Opportunities and risks of LLMs in survey research

David Rothschild, James Brand, Hope Schroeder, Jenny Wang

October 28, 2024

Abstract

Recent advances in the development of large language models (LLMs) bring both disruptive opportunities and underlying risks to survey research. LLMs' capabilities for content generation and summarization tasks have already led to fast-paced innovation across social science research communities, including survey and market research, both academically and in practice. In this research note, we outline opportunities for LLMs to assist in survey creation, testing, analysis, and reporting. Backed by both practical examples and academic literature, we identify areas for research and development, distinguishing between challenges related to survey methods and the tools used to deploy surveys—a distinction necessary for the field to benefit from potential opportunities while minimizing potential risks. Further, we emphasize how different advances affect the degree of agency for the researcher. Overall, we are cautiously optimistic that LLM-based tools will augment, as opposed to replace, the researcher in the long-run, and will allow the survey research industry to scale.

Introduction

Surveys are a proven method of collecting real-world information from target populations to better understand their sentiment, knowledge, and actions. These results are critical for facilitating informed decision-making. Surveys are versatile instruments, used extensively in research across industry, government, and academia. In market research, they can play a crucial role in determining whether to invest in a product, or in identifying target demographics in an advertising campaign. In political research, surveys monitor trends in support for political candidates, elected officials, and policies. Governments also rely heavily on survey research; for example, they conduct employment surveys that provide data to help businesses make informed macroeconomic decisions. In academia, surveys are particularly critical for testing hypotheses in the social sciences.

The survey research community experienced its most recent major transformation with the gradual acceptance of internet-based respondents in the early 2000s through 2010s. Online survey instruments dramatically altered the per-respondent cost of surveys, reducing time and financial costs [1]. However, while this technological shift affected the tools that researchers use to collect data, executing a successful survey has remained a relatively stable multi-step process. We stylize this process as follows:

0. **Measure Ideation**: Researchers determine the measures they want to track.
1. **Survey Ideation**: Researchers turn measures into tractable survey questions.
2. **Audience Response**: Researchers launch a survey instrument to the target audience.
3. **Data Analytics**: Researchers clean the data and analyze the results.
4. **Reporting to Stakeholders**: Researchers convey the results back to relevant stakeholders.

While the advent of online survey platforms impacted the distribution and costs of launching surveys to various audiences (Step 2), it did not significantly affect other main stages of the survey research process, including measure ideation (Step 0), survey ideation (Step 1), data analytics (Step 3), and reporting to stakeholders (Step 4).[1] We note measure ideation as Step 0, as some researches

[1] Of course, there were some changes: online infrastructure changed the nature of some questions from other common modes. Internet-based surveys ushered in a new wave of research focused on developing more sophisticated analytical methods to offset the challenges of lower quality samples [2].

may think of this as outside of the survey pipeline and we will not go into this step in depth in this note, but it is necessary for a meaningful survey and can take a lot of time and effort. In short, the last major disruption to survey research spurred growth and cut some costs, but the process remains costly in time and money due to the labor-intensive work still necessary to ensure high quality results in an increasingly complex and evolving process prone to a variety of bias and errors.

The emergence of large language models (LLMs) has prompted an opportunity to rethink the survey creation process. The popular release of ChatGPT spurred academic researchers and companies alike to quickly adopt these tools into their survey frameworks [3]. An initial major focus in both the academic research community and the survey industry was to assess the ability of LLMs to generate synthetic responses that mimic human respondents [3]. However, while synthetic participant data can be effective under some conditions (e.g., synthetic estimates were comparable to average human responses on willingness-to-pay for well-known product types [4]), its adoption in the market is limited so far due to a mix of consistency and representational concerns, along with practical business questions of risk and marginal value. Given these concerns, some teams have pivoted to using LLMs in later stages of the survey pipeline (e.g., start-ups like RoundTable.ai shifted their product focus from synthetic responses to AI-assisted data cleaning for surveys). Others have begun to explore research into survey question writing and analytics for unstructured responses [5]. Among all LLM-related papers in the 2024 AAPOR conference (in May 2024) relating to the survey research process, three dealt with writing questions, six dealt with coding open-ended answers, and just one focused on synthetic responses.[2]

LLMs have features that are potentially beneficial to survey research, but they also have features that can complicate their use for rigorous survey research. LLMs are models that have been pre-trained on massive amounts of data. The contents of proprietary or closed models' training data are largely unknown, and therefore it is unknown what survey data (individual or aggregated), literature, and research have been ingested in the training process. After training, many commercial LLMs are then subject to processes where the base model is modified to improve it along some dimension, for example, to increase knowledge of a certain domain, or to improve a user's experience. This can include steps like instruction and safety-tuning, wherein the model learns to follow instructions, or be "safer" by avoiding certain sensitive topics or stereotypes. That model may then be served to users through a public-facing web interface (like ChatGPT or CoPilot), through an API (a protocol that lets software communicate directly with the model), or in the case of open models, through a published model that allows someone to host the model themselves. The training data, training process, and post-training modifications of a model are often opaque in closed models but can be more transparent in open models. Fine-tuning can make LLMs much better for some purposes, like improving their conversational capabilities through reinforcement learning from human feedback (RLHF). These fine-tuning processes can have unintended side-effects for some use cases, including in survey applications. For example, if the model is fine-tuned to avoid certain sensitive topics for safety reasons when interacting with human users, the model may be less likely to generate a sound question on a sensitive topic, or less likely to provide an "anti-social" response to a survey question that could significantly skew a researcher's findings.

In this note, we use several key categorizations to organize the disruptions we anticipate LLMs bringing to the survey research industry. First, we consider their position within the survey pipeline. Second, we consider whether the impacts we envision are likely to be short-term or long-term in nature. A short-term disruption is one which is already relatively well-understood and generally ready to deploy. A longer-term disruption, by definition, has at least one major hurdle preventing or slowing implementation at present. Some of these hurdles are methodological questions about how LLMs can function in the survey process, and other hurdles reflect the practical integration of LLMs into the tools survey creators use to design and deploy surveys. We make a distinction between these two sets of concerns, as they introduce two kinds of research questions. *Methodological* concerns include changes to the types of survey instruments we construct and their validity, as well as questions about how the models themselves may influence our survey methods. This includes how the LLM's training data or additional fine-tuning affects its responses (e.g., to follow instructions or produce chat responses).

