprint arXiv:1802.02163*.

Elliott, K. J. (2020). Democracy's pin factory: Issue specialization, the division of cognitive labor, and epistemic performance. *American Journal of Political Science*, 64(2):385–397.

Hadad, V., Hirshberg, D. A., Zhan, R., Wager, S., and Athey, S. (2021). Confidence intervals for policy evaluation in adaptive experiments. *Proceedings of the national academy of sciences*, 118(15):e2014602118.

Inan, H., Upasani, K., Chi, J., Rungta, R., Iyer, K., Mao, Y., Tontchev, M., Hu, Q., Fuller, B., Testuggine, D., et al. (2023). Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*.

Jennings, W. and Wlezien, C. (2011). Distinguishing between most important problems and issues? *Public Opinion Quarterly*, 75(3):545–555.

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., de las Casas, D., Hanna, E. B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L. R., Saulnier, L., Lachaux, M.-A., Stock, P., Subramanian, S., Yang, S., Antoniak, S., Scao, T. L., Gervet, T., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. (2024). Mixtral of experts.

Key, V. O. (1961). Public opinion and the decay of democracy. *The Virginia Quarterly Review*, 37(4):481–494.

Krosnick, J. A. and Berent, M. K. (1993). Comparisons of party identification and policy preferences: The impact of survey question format. *American Journal of Political Science*, pages 941–964.

Lee, A. Y., Moore, R. C., and Hancock, J. T. (2023). Designing misinformation interventions for all: Perspectives from aapi, black, latino, and native american community leaders on misinformation educational efforts. *Harvard Kennedy School Misinformation Review*.

Leech, B. L. (2002). Interview methods in political science. *PS-WASHINGTON-*, 35(4):663–664.

Lopez, J. and Hillygus, D. S. (2018). Why so serious?: Survey trolls and misinformation. Available at SSRN: https://ssrn.com/abstract=3131087 or http://dx.doi.org/10.2139/ssrn.3131087.

McCombs, M. and Zhu, J.-H. (1995). Capacity, diversity, and volatility of the public agenda: Trends from 1954 to 1994. *Public Opinion Quarterly*, 59(4):495–525.

Mellon, J., Bailey, J., Scott, R., Breckwoldt, J., and Miori, M. (2022). Does gpt-3 know what the most important issue is? using large language models to code open-text social survey responses at scale. *Using Large Language Models to Code Open-Text Social Survey Responses At Scale (December 22, 2022)*.

Mena, P. (2019). Principles and boundaries of fact-checking: Journalists' perceptions. *Journalism practice*, 13(6):657–672.

Montgomery, J. M. and Cutler, J. (2013). Computerized adaptive testing for public opinion surveys. *Political Analysis*, 21(2):172–192.

Montgomery, J. M. and Rossiter, E. L. (2020). So many questions, so little time: Integrating adaptive inventories into public opinion research. *Journal of Survey Statistics and Methodology*, 8(4):667–690.

Montgomery, J. M. and Rossiter, E. L. (2022). *Adaptive Inventories: A Practical Guide for Applied Researchers*. Cambridge University Press.

Offer-Westort, M., Coppock, A., and Green, D. P. (2021). Adaptive experimental design: Prospects and applications in political science. *American Journal of Political Science*, 65(4):826–844.

Offer-Westort, M., Rosenzweig, L. R., and Athey, S. (2022). Battling the coronavirusinfodemic'among social media users in africa. *arXiv preprint arXiv:2212.13638*.

Ostrom Jr, C. W., Kraitzman, A. P., Newman, B., and Abramson, P. R. (2018). Polls and elections: Terror, war, and the economy in george w. bush's approval ratings: The importance of salience in presidential approval. *Presidential Studies Quarterly*, 48(2):318–341.

Page, B. I. and Shapiro, R. Y. (2010). *The rational public: Fifty years of trends in Americans' policy preferences*. University of Chicago Press.

Palmer, A., Smith, N. A., and Spirling, A. (2023). Using proprietary language models in academic research requires explicit justification. *Nature Computational Science*, pages 1–2.

Perens, B. et al. (1999). The open source definition. *Open sources: voices from the open source revolution*, 1:171–188.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training.

Rheault, L. and Cochrane, C. (2020). Word embeddings for the analysis of ideological placement in parliamentary corpora. *Political Analysis*, 28(1):112–133.

Ryan, T. J. and Ehlinger, J. A. (2023). Issue publics: How electoral constituencies hide in plain sight. *Elements in Political Psychology*.

Salganik, M. J. and Levy, K. E. (2015). Wiki surveys: Open and quantifiable social data collection. *PloS one*, 10(5):e0123483.

Sanderson, Z., Brown, M. A., Bonneau, R., Nagler, J., and Tucker, J. A. (2021). Twitter flagged donald trump's tweets with election misinformation: They continued to spread both on and off the platform. *Harvard Kennedy School Misinformation Review*.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Velez, Y. and Liu, P. (2023). Confronting core issues: A critical test of attitude polarization.

Velez, Y. R. (2023). Trade-offs in latino politics: Exploring the role of deeply-held issue positions using a dynamic tailored conjoint method. *Aletheia*.

Velez, Y. R., Porter, E., and Wood, T. J. (2023). Latino-targeted misinformation and the power of factual corrections. *The Journal of Politics*, 85(2):789–794.

Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. *science*, 359(6380):1146–1151.

Weller, S. C. (1998). Structured interviewing and questionnaire construction. *Handbook of methods in cultural anthropology*, pages 365–409.

Wlezien, C. (2005). On the salience of political issues: The problem with 'most important problem'. *Electoral studies*, 24(4):555–579.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena.

Appendix

A Algorithm Details

In Thompson Sampling, a Beta distribution is used as the conjugate prior and a Bernoulli distribution is used as the likelihood. This combination yields a Beta posterior distribution. The Beta distribution is ideal for binary outcomes, and could be used in CSAS settings where binary items are the most appropriate way of measuring the outcome (see (Offer-Westort et al., 2021) for an implementation). However, given the prevalence of rating scales in political science, I draw on Agrawal and Goyal (2017) to implement Thompson sampling with Gaussian priors (also referred to as Gaussian Thompson sampling). In contrast to traditional Thompson sampling, GTS assumes a normal conjugate prior and allows for the modeling of non-binary outcomes. The implementation of the algorithm is presented below:

Algorithm 1 Gaussian Thompson Sampling Algorithm

Initialization: Initialize "seed questions."
for $i = 1, 2, \ldots, I$ **do**
  if $i = 1$ **then**
    Assign initial seed questions with uniform probability.
  else
    if new questions are added by respondents and pass the filtering step **then**
      Add new questions to the item pool.
    end if
    a) Draw Samples:
    for each question q in the item pool **do**
      Draw $\theta_q \sim N\left(\hat{\mu}_{q,i-1}, \frac{1}{n_{q,i-1}+1}\right)$.
    end for
    b) Calculate Probabilities:
    $q_w = \frac{1}{M}\sum_{m=1}^{M} 1\left\{w = \arg\max_w \{\theta^{(1)}, \ldots, \theta^{(|Q|)}\}\right\}$.
    c) Apply Probability Floor and Rescale:
    $e_q = \max(q_w, .01)$ for each q
    Rescale $e_1, \ldots, e_Q$ to sum to one.
    d) Assign Items:
    $q_i \sim$ Multinomial $(e_1, \ldots, e_Q)$.
    e) Observe Response: Observe response and update parameters.
  end if
end for

In plain language, the first participant is randomly assigned to seed questions with a uniform probability and their ratings are used to update the initial parameters. $\hat{\mu}$ is equal to the mean rating for a given question, whereas $\hat{\sigma}$ is equal to $\frac{1}{n+1}$ where n is equal to the number of ratings for a given question. Since each participant rates their own submitted item and a set of adaptive items, parameters for existing and newly added items, assuming those items have passed the filtering step, are set for the subsequent participant. As participants enter the survey, dynamic items are presented with a probability determined by the ratings of the previous participants. This probability is calculated by simulating draws from the normal distribution for each question and then calculating the proportion of the time with which

each question emerges as the most preferred within a given simulaton. In essence, as each participant interacts with the survey, their responses influence an overall estimate of the question's popularity. This method of updating probabilities ensures that the questions most likely to elicit higher ratings are more frequently presented to future participants.