We separate these methods questions from concerns about how we can best design effective *tools* and platforms for survey creators with assistance from LLM-based features. The design and implementation of tools that survey creators use will, in turn, significantly influence how survey research is actually carried out at scale. Survey creators may have different needs and goals, and as such, LLM-based survey tools may support these different audiences differently. For instance, survey ex-

[2]Program for 2024 AAPOR Conference can be found here: https://aapor.confex.com/aapor/2024/meetingapp.cgi.

perts have different support needs than novice survey creators, and intentional tool design can help support them differently. Those using LLM-based survey tools may have different priorities across industries and contexts that inform their decision-making about the relative costs and benefits of using these new tools. For example, government surveys might prioritize thoroughness, with long, well-established questionnaires, while surveys in industry may focus on speed and efficiency due to competitive pressures.

A third key categorization involves the level of agency given to the survey creator, which is significantly influenced by how researchers design interactions with LLMs. This design choice carries important implications for both the survey creator's control and potential downstream implications for the outcome analysis. For example, using an LLM to get feedback on a question written by a human places the researcher squarely in a position of agency over the work. By contrast, prompting an LLM to generate a survey question gives more agency to the LLM's suggestions in a survey. Finally, when an LLM is used to conduct a survey end-to-end, the tool is granted the most agency. While it may be instructed to follow specific goals, any engagement with respondents without direct human oversight introduces new risks that need to be mapped and ameliorated (or otherwise weighed) along with the potential benefits.

Considering these trends in both product design and academic research on these topics, along with our own experiences using LLMs in surveys, we argue that an under-explored use of LLMs is for augmenting human experts (relative to replacing human respondents) especially in Step 1 (creating and testing surveys prior to deployment) and in Step 3 (data cleaning and analytics). We outline potential opportunities for integrating LLMs into survey methods and tools, review literature on the risks of each potential application to the survey creation process, and suggest recommended areas for future research. At each stage, we synthesize knowns and unknowns into outstanding questions for the survey research community in order to engage a robust dialogue around LLMs in survey research. The next three sections progress from lower to higher risk, with decreased agency for humans but increasing potential opportunities in each successive section: editing questions with LLMs, generating questions with LLMs, and interactive surveys run by LLM-based agents. After that, we explore the use of synthetic responses, not to replace human responses (i.e., not to jump into Step 2), but to enhance pilot testing, focus targeting, and prepare data analytics (i.e., as a supplement to Steps 1 and 3).

Editing survey questions

Question formulation is an ongoing, well-known problem in survey research [6]. For survey creators who already have a draft of their questions, LLMs may be able to provide feedback that strengthens the survey instrument according to known best practices in survey design and a deep knowledge of common questions.

The way a survey question is worded can have major effects on the response quality [1]. Survey creators may use vague wording or fail to consider how the question may be misinterpreted, and inexperienced survey creators may be at particular risk of this kind of survey design misstep. Tourangeau, Rips, and Rasinski (2000) outline seven main kinds of comprehension issues that plague survey designs, including grammatical ambiguity, excessive complexity, faulty presupposition, vague concepts, vague quantifiers, terms unfamiliar to a reader, and false inferences [7]. Different question formats, like open-ended or closed questions, also each have their own known risks. Groves draws on Sudman and Bradburn (1982) to further provide recommendations for formatting questions about sensitive behavior and attitudes, including advice regarding avoiding "loading" questions for a particular response, avoiding vagueness, and reducing cognitive burden [8]. Lenzner et. al shows that vagueness and word choice significantly affect survey response success [9].

LLMs have been shown to give feedback on student writing that is comparable by multiple measures to feedback from experienced human editors, without any fine-tuning for the task [10, 11]. Researchers have also found LLM-generated feedback to be useful when editing scientific papers [12]. This preliminary evidence from research on essay and research paper writing tasks suggests that LLMs could provide useful editing feedback for writing in other contexts as well, but the context of survey question formulation requires additional investigation.

Anecdotally, interactions with LLMs in chat-based interfaces suggest they demonstrate the ability to give helpful feedback in the survey creation process, even absent any survey-specific fine-tuning. LLMs have become ubiquitous through chat-based interfaces that can provide responsive feedback to a given

input for low cost. Early this year, our team was interested in studying the Israel-Gaza War's impact on the 2024 presidential election. We asked Microsoft Copilot for feedback on our initial question, "Do you approve of Joe Biden's handling of the Israeli-Palestinian conflict?". The tool responded with suggestions to improve the survey question, recommending that we ask a pre-question to gauge respondents' familiarity with Biden's policy on the Israeli-Palestinian conflict, specifying a timeframe and particular actions within the question, and adding a scale of approval. These suggestions mirror recommendations from Schaeffer and Dykema (2011) for question formulation that avoids vagueness [13].

Interactions with a chat-based LLM also provided us with useful feedback on the wording of survey response choices, another critical part of creating a sound survey instrument. In another project where we measured the frequency of political discussions among respondents, we noticed a discrepancy between our results and existing trends in news consumption. Specifically, over 40% of respondents indicated that they spoke about politics "Once or Twice" and nearly 25% selected "Daily" to the question, *How often did you talk politics with your friends and family in the last week?* [14]. This raised concerns about the potential influence of social desirability bias, suggesting that our answer choices—*Multiple times a day, Daily, Once or twice, and Never*—were causing respondents to over-report political news consumption behaviors they perceived to be more socially acceptable [15]. We decided to revise our answer choices to mitigate this form of bias and better capture the expected variation in responses. We prompted GPT-4 for feedback, asking:

> **Prompt**: How can we improve the answer choices for the survey question "How often did you talk politics with your friends and family in the last week?" The current options are "Multiple times a day," "Daily," "Once or twice," and "Never." We have noticed that more respondents choose "Daily" and "Once or twice" than expected.