B Large Language Model Prompts

B.1 Misinformation Filter

- model: ada
- temperature: 0
- prompt:

> Is the following text a specific, verifiable claim or is it a value judgment/gibberish/evaluative statement? Use 1 to indicate a verifiable claim; 0 to indicate otherwise.
>
> Examples:
> Biden is corrupt.->0
> Bidenś son had profitable business deals in Ukraine.->1
> Trump was a horrible president.->0
> Trump kept classified documents in his home.->1
> George Soros funded the Black Lives Matter protests.->1
> Republicans are hiding a lot of information.->0
> The 2020 election was stolen by the Democrats.->1
> Republicans are racist.->0
> Republicans are in the pocket of big business.->0
> Democrats lie.->0
> Republicans are the best party.->0
> Democrats are the best party.->0
> Democrats hate the country.->0
> Republicans should ban Donald Trump from social media.->0
> Bill Clinton and Hillary Clinton were friends with Jeffrey Epstein.->1
> {text}->{completion}

B.2 Claim Summary

- model: davinci-002
- temperature: 0
- prompt:

> Produce a one-sentence summary of the following point of view. Only summarize the most important claim. This claim is about {party}.
> Claim: {text}
> Summary: {completion}

B.3 Important Issue Summarization

- model: gpt-4
- temperature: 0
- prompt:

> Please extract the political issue or concern mentioned by the respondent using

one to three words. Be descriptive and stay true to what the user has written.
Select only one issue, concern, or topic. Never ask about two issues.
If the issue is not politically relevant, return the phrase Room Temperature Superconductors.
If a related issue or theme has already been mentioned, return the same issue or theme as the output. Do not duplicate broad issue areas or themes.
Examples:
Previously Mentioned Issues () I care about the environment.->Environment
Previously Mentioned Issues (Taxation) My taxes are too high.->Taxation
Previously Mentioned Issues () Abortion should be legal under all circumstances.->Abortion
Previously Mentioned Issues (Immigration) Close the borders.->Immigration
Previously Mentioned Issues (Inflation) I am concerned about rising prices.->Inflation
Previously Mentioned Issue ({matches}) {text}->

Note: A comma-separated list of the five closest matches, labeled as matches, is generated using embeddings. These matches are identified from the question bank based on their cosine similarity.

C Open-Source and Proprietary Language Model Comparisons

OpenAI's large language models are used throughout the manuscript to filter and summarize open-ended responses before submitting them to the question bank. Though models such as OpenAI's GPT-3.5 and GPT-4 score at the upper end of various coding and natural language benchmarks (Achiam et al., 2023), open-source models such as Meta's Llama 2 (Touvron et al., 2023) and MistralAI's Mixtral (Jiang et al., 2024) have also demonstrated significant capabilities in these areas.[9] Moreover, unlike proprietary models such as GPT-4 that tend to update their weights with increasing training data or as a result of modifications to inference routines, base model weights are fixed and accessible, facilitating transparency and reproducibility (Palmer et al., 2023).[10]

To assess whether open-source models are up to the task of filtering and summarizing open-ended responses, LLM-generated output for Mistral's *Mixtral* model, a 46.7 billion parameter large language model with a unique Mixture of Expert (MOE) architecture was used. Mixtral has shown strong performance on various benchmarks, exceeding Meta's 70 billion parameter Llama model on Huggingface's Open LLM Leaderboard, which features a variety of coding, factual accuracy, and mathematics benchmarks (Beeching et al., 2023). Moreover, Mixtral has been found to compete favorably against GPT-3.5 on thousands of blind assessments of chat output (Zheng et al., 2023).

The performance of OpenAI's GPT models and Mistral AI's Mixtral were directly compared using the same open-ended responses and prompts from both studies.[11] Completions were classified as identical, comparable, or conflicting. Completions were marked as comparable when they returned outputs that were similar in meaning (e.g., categorizing an open-ended response about universal healthcare as healthcare versus universal healthcare), whereas conflicting outputs consisted of entirely different concepts (e.g., classifying a response about a cryptocurrency investor who was arrested as "corruption" versus cryptocurrency).