GPT-4 suggested that we rewrite the answer choices to include an option that implies engaging in political discussions with an *irregular* frequency. Our revised answer set included the following choices: *Multiple times a day, Once a day, Several times a week, A few times, Irregularly / Occasionally,* and *Not at all.* In a subsequent pilot test using the improved answer set, we noticed the distribution of responses was more closely aligned with existing data on news consumption trends. Notably, inclusion of the "Irregularly/Occasionally" option allowed us to capture a more accurate measure of political discussion frequency, as respondents were more willing to admit occasional engagement in political discussions than to claim they did not engage at all.

These anecdotes show the potential for LLMs to provide interactive feedback in the survey question creation in the form of a chat-based interaction, while maintaining agency for the researcher. This could be especially useful to inexperienced survey creators learning classic theories of survey research and pitfalls in the survey design process [16], but also for experienced survey creators who can benefit from editing feedback on their work. While it is not possible to judge the quality of a question in absolute terms, the question noted above performed consistently, correlating with other variables of interest in a long-running repeated cross-sectional study. Our simple tests found that the LLM's feedback may align with known question-wording recommendations from the survey creation literature, which may be useful in providing helpful assistance on question wording.

Another under-explored opportunity is the ability of LLMs to effectively paraphrase or rephrase text while preserving its semantic meaning [17]. Given the sensitivity of survey questions to formulation and wording, LLMs could be used to generate many wordings of the same question for subsequent testing. Further, this chat-based editing does not need to be done question-by-question, but tools can engage on the full survey, potentially exploring issues that may be overlooked (for example in the interaction of questions), even by experienced researchers.

Yet, there remain unknowns and risks. LLMs fundamentally rely on the data on which they are trained, which may (or may not) contain literature on survey question creation, answer formulation, and survey design. Additionally, off-the-shelf LLMs produce probabilistic responses, which means feedback on survey creation or question rewording from an off-the-shelf LLM can be inconsistent. Relatedly, responses may be preferentially generated according to one theory in the survey creation literature more often than others, depending on what is present in the training data. More niche insights on survey creation are less likely to be surfaced than common ones. If an LLM repeatedly gives the same editing advice, survey response quality could improve or we might observe homogenization, which has been seen in other brainstorming tasks that involve AI [18, 19]. Homogenization may sound

benign, assuming there is a clear "best" way to ask a question, but often there is no best question to ask, and homogenization may reduce valuable exploration. On the other hand, LLMs could also be used to augment the process of exploring and converging on optimal questions.

Future Research Questions

- Methods: What knowledge do proprietary and open-source LLMs demonstrate about survey design off-the-shelf (e.g., what information does it have on examples of measurement and specification biases), and can their knowledge be augmented for the purposes of designing high quality survey questions and answer choices?
- Tools: Can LLMs be used to provide interactive feedback on survey design, through chat-based interfaces or existing platforms? How might survey creation tools assist in teaching novices to create well-constructed surveys according to known best practices through providing feedback, suggestions, and editing? How might survey creation tools be designed to consider interesting or productive alternatives in survey design, rather than homogenizing towards known paradigms and defaults alone?

Question generation

The previous section focused on the case where a human has already written a draft of a survey question and seeks review from a LLM in a chat-based interface. Using LLMs for question generation itself, not just using them for feedback on human-written questions, is an obvious use case for LLMs in survey creation step of the pipeline. LLMs have already shown capability in writing multiple choice questions for education contexts [20] but their use in generating survey questions has not yet been systematically characterized. Furthermore, there are ways of operationalizing LLM-based survey question creation that differentially align with survey goals, grant agency and support to the survey creator, and have the potential to affect the study methodologically.

Writing a good survey question that addresses a research question and also avoids common pitfalls is time-consuming; it is easy to spend months on expert review, focus groups, and cognitive testing of surveys. In time-sensitive survey deployment contexts, or in contexts where many unique survey questions need to be generated at a time (i.e., for a large set of products or content), survey methods have largely been difficult to scale. Question templates can be leveraged to keep researchers' goals at the forefront of the work while increasing efficiency, but performance for both question and answer choice generation can and should be extensively vetted before deployed use.

An early experiment we deployed sought to understand how the same piece of breaking news is differentially absorbed by consumers belonging to different US political parties. We used an LLM to generate the question and answer choices for a unique single poll question for each article, which we launched using Civic Science's poll publishing system. Crucially, we generated a question of a specific templated format *Which of the author's points in this article about [TOPIC] are you likely to remember in a month?*. We used GPT-4's proven news summarization capabilities [21] to generate four takeaway points from the article. We extensively vetted the LLM's ability to generate questions based on this specified format with those four takeaways as answer options. We ran a human subjects evaluation with hundreds of participants, which showed that participants judged GPT-generated takeaway choices to be of similar faithfulness and factuality as standard summary points of the article, justifying our use of this method for this purpose. The question and answer choices were human-reviewed before launching the question to fully ensure quality. Despite the fact we maintained human review, using templated generation reduced the time to survey launch substantially, allowing the study to capture survey responses from as many online respondents as possible once the article was viral and before interest in the article quickly receded.[3] Figure 1 provides an example.

In this example, we used vetted, templated questions to scale customized surveys, but others may be tempted to prompt a LLM in a more open-ended way in order to generate questions without guidance on structure, perhaps for brainstorming purposes. Some of this is already happening: major

[3] https://civicscience.com/about-civicscience-polls/. In a more general test of LLM-created polls with human editorial review, for their daily polling Civic Science workers write an average of 36.2 questions per hour with LLM-augmented question writing, a marked increase from the average of 11.5 questions per hour prior to its implementation, with no loss of quality.