The models agreed the vast majority of the time. In the misinformation study, models produced identical or comparable output 86% of the time; whereas this number was 97% in the most important issue study. The higher share of comparable outputs in the most important issue (41% versus 11% in the misinformation study) is mostly a function of how the two models handled specificity versus generality when summarizing issue positions. For example, Mixtral had a tendency to classify any immigration-related response as "Immigration," whereas GPT-4 was more likely to distinguish between different kinds of immigration responses

[9] In this paper, I use the term "open-source" to refer to models that share their weights publicly. Mixtral is licensed under Apache 2.0, which is a common open-source license. However, Meta's Llama 2 has some limitations on commercial use, which contradicts the "Open Source Definition" that states that open-source software should not impose any restrictions on its users (Perens et al., 1999).

[10] Of course, stochastic outputs from large language models are largely unavoidable due to their probabilistic sampling of tokens. Even with zero temperature, a parameter that minimizes the randomness of outputs, stochastic outputs in large language models can also emerge due to technical factors such as floating-point errors and parallel core sequencing.

[11] As mentioned in the manuscript, the more capable model, GPT-4, was used to summarize issue positions in Study 2, whereas GPT-3 was used in Study 1.

(i.e., "Legal Immigration," "Border Security").[12] When classifying the output into better or worse categories, I coded cases where the summary accurately reflected the user response (e.g., "Class Exploitation" when a user specifically describes the exploitation of lower classes, versus "Inequality") as the "better" completion. In the misinformation study, the OpenAI model produced "better" output 57% of the time, whereas this number was 65% in the issue importance study.

C.1 Toxicity Detection across Model Types

Overall, the performance of Mistral AI's Mixtral was comparable to OpenAI's models, but OpenAI generally had an edge in terms of producing higher quality completions. It is possible that modifications to the prompt could yield even higher quality completions, given that Mixtral produced output that matched or resembled the output of the proprietary model. This suggests Mixtral could potentially be used as a "drop-in replacement" for GPT-3.5 or GPT-4 when implementing the CSAS method.

Still, toxicity filtering remains an area where open-source models like Mixtral may lag behind proprietary models like OpenAI's GPT-3 and GPT-4. Proprietary models often benefit from extensive tuning and "red-teaming" efforts where toxic content is deliberately produced to improve detection of harmful outputs, leading to more sophisticated toxicity filters. This is crucial, especially when dealing with sensitive topics such as misinformation where exposure to harmful content may be a risk. However, this is a quickly evolving area. Toxicity filters using the Llama model have been recently developed, and could serve as a potential open-source replacement. For example, Inan et al. (2023) use the 7-billion parameter version of Llama 2 to train a moderation model, and find that its classification accuracy rivals OpenAI's moderation endpoint. These moderation models are publicly accessible using various inference APIs and allow scholars to build fully open-source routines that manage completions and toxicity filters.

Despite the potential risks of harmful content, it is important to acknowledge that instances of toxic *output* are relatively rare in most applications. For instance, across the two studies, there were zero instances of LLM completions involving derogatory terms or slurs, and risk of harmful completions may be relatively low in settings such as gauging issue importance. In such contexts, open-source models may still be suitable even without sophisticated toxicity filters, given their strong performance in classifying and summarizing text. Nevertheless, when it comes to areas such as misinformation, the use of open-source models might carry higher risks. The lack of advanced toxicity filters could potentially allow harmful content to slip through, making these models less ideal. Testing and refining open-source moderation tools may be a necessary step before these alternatives are implemented. Appendix E describes how to deploy the CSAS method, and a GUI has been developed that allows scholars to carry out

[12] It is important to note that classifications need not be viewed as 'right' or 'wrong,' as this could depend on research objectives. If the goal is to extract general issue topics, Mixtral may be preferred due to its tendency to err on the side of broader themes, whereas if more niche issue positions are desired, GPT-3.5 or GPT-4 may be ideal.

their own “red teaming” efforts by assessing how models handle different types of adversarial inputs.

C.2 Testing Proprietary and Open-Source Models in Real-Time

A study with 200 participants was conducted on CloudResearch Connect to estimate costs across different API providers and assess the real-time performance of open-source models on January 6, 2024. 100 participants completed a CSAS survey that summarized their open-ended responses using GPT-4, wheres another 100 participants completed a CSAS survey using Mixtral. As mentioned in the manuscript, the inference cost per participant was approximately .005 cents for Mixtral using AnyScale’s ChatCompletion endpoint and .01 cents for GPT-4. Overall, GPT-4 and Mixtral produced similar completions. Though, Mixtral was slightly more prone to completion errors. Whereas all of GPT-4 summaries involved a single issue, there were four instances of completions using Mixtral that included multiple issues (e.g., “Abortion and vaccine choice”) and three instances of completions that did not return an issue, but instead continued a conversation (e.g., “In your examples, you have provided specific political issues…”).

In sum, while open-source models like Mixtral exhibit commendable performance in many areas, their application in contexts requiring stringent content moderation should be approached with caution. Until these models achieve parity with proprietary models in toxicity filtering, their use should be carefully considered, especially in settings where misinformation or sensitive content is a concern.

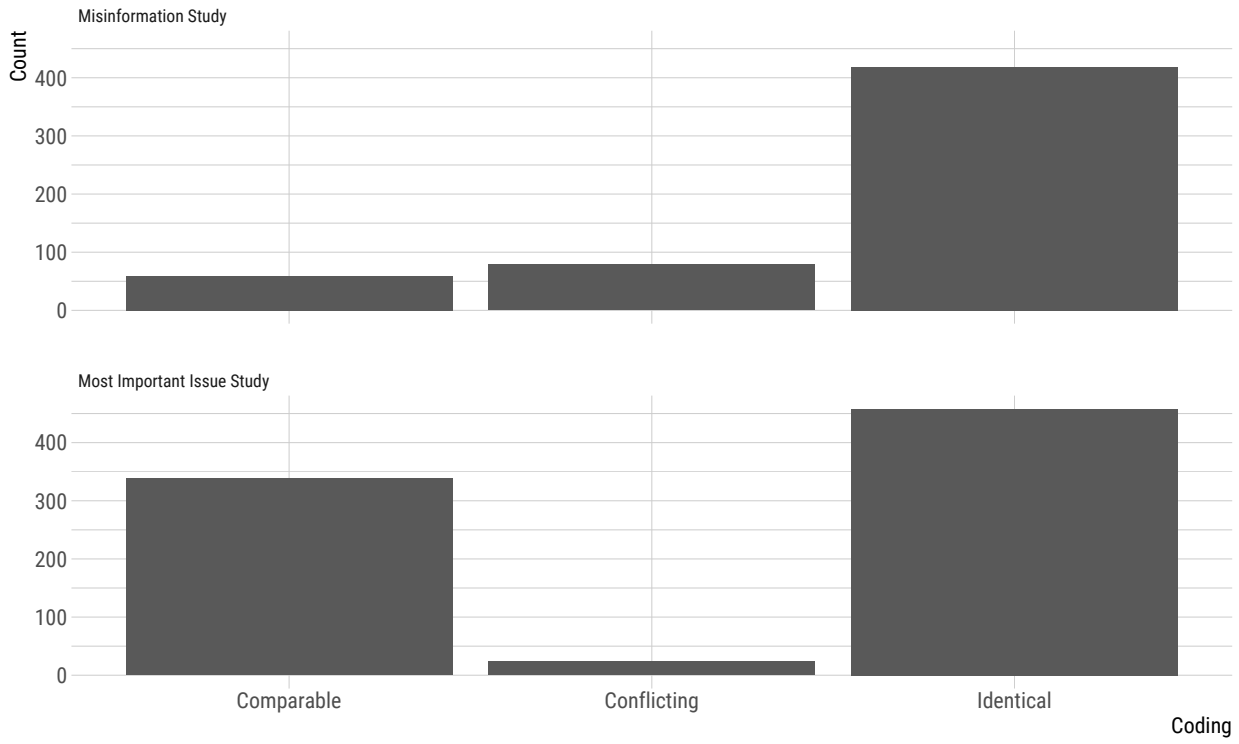

Figure C1: Comparison between Mixtral (Mistral 8x7B Instruct 0.1) and GPT-3.5/4

D Toxicity Filters

Toxicity filters prevent harmful content from reaching online participants, but they may also filter out claims that could reveal important aspects of public opinion or misinformation. Across the two studies, zero instances of toxic content involving slurs or hateful content in the open-ended responses or the completions were observed. However, some items that were flagged as toxic by OpenAI's moderation endpoint were not added to the question bank, even though they could have been relevant for studying misinformation. These items included topics such as child abuse, drug use, and violence (e.g., "Ted Cruz is the Zodiac killer").

Given that viral misinformation often includes these themes, and tracking the nature of false claims across groups and time is a vital goal of much of this research, toxicity filters that are overly sensitive to certain concepts could limit what we learn from the CSAS method. That being said, if enough submitted items are present that possess similar themes or claims, one could incorporate these items into future surveys, much like one would learn from open-ended responses in traditional surveys.

One could also potentially prompt LLMs to rephrase a piece of sensitive content until it can be included in a question bank. This iterative approach has been used in other research examining the impact of tailored counter-arguments on issue attitudes, where completions are "sanitized" before being presented to participants (Velez and Liu, 2023). Though the

possibility that toxicity filters may produce "false negatives" is concerning when studying topics such as misinformation, it may not be as significant a problem for other applications such as issue importance. In these applications, the primary focus is not to delve into the specifics of controversial or potentially harmful claims about politics, but to reveal policies or issue areas that are important to participants. As such, the likelihood of encountering content that would be flagged by toxicity filters is inherently lower in these contexts.