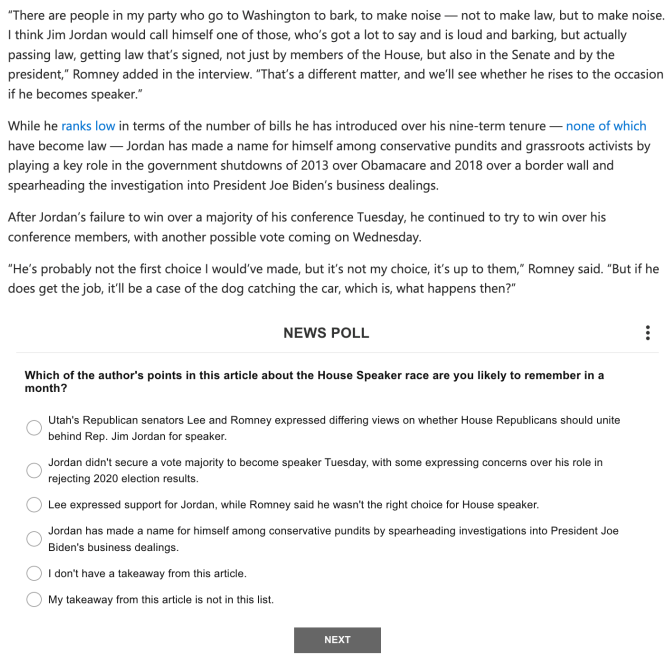

Figure 1: Screenshot of a poll question at bottom MSN article using Civic Science's polling module (note that the text is just the bottom of the article). All questions in the study were generated with LLM using preset structure, then reviewed and edited (if necessary) by a human before launching.

companies in the survey industry including Qualtrics and Pollfish have launched LLM-based question generation features directly into their platforms. For instance, after a user types in their survey goal in Pollfish, the platform uses LLMs to generate questions based on what the user stated they want to learn. Then, the user can edit the auto-generated questions, lowering the barrier to starting a draft.[4] However, literature from the classic psychological and behavioral research as well as human-AI interaction research that humans tend to anchor on suggestions [22]. The anchoring effect tends to be strong for the inexperienced, and for suggestions that are close to the expected range (rather than outliers) [23]. Because LLMs in 2024 generate highly fluent responses to most kinds of queries, even if the responses are in some ways different than what a human might have said, the chance that they provide suggestions on which humans then anchor is high, and the impact of this should be investigated [24]. Additionally, there is extensive evidence that LLMs are highly sensitive to prompting [25], another aspect of LLM survey generation that needs to be explored.

As tooling is developed, whether it is to scale individual questions or create entire surveys, we hope that developers consider simple features that can increase agency and augmentation of researchers, rather than trying to replace them, adding unnecessary downside risk [26]. Some research on tools providing LLM-based support in data analysis has found that users may prefer to interface with the problem themselves before receiving assistance [27]. Developers of analysis tools for researchers consistently note researchers' desire to retain agency over the research process [28]. Some work building data analysis platforms with LLM support builds opt-in suggestion features with this principle in mind in order to promote researcher agency [29].

Future Research Questions

- Methods: What kinds of questions do LLMs generate, especially when survey goals and task templates are underspecified? What kinds of tendencies do LLMs have, influenced by their training data, when generating questions and answers that could affect surveys? How does question generation performance vary by topic domain or question type? How do we consider the impact of anchoring, especially with the potential variation that LLM-based tools may introduce.
- Tools: Can survey tools be designed to generate questions according to researchers' goals or

[4] https://www.pollfish.com/

according to a template? Can survey tools be designed to support a brainstorming stage of question generation while minimizing risks of negative or undesired influence on survey design?

Interactive surveys with LLM-based agents

The previous two sections focus on using LLMs in common fixed survey paradigms. In this section, we explore the possibility that LLMs can create more engaging survey experiences that involve deeper interactions through open-ended questions and conversations. Fully interactive survey interfaces are already being deployed by several start-ups, along with many more companies that focus on reading and categorizing open answer responses. Unlike traditional surveys, which rely on predefined survey flows and static question formats, LLMs can adapt dynamically to respondents' answers by generating follow-up questions that probe deeper into potentially informative feedback.

Before considering more complex interactivity, LLMs can help overcome some key challenges of open-ended survey questions: coding and quantifying response data. Labelling responses to open-ended questions with humans is costly, and consistency across human labelers can vary wildly. Transforming those open-ended survey responses into useful indicators can be difficult, for humans or models, a challenge stemming from the very heterogeneity that makes them rich sources of insight. LLMs are rapidly transforming text analysis methods relevant to this problem. Methods of generating topics and themes from a corpus with the assistance of a LLM are becoming ubiquitous, and provide an opportunity for survey researchers to make sense of open-ended survey data. TopicGPT, GoalEx, TNT-LLM, and LLooM all propose methods of using LLMs to generate media and research question-specific topics, taxonomies, and thematic insight [30, 31, 32, 33]. They provide additional flexibility compared to traditional topic modeling methods like latent dirichlet allocation (LDA) [34], but unlike LDA, their biases and risks to methodological validity have been less well-studied. In short, it is critical to consider the robustness, replication, and effectiveness of the LLM-based analytics with the right counter-factual: other modeling techniques, humans, or nothing. LLMs compare favorable with humans due to the cost-effective scaling and favorable to other models with their robustness and flexibility. Just like teams of humans, the models may give different answers to the same data, especially over time. But, in many cases they are already being use when the true counter-factual is something else: nothing, as the cost to reading categorizing the open-ended would be just too high for many research projects. Finally, human involvement is still critical when using LLMs to maintain ensure the task in on target, especially as various methodological questions are still unanswered

Retrieval-augmented generation (RAG) systems also facilitate new ways of understanding, retrieving, and analyzing data: disrupting not just the analysis phase of the survey pipeline (step 3), but reporting on results as well (step 4). They combine LLMs with data retrieval systems to make some types of survey data analysis easier, by making data retrieval requests in natural language possible [35, 36]. For example, Civic Science has developed a proprietary chat-bot which can answer questions about the data in natural language, tying in current polling data to both earlier polling data and open panel questions for quick analysis. Further, the novel interfaces between researchers and their data also serves to disrupt the normal, static, delivery of data to end stakeholders.