E Implementing the CSAS Method

Application of the CSAS method requires a database. The implementation in the manuscript uses an external SQL database hosted by a web provider, Web Hosting Hub, that stores the items, along with the number of ratings and average rating for each question. The latter two statistics are sufficient for implementing Thompson sampling, since the variance is calculated as the inverse of the total number of ratings. PHP is used to send information between the SQL database and Qualtrics. Given that one may want to implement CSAS without using a long-term hosting provider, it is worth exploring alternatives that can provide flexible and affordable services on a project-by-project basis. To that end, a Django implementation of the input, update, and sampling routines has been developed that can be self-hosted or hosted on popular cloud providers such as Replit.

Below, I walk through the steps of setting up a CSAS survey using the cloud hosting provider, Replit, as well as a self-hosted version. Though the latter is cost effective and does not require an external database, it may not be feasible to host locally due to computational resources. Before launching the server, pilot tests are recommended to simulate the expected traffic. In Qualtrics, the "Generate Test Responses" function can be used to run these simulations. Running CSAS locally also carries risks, since it makes local servers publicly accessible. Thus, external hosting may be preferable from a security vantage point.

The CSAS application currently includes the option to use OpenAI's GPT-3.5 turbo and Mistral AI's Mixtral models. The former uses OpenAI's API to generate embeddings and flag content as toxic as part of the "redundancy" and "toxicity" filters described in the paper, whereas the latter uses Anyscale, an LLM hosting provider, to carry out the filtering process using open-source alternatives. These open-source models are Llama Guard for moderation, and General Text Embeddings (gte-large) for redundancy. Therefore, one can choose to either implement a fully proprietary or open-source model "stack."

E.1 "Easy CSAS" Replit Example

- Sign up for a free Replit account.
- Visit https://replit.com/@yrvelez/CSAS-Helper
- Fork the repository by clicking "Fork" on the top right.
- Update secrets

  SECRET_KEY: This can be anything, but preferably a strong and random string that is hard to guess.

  OPENAI_API_KEY: To use OpenAI's models, an API key is needed. This can be obtained by following this guide.

  ANYSCALE_API_KEY: AnyScale is a provider that hosts open-source models such as Llama 2 and Mixtral. One can sign up for an account at this page.
- After updating secrets, clicking Run will enable the Django web server and open the "CSAS Helper."
- Save the relevant parameters (e.g., minimum scale value, maximum scale value, API) and

add seed items. **Note: The number of seed items must be equal to or greater than the number of dynamic items. Otherwise, there will not be enough items to sample from in the database, and an error will be produced.**

- Copy and paste the HTML from the "Easy CSAS" section into the HTML view of a Text/Graphic Question Type in Qualtrics.
- Launch the survey.