When describing "interactiveness" in current literature, survey researchers often refer to surveys designed to facilitate dynamic engagement through branching logic or real-time feedback. Branching or conditional logic is often used to display only relevant questions or customize unique question paths for different respondents. For example, in the American National Election Studies (ANES), survey branching is used to measure more specific responses to political party identification on a rating scale [37]. This approach is based on the idea that decomposing a problem into smaller components leads to increased response accuracy and can improve the validity and reliability of responses to rating scale questions [38, 37]. In addition to branching logic, existing literature often discusses "interactive" surveys as those that develop visual representations of questions to make the content more engaging and increase the data quality [39, 40]. Or, they discuss designs that enable the survey creator to add attention checks or prompts that encourage respondents to slow down and provide more thoughtful responses [41, 42]. In these survey designs, the survey creator retains full control over the interaction, deciding how and when to implement these tools.

Over the past few years, the human-computer interaction (HCI) community has explored the potential of chatbots to conduct surveys. These tools have been shown to significantly improve participant engagement and elicit higher-quality responses [43]. Early HCI research on chatbots compared

web-based and chat-based interfaces using identical questions, focusing on how the survey instrument affected response quality [44]. For instance, AI chatbots can mimic active listening skills and encourage respondents to provide more informative open-ended responses [45]. These interactive capabilities allow chat-based surveys to gather richer information compared to traditional web surveys with standardized question sets.

While AI chatbots are one approach, LLMs have opened up new possibilities beyond chat interfaces. Interactive LLM-based questions can address open-ended questions beyond labeling and coding the responses themselves. Respondents tend to give short answers to open-ended questions, which vary in quality. LLM-based systems could detect low quality answers, and ask appropriate follow-ups if necessary. They can be conducted in a mix of modes, such as chat or voice, depending on what is most comfortable for respondents or what mode-specific biases the researcher wants to address. Already, companies like Voiceform[5] and Outset[6] allow researchers to conduct hundreds of surveys in parallel using AI to probe respondents to dig deeper on specific topics (in chat or voice), then use LLMs to synthesize qualitative results.

While the potential benefits are clear, they come with corresponding risk and many outstanding research questions. Beyond the concerns noted above regarding the labelling and quantifying of open-ended response data, from an agency perspective, the transition from creating static surveys to using LLMs to dynamically-generate surveys impacts the level of control that survey creators have over the question flow. This shift raises questions about oversight of the tool and the ethical implications of using AI to guide respondents' answers. Additionally, there exist many unknowns related to how LLM conversations could play out. Potential risks to participants include the possibility of LLMs exploiting respondents, tiring them out, or asking unnerving questions that break rapport. For survey creators, there are risks that the LLM might act inappropriately, perform out of alignment with the survey's goals, or produce less interpretable responses. This creates liability for the survey creator and may undermine the survey's validity. To lessen this risk, surveys creators should consider mixing LLM-based questions with standard online survey frameworks. This may minimize both human respondent labor and potential biases, to capture as much information as possible as efficiently as possible. However, this is still an area that requires deeper exploration. We must be careful to maintain a balance between flexibility of interactive modality and the ability to obtain actionable responses.

Future Research Questions

- Methods: How might research rigorously leverage LLMs in text analysis to produce insight from open-ended survey questions? How might underlying training data in LLMs affect their performance in the generation of dynamic surveys, including variation in their performance due to their uneven distribution of knowledge on certain topics, cultures, and customs? How might the design specification of interactive surveys align with a researcher's goals while striking a balance between LLMs' flexibility and standard questionnaires' rigidity? How might participants be protected from potential risks of LLM-based interactive surveys, including exploitation, regrettable disclosure, and fatigue?
- Tools: How might interactive survey tools be designed to increase inclusion and response from groups of interest, including underrepresented groups, according to survey goals? What guardrails are built into the tools to minimize risk of dangerous, offensive, or otherwise unproductive questions?

Pilot testing

LLM-based synthetic response data may allow researchers to generate datasets with fewer constraints on time, cost, and privacy concerns [46]. As increasingly powerful LLMs have been introduced, researchers have studied their ability to mimic demographic sub-populations for surveys [3], explored them as tools for market research [4], and applied them in behavioral economics experiments [47]. The potential benefits explored by these studies highlight why synthetic response data has received significant attention within the social science research community.

[5] https://www.voiceform.com/
[6] https://outset.ai/

While many of these LLM-based simulation tools have demonstrated potentially promising results, other studies have honed in on several fundamental epistemic and representational risks of using LLMs to provide synthetic responses [48]. The beliefs, opinions, and perspectives espoused by LLM responses are a function of the data on which the language model was trained (and then fine-tuned). This can result in synthetic responses amplifying biases that do not accurately represent the diversity of human perspectives normally accessible through surveys. Extensive research shows that language models tend to exhibit variable beliefs, with Hartmann et al. finding a left-libertarian bias in ChatGPT (GPT-3.5), and Santurkar et al. identifying significant skews in the views expressed by multiple base and human-feedback tuned models from OpenAI and AI21 labs [49, 50]. Furthermore, Röttger et al. shows the lack of robust evaluation methods of the political stances expressed by LLMs, and their unstable responses in closed versus open-ended responses [51]. Wright et al. shows that LLMs' expressed stances are highly-sensitive to demographic features included in the prompts [52]. Research has also found that groups of marginalized people are susceptible to flattened caricatures in LLM simulations [53], as well as frequently misportrayed as stereotypical imitations from an out-group rather than a representative of that minority group [5], calling into question the ability of "persona"-based simulations to correct for biases in LLMs. These potential limitations mean researchers need to weigh the benefits of increased research speed, reduced costs, and decreased privacy exposure with the potential risks, both empirical and ethical. More work is needed to establish the validity of synthetically generated data for hypothesis testing across different domains. However, even given the current state of the literature on synthetic data, we believe a conservative and practical application aligns with our proposed survey creation pipeline: using synthetic data for survey pilot testing.

Though these issues suggest some reasons to be skeptical about LLMs as a replacement for humans in research, especially in the short term, they already allow researchers to run pilot tests in their current state. Pilot testing is an under-discussed but crucial aspect of survey development [54]. We have already discussed the idea of LLMs reviewing survey questions (and the complete survey) to ensure it is well-written. Beyond this, pilot testing may involve survey creators using LLM personas to answer questions while representing particular demographic groups (e.g., a researcher may plug in 10 key target populations and query synthetic responses to indicate how each group might respond to their survey). This can be used to create synthetic data, to set-up and test analysis prior to data collection (critical to a successful analysis, as it is easy to forget to capture or mis-specify something necessary for the analysis), and for considering which sub-populations may be interesting for over-sampling (i.e., it may be worth exploring populations that the LLM thinks may be interesting, or for which the LLM performs poorly when comparing to previously run surveys). As discussed by Bail, [55], social scientists are in a position to determine whether biases in LLMs are inherently *bad* or can be carefully controlled for in empirical tests [55].

In some ways, we think of LLMs in this context as an expansion of existing estimation, modeling, and imputation strategies, rather than a fundamentally new paradigm. For instance, existing statistical techniques which correct estimates from non-probability samples, such as multilevel regression with poststratification (MRP) which accounts for demographic and geographic variations in survey data to obtain estimated average responses from any demographic subpopulation, function very similarly [2]. More generally, survey modeling very often requires extrapolations in one form or another. The question, then, is whether LLMs allow us to extrapolate more flexibly or more confidently than existing methods, and how best to combine the information therein with the data we collect from human participants.

In the longer-run, while we expect human respondents to always be critical to surveys, the exact point where synthetic respondents end and human respondents begin will shift. This is already a fuzzy line: while it is natural to consider answers to surveys produced by LLMs out-of-the-box (without any human responses provided) "synthetic", what should we call LLM-imputed answers to surveys after seeing hundreds of human responses to similar questions? Over time, firms may also develop their own proprietary LLMs which address some of the biases discussed above by fine-tuning their model on thousands of survey questions with millions of responses: is the resulting data "synthetic" if we asked the resulting fine-tuned model how someone would answer a question today after it saw the results of 1,000 human respondents yesterday? We also expect that, for pricing experiments and other repetitive, highly correlated tasks, these hybrid responses could assist organizations in testing responses. In time, these surveys may accurately predict necessary target populations with fewer human respondents, and this could spur more surveys to be run in total.

Future Research Questions

- Methods: How do we improve the accuracy of synthetic responses: could combining the output of multiple models create more diverse populations of LLMs that are less at-risk of homogeneity? Is that more or less effective than fine-tuned models with additional survey data? Can LLMs help us establish baseline distributions that serve as better priors for Bayesian models? Can concerns about representational harms and bias with these methods be addressed with the right iterative process that mixes in real human responses with the models?

- Tools: Can tools be created to make it easy for survey researchers to use LLM synthetic responses for pilot testing (helping to create more efficient questions, target population, and analytic methods), but which also guard against misuse by researchers in replacing human respondents for main results?

Discussion

Short term benefits: After initial forays into synthetic responses, we have seen that there are promising opportunities for using LLMs in survey ideation and data analytics in the short-term. Question editing, and even question writing, are already impacting the industry with readily available tooling, but questions remain about how to minimize risks to validity and maintain the survey creator's control over the survey. LLMs open up new opportunities for surveys, they can be used to create customized questions that can be deployed in real-time in situations that would previously been too costly to survey. They also may help facilitate new types of analysis, especially for open-ended questions, flexible queries made in natural language, and improving reporting deliverables. And, everything can be tested and prepped very efficiently with pilot testing on synthetic respondents.

Long term benefits: In the long-term, (LLM-powered) dynamic, open-ended questions with LLM-powered labeling may be a much larger disruption, handing over even more agency to the models for the benefit of cost-savings and new markets for surveys to explore. We expect chat-based systems to create mixed-mode experiences that allow new scale and impact of surveys. Ideally, systems will integrate market intelligence, allowing stakeholders to explore survey results in the context of other key data to create a more continuous data stream. Augmenting survey creators' capacity to create new analysis, rather than replacing them, can help ensure this technology is empowering instead of displacing. Sometimes it is very subtle how a tool augments rather than replaces [26], by focusing on allowing the researcher to do new things (rather than old things faster), and learn in the process.

Proprietary models are not transparent: Our anecdotes suggest that off-the-shelf LLMs show some promise as survey writing aides, but none of the major public models have been trained or rigorously evaluated for this purpose. The data on which they were trained (and subsequent fine-tuning with guardrails added to the model) remain unknown, which can impact the suitability and performance of using LLMs for this purpose. Borrowing from the more general literature on LLM's impact on productivity (e.g., [56, 57], academic work can accept that limitation, and instead focus on the impact of LLM-based tools at each step on quality (however key stakeholders define it), effort (for the proper stakeholder), and cost (in time and money). Developing consistent metrics and continuous testing as the models and tools evolve can help responsibly integrate these developments. However, in general, research should also strive to focus on assessing, developing, and improving open models for the good of the wider research community [58], plausibly fine-tuned with tranches of survey data, where there are fewer unknowns in the sources, structure, and composition of the resulting model's data.

Platforms should be transparent: Survey companies and platforms should be transparent regarding their use and evaluation of LLM-enabled tools so that users can make informed decisions about adopting their suggestions and other stakeholders can understand what those decision were. Similar to the way transparent surveys now outline the procedure for identifying their audience[7], they should report back their LLM usage as best as possible. The more open the model, the lower the level of black-box fine-tuning, meta-prompts, and additional guardrails, the more stable and clean this transparency can be.

[7]Like the AAPOR transparency initiative https://aapor.org/standards-and-ethics/transparency-initiative/

Human researchers remains central to survey research: We have deliberately skipped over Step 0 in the survey research pipeline: we did not address cases in which a researcher is still unsure of what they want to know. Involving an LLM in the survey process before having a clear research agenda is difficult to explore formally and may be riskier to implement, because LLM-generated suggestions may impact many difficult-to-quantify aspects of research if they contribute to both the development of the question and to evaluation/measurement. And, we have also highlighted how the researcher is still central to the questions under any regime, still prompting and reviewing in the most extreme cases of interactive surveys. The researcher will continue to oversee the execution of the survey, the analytics, and the reporting back to stakeholder, even as the LLMs will augment their work at each step.