E.2 More Customizable Replit Example

- Sign up for a free Replit account.
- Visit https://replit.com/@yrvelez/CSAS-Helper
- Fork the repository by clicking "Fork" on the top right.
- Update secrets

  SECRET_KEY: This can be anything, but preferably a strong and random string that is hard to guess.

  OPENAI_API_KEY: To use OpenAI's models, an API key is needed. This can be obtained by following this guide.

  ANYSCALE_API_KEY: AnyScale is a provider that hosts open-source models such as Llama 2 and Mixtral. One can sign up for an account at this page.
- After updating secrets, clicking Run will enable the Django web server and open the "CSAS Helper."
- The "CSAS Helper" allows users to test how input is processed by different LLMs, revise prompts, set default APIs, and the number of dynamic items. It also provides a set of Qualtrics links that can be used to implement the input, update, and sampling routines of the CSAS method.
- In Qualtrics, the sampling step can be implemented by creating a Web Service that points to the sampling URL. Clicking "Test" will produce the relevant fields. To quickly save these fields, one should click Select All and Add Embedded Data.
- Before the input block, an open-ended question should query the user about a topic. Another Web Service should be created after this question that points to the input URL. The query paramer should be input, and it should be set equal to the internal question ID.
- Ratings should be provided to participants by piping in embedded data fields (e.g., ${e://Field/q_1}). The user-submitted question is saved in Qualtrics as ${e://Field/completion}.
- To implement the update routine, the update URL should be used with a list of query parameters (${e://Field/q_1}) and ratings using Qualtrics' Recode format (e.g., {q://QID3/SelectedAnswerRecode/2}).

  A Qualtrics template for a "three dynamic item" survey is provided here to facilitate setup.

E.3 More Customizable Local Hosting Example

- Install GitHub (https://github.com/git-guides/install-git), Python 3.0 (https://www.python.org/), and Ngrok (https://ngrok.com/download)

- Open Terminal/Shell
- Clone the GitHub repo (https://github.com/yrvelez/csas.git)
- Create a virtual environment: virtualenv env
- Active the virtual environment: venv source/bin/activate
- Install requirements.txt (pip install -r requirements.txt)
- Change directory into csas_files
- Update API keys by using the export command in Terminal (e.g., export SECRET_KEY=foobar).

  SECRET_KEY: This can be anything, but preferably a strong and random string that is hard to guess.

  OPENAI_API_KEY: To use OpenAI's models, an API key is needed. This can be obtained by following this guide.

  ANYSCALE_API_KEY: AnyScale is a provider that hosts open-source models such as Llama 2 and Mixtral. One can sign up for an account at this page.
- Check for changes to models (databases) and create migration files: python manage.py makemigrations
- Apply changes to database: python manage.py migrate
- Launch Django app: python manage.py runserver localhost:8080
- Launch ngrok: ngrok http 8080
- Click "Visit Site" after ngrok disclaimer
- The "CSAS Helper" allows users to test how input is processed by different LLMs, revise prompts, set default APIs, and the number of dynamic items. It also provides a set of Qualtrics links that can be used to implement the input, update, and sampling routines of the CSAS method.
- In Qualtrics, the sampling step can be implemented by creating a Web Service that points to the sampling URL. Clicking "Test" will produce the relevant fields. To quickly save these fields, one should click Select All and Add Embedded Data.
- Before the input block, an open-ended question should query the user about a topic. Another Web Service should be created after this question that points to the input URL. The query paramer should be input, and it should be set equal to the internal question ID.
- Ratings should be provided to participants by piping in embedded data fields (e.g., ${e://Field/q_1}). The user-submitted question is saved in Qualtrics as ${e://Field/completion}.
- To implement the update routine, the update URL should be used with a list of query parameters (${e://Field/q_1}) and ratings using Qualtrics' Recode format (e.g., {q://QID3/SelectedAnswerRecode/2}).

A Qualtrics template for a "three dynamic item" survey is provided here to facilitate setup.