Human respondents remain central to survey research: Surveys remain some of the most important tools used across fields to gather data and learn more about people. If implemented rigorously, LLMs can improve survey research by easing and accelerating the path through ideation, data collection, analytics, and reporting generation. Implemented carelessly, we risk introducing errors into the survey pipeline that could decrease quality and/or homogenize the insights gained from surveying diverse populations: this is why human respondents will always be a critical part of the survey pipeline, whether they are the primary source of data or part of an iterative process where they help to update and calibrate models. In a world which lacks ground-truth, which relies on trust and transparency to ensure acceptance and impact, we are excited to see how the methodology and tooling of survey research evolve to meet this awesome disruption.

References

[1] Robert M. Groves. Three Eras of Survey Research. *Public Opinion Quarterly*, 75(5):861–871, 12 2011.

[2] Wei Wang, David Rothschild, Sharad Goel, and Andrew Gelman. Forecasting elections with non-representative polls. *International Journal of Forecasting*, 31(3):980–991, 2015.

[3] Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.

[4] James Brand, Ayelet Israeli, and Donald Ngwe. Using gpt for market research. Technical Report 23-062, Harvard Business School, April 2023. Revised July 2023.

[5] Angelina Wang, Jamie Morgenstern, and John P. Dickerson. Large language models cannot replace human participants because they cannot portray identity groups, 2024.

[6] Bernard CK Choi and Anita WP Pak. Peer reviewed: a catalog of biases in questionnaires. *Preventing chronic disease*, 2(1), 2005.

[7] Roger Tourangeau, Lance J Rips, and Kenneth Rasinski. The psychology of survey response. 2000.

[8] Norman M Bradburn, Seymour Sudman, and Brian Wansink. *Asking questions: the definitive guide to questionnaire design–for market research, political polls, and social and health questionnaires*. John Wiley & Sons, 2004.

[9] Timo Lenzner, Lars Kaczmirek, and Alwine Lenzner. Cognitive burden of survey questions and response times: A psycholinguistic experiment. *Applied Cognitive Psychology*, 24:1003–1020, 2010.

[10] Jacob Steiss, Tamara Tate, Steve Graham, Jazmin Cruz, Michael Hebert, Jiali Wang, Youngsun Moon, Waverly Tseng, Mark Warschauer, and Carol Booth Olson. Comparing the quality of human and chatgpt feedback of students' writing. *Learning and Instruction*, 91:101894, 2024.

[11] Juan Escalante, Austin Pack, and Alex Barrett. Ai-generated feedback on writing: insights into efficacy and enl student preference. *International Journal of Educational Technology in Higher Education*, 20(1):57, 2023.

[12] Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Smith, Yian Yin, et al. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *arXiv preprint arXiv:2310.01783*, 2023.

[13] Nora Cate Schaeffer and Jennifer Dykema. Questions for Surveys: Current Trends and Future Directions. *Public Opinion Quarterly*, 75(5):909–961, 12 2011.

[14] Naomi Forman-Katz. Americans are following the news less closely than they used to. *Pew Research Center*, 2023. Accessed: 2024-07-17.

[15] Anton J Nederhof. Methods of coping with social desirability bias: A review. *European journal of social psychology*, 15(3):263–280, 1985.

[16] Alice Cai, Ian Arawjo, and Elena L. Glassman. Antagonistic ai, 2024.

[17] Sam Witteveen and Martin Andrews. Paraphrasing with large language models. In *Conference on Empirical Methods in Natural Language Processing*, 2019.

[18] Barrett R Anderson, Jash Hemant Shah, and Max Kreminski. Homogenization effects of large language models on human creative ideation. In *Proceedings of the 16th Conference on Creativity & Cognition*, pages 413–425, 2024.

[19] Lisa Messeri and MJ Crockett. Artificial intelligence and illusions of understanding in scientific research. *Nature*, 627(8002):49–58, 2024.

[20] Jacob Doughty, Zipiao Wan, Anishka Bompelli, Jubahed Qayum, Taozhi Wang, Juran Zhang, Yujia Zheng, Aidan Doyle, Pragnya Sridhar, Arav Agarwal, Christopher Bogart, Eric Keylor, Can Kultur, Jaromir Savelka, and Majd Sakr. A comparative study of ai-generated (gpt-4) and human-crafted mcqs in programming education. In *Proceedings of the 26th Australasian Computing Education Conference*, ACE '24, page 114–123, New York, NY, USA, 2024. Association for Computing Machinery.

[21] Tanya Goyal, Junyi Jessy Li, and Greg Durrett. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*, 2022.

[22] Timothy D Wilson, Christopher E Houston, Kathryn M Etling, and Nancy Brekke. A new look at anchoring effects: basic anchoring and its antecedents. *Journal of Experimental Psychology: General*, 125(4):387, 1996.

[23] Gretchen B Chapman and Eric J Johnson. The limits of anchoring. *Journal of Behavioral Decision Making*, 7(4):223–242, 1994.

[24] Valdemar Danry, Pat Pataranutaporn, Matthew Groh, Ziv Epstein, and Pattie Maes. Deceptive ai systems that give explanations are more convincing than honest ai systems and can amplify belief in misinformation, 2024.

[25] Shubham Atreja, Joshua Ashkinaze, Lingyao Li, Julia Mendelsohn, and Libby Hemphill. Prompt design matters for computational social science tasks but in unpredictable ways, 2024.

[26] Jake M. Hofman, Daniel G. Goldstein, and David Rothschild. A sports analogy for understanding different ways to use ai. *Harvard Business Review*, 2023.

[27] Michelle S. Lam, Janice Teoh, James A. Landay, Jeffrey Heer, and Michael S. Bernstein. Concept induction: Analyzing unstructured text with high-level concepts using lloom. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery.

[28] Jialun Aaron Jiang, Kandrea Wade, Casey Fiesler, and Jed R Brubaker. Supporting serendipity: Opportunities and challenges for human-ai collaboration in qualitative analysis. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–23, 2021.

[29] Cassandra Overney, Belén Saldías, Dimitra Dimitrakopoulou, and Deb Roy. Sensemate: An accessible and beginner-friendly human-ai platform for qualitative data analysis. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, pages 922–939, 2024.

[30] Chau Minh Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. TopicGPT: A Prompt-based Topic Modeling Framework, April 2024. arXiv:2311.01449 [cs].

[31] Zihan Wang, Jingbo Shang, and Ruiqi Zhong. Goal-Driven Explainable Clustering via Language Descriptions, November 2023. arXiv:2305.13749 [cs].

[32] Mengting Wan, Tara Safavi, Sujay Kumar Jauhar, Yujin Kim, Scott Counts, Jennifer Neville, Siddharth Suri, Chirag Shah, Ryen W. White, Longqi Yang, Reid Andersen, Georg Buscher, Dhruv Joshi, and Nagu Rangan. TnT-LLM: Text Mining at Scale with Large Language Models, March 2024. arXiv:2403.12173 [cs].

[33] Michelle S. Lam, Janice Teoh, James A. Landay, Jeffrey Heer, and Michael S. Bernstein. Concept Induction: Analyzing Unstructured Text with High-Level Concepts Using LLooM. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–28, Honolulu HI USA, May 2024. ACM.

[34] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022, mar 2003.

[35] Gautier Izacard and Édouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, 2021.

[36] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

[37] Neil Malhotra, Jon A. Krosnick, and Randall K. Thomas. Optimal Design of Branching Questions to Measure Bipolar Constructs. *Public Opinion Quarterly*, 73(2):304–324, 05 2009.

[38] Emily E. Gilbert. A Comparison of Branched Versus Unbranched Rating Scales for the Measurement of Attitudes in Surveys. *Public Opinion Quarterly*, 79(2):443–470, 05 2015.

[39] Sara Dolnicar, Bettina Grün, and Venkata Yanamandram. Dynamic, interactive survey questions can increase survey data quality. *Journal of Travel & Tourism Marketing*, 30(7):690–699, 2013.

[40] Adeline Delavande and Susann Rohwedder. Eliciting Subjective Probabilities in Internet Surveys. *Public Opinion Quarterly*, 72(5):866–891, 11 2008.

[41] R Michael Alvarez and Yimeng Li. Survey Attention and Self-Reported Political Behavior. *Public Opinion Quarterly*, 86(4):793–811, 02 2023.

[42] Frederick G Conrad, Mick P Couper, Roger Tourangeau, and Chan Zhang. Reducing speeding in web surveys by providing immediate feedback. In *Survey Research Methods*, volume 11, page 45. NIH Public Access, 2017.

[43] Ziang Xiao, Michelle X. Zhou, Q. Vera Liao, Gloria Mark, Changyan Chi, Wenxi Chen, and Huahai Yang. Tell me about yourself: Using an ai-powered chatbot to conduct conversational surveys with open-ended questions. *ACM Trans. Comput.-Hum. Interact.*, 27(3), jun 2020.

[44] Soomin Kim, Joonhwan Lee, and Gahgene Gweon. Comparing data from chatbot and web surveys: Effects of platform and conversational style on survey response quality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–12, New York, NY, USA, 2019. Association for Computing Machinery.

[45] Ziang Xiao, Michelle X Zhou, Wenxi Chen, Huahai Yang, and Changyan Chi. If i hear you correctly: Building and evaluating interview chatbots with active listening skills. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020.

[46] James Jordon, Lukasz Szpruch, Florimond Houssiau, Mirko Bottarelli, Giovanni Cherubin, Carsten Maple, Samuel N Cohen, and Adrian Weller. Synthetic data-what, why and how? 2022.

[47] John J Horton. Large language models as simulated economic agents: What can we learn from homo silicus? Working Paper 31122, National Bureau of Economic Research, April 2023.

[48] William Agnew, A Stevie Bergman, Jennifer Chien, Mark Díaz, Seliem El-Sayed, Jaylen Pittman, Shakir Mohamed, and Kevin R McKee. The illusion of artificial inclusion. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2024.

[49] Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. The political ideology of conversational ai: Converging evidence on chatgpt's pro-environmental, left-libertarian orientation, 2023.

[50] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[51] Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Rose Kirk, Hinrich Schütze, and Dirk Hovy. Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models, 2024.

[52] Dustin Wright, Arnav Arora, Nadav Borenstein, Srishti Yadav, Serge Belongie, and Isabelle Augenstein. Revealing fine-grained values and opinions in large language models. *arXiv preprint arXiv:2406.19238*, 2024.

[53] Myra Cheng, Tiziano Piccardi, and Diyi Yang. CoMPosT: Characterizing and evaluating caricature in LLM simulations. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10853–10875, Singapore, December 2023. Association for Computational Linguistics.

[54] Jane Brooks, Deborah M Reed, and Barbara Savage. Taking off with a pilot: The importance of testing research instruments. In *ECRM2016-Proceedings of the 15th European Conference on Research Methodology for Business Management": ECRM2016. Academic Conferences and publishing limited*, pages 51–59, 2016.

[55] Christopher A. Bail. Can generative ai improve social science? *Proceedings of the National Academy of Sciences*, 121(21):e2314021121, 2024.

[56] Sofia Eleni Spatharioti, David M Rothschild, Daniel G Goldstein, and Jake M Hofman. Comparing traditional and llm-based search for consumer choice: A randomized experiment. *arXiv preprint arXiv:2307.03744*, 2023.

[57] Erik Brynjolfsson, Danielle Li, and Lindsey R Raymond. Generative ai at work. Technical report, National Bureau of Economic Research, 2023.

[58] Alexis Palmer, Noah A Smith, and Arthur Spirling. Using proprietary language models in academic research requires explicit justification. *Nature Computational Science*, 4(1):2–3, 2024.