Webinar Q&A: Artificial Intelligence and the Future of Survey Research

1. Very interested in the June Webinar as I might have had a conflict and couldn't attend. Can you provide the link to the recording and Q&A document?

   RESPONSE: You may access that information here: https://www.rti.org/event/small-area-estimation-survey-research

2. With cost concerns regarding token usage, what would be the downside of hosting an open source LLM with something like Ollama on local hardware instead of relying on ChatGPT/Claude/etc.?

   RESPONSE: The open source LLM would need to be installed on hardware and would need to be maintained so there would be costs associated with hosting an open source LLM.

3. How much time would you estimate you spent on the LLM approach vs. the human approach?

   RESPONSE: About half the time. Most of the time reduction came from not having to directly write R code.

4. For the frame development, what do you think contributed to the LLM being wrong ~25% of the time?

   RESPONSE: When searching school names to identify schools embedded in hospitals and correctional facilities, the keywords the LLM used are highly, but not perfectly, correlated with those types of schools. So, a school name with the word "health" in the name is not necessarily a school embedded in a hospital.

5. Related to the frame development, how were the number of categories in the classification scheme selected?

   RESPONSE: There are different types of classification variables including those used for sample stratification (both explicit and implicit) and those used to inform sample recruitment. The sample stratification variables were constructed to ensure that the school sample would be distributed geographically and across domains relevant to

student math and science achievement. The sample recruitment classifiers are generally comprised of simple binary classifiers that capture information about schools that could impact the ability to recruit schools.

6. How were the number of categories in the classification scheme selected?

   RESPONSE: We developed the coding scheme to meet the usual requirements of exhaustiveness and mutual exclusivity. We started from a set of labels drawn from the existing literature and we refined the scheme in two iterative rounds with human coders. We tried to adjust category definitions and boundary rules in response to disagreements that we had.

7. How did you organize your iterations? LLM tasks require a lot of back and forth - at what time did you decide to stop (basically say, this is good enough - no more revisions required)?

   RESPONSE: We did not go into the development with a hard rule for determining when to stop. Based on initial iterations, we expected almost perfect agreement between the LLM-generated frames and the person-generated frames on all variables except those where keyword searches of school name were used to create the variables. We iterated until the non-keyword search variables agreed except in about 10 cases but we allowed for a couple of hundred differences in the keyword search variables (due to the expected imperfect correlation between keywords and classification outcome).

8. How can we ensure for the correct weights in case of surveys including stratification and clustering?

   RESPONSE: It is unclear how this question relates to LLM-assisted frame development. If the question is about how to determine if weights are constructed correctly for a given sample design, then some common checks include: 1) For each stage of sampling, check to make sure that the sum of the weights matches the frame total within each stratum and overall and 2) that each weight is positive.

9. Did you test for any sensitivity in the fine-tuning (rather than the prompt-only methods) with the formatting of the input prompts (e.g. first-person vs. third-person; description, etc.)?

RESPONSE: Thank you! No, we did not. Mainly due to time constraints, as the first author had to submit her PhD thesis. But we also felt that performance might already have reached a ceiling.

10. Did you try fine-tuning the open source LLMs? Just wondering how good they could get. I think I only saw finetuning for ChatGPT.

    RESPONSE: Sadly no, mainly due to time constraints, as the first author had to submit her PhD thesis. However, my impression from other small, unpublished projects we conducted recently is that today's open-source LLMs may achieve similarly strong performance to the fine-tuned GPT model from 2025.

11. I imagine it would be difficult to catch certain mistakes that the language learning model can make when constructing a data frame (e.g., the school name). What methods would you suggest to flag these nuances? Would it be best to subset and manually check just so you can pick up on some of these issues? Can you ask for the model to not resolve these things autonomously and to flag it instead?

    RESPONSE: I have the LLMs generate R code, add documentation in the R code for various things like variable construction and merging, and then include in the output cross-tabulations and some record lists to check those sorts of things. Just how I would review work done by someone else (or myself).

12. (1) Can you talk a little more about how you did the fine tuning? How many/what percent of cases did you use for the fine-tuning model? (2) Was the improvement you saw from the fine tuning equally distributed across categories? Or did you see more improvement in, for example, the "other" category or other "catch-all" buckets? (3) Also, can you confirm that only one code was applied per comment analyzed?

    REPONSE: (1) We used several thousand cases for fine-tuning because we already had them available. However, the token-accuracy and loss curves suggest that just over 100 examples would likely have been sufficient. The figures are shown at the very end of the article's online appendix, if you would like to check them!
    (2) Fine-tuning mainly helped with fuzzier responses assigned to categories such as "help in general," "importance," "no reason," and "other." The other LLMs struggled with responses such as "no particular reason."
    (3) Yes, we assigned only one code per response. We specifically checked for responses mentioning two or more reasons but found only a handful among the 5,000 manually coded responses.

13. Do you have specific stats on the ROI of using these LLMs (labor hours needed with and without AI, including time spent learning and fine-tuning use of the tool and reviewing output; cost of tools/tokens; accuracy of output and frequency of errors) for your project?

    RESPONSE: ROI is very difficult to estimate in an unbiased way because we don't randomize tasks at work in a way that would allow us to. That said, from our experience, we find that we can take on more things, complete them on-time and to the satisfaction of stakeholders without the upticks in stress and overtime hours that would have once entailed. That's not a direct measurement, and not necessarily representative, but still quite meaningful. Some things in our own work life can be very clearly done more quickly, but sometimes tasks take just as long but are done to a higher level (say, with more thorough documentation or with extra QC). With respect to coding, specifically, the frequency of errors has dropped off quite significantly, especially with the latest frontier models like Claude Fable and GPT 5.6 Sol. At the same time, our level of intimacy with coding products is much lower, and thorough validation requires new techniques. In general, we are finding more and more that effective AI use is a significant skill unto itself.

14. Considering confidentiality restrictions, to what extent would the trade-off between explainability and accuracy impact the selection of an LLM over a more discriminative approach (e.g., LDA) to analyze unstructured data from national datasets?

    RESPONSE: In my experience the confidentiality constraint usually resolves the question before the explainability–accuracy trade-off becomes relevant: open-ended responses in national datasets carry high disclosure risk, so if the data cannot leave a secure environment, proprietary APIs are simply out of the choice set.

15. Does the agent already understand what the XML Tags mean or do they need to be defined in advance?

    RESPONSE They do not require any prior explanation, and the agents consistently act on them in obvious and effective ways.

16. Are there benefits in your workflow you have found to using xml for the framing compared to something like markdown, or is this just personal preference?

    RESPONSE: The xml tags were part of formal prompting guides from Anthropic a couple of years ago, and I picked it up secondhand by way of Reddit. At this stage,

the models are smart enough to interpret any sort of explicit request segmentation device we might find most convenient. We use the xml tags for direct prompting in conjunction with workspace-level markdown files.

17. What are your thoughts about how to address issues/limitations raised by using LLMs in academic papers? I'm writing a manuscript for a scoping review where we used an LLM to extract the data. We did this due to the large number of included studies (>2000) and limited human resources. We plan to discuss our use of the LLM in the limitations section, but we're not sure what we should say. One idea is to mention the error rates that we identified from human validation of the extracted data in 10% of the studies. What are your thoughts?

    RESPONSE: The standards of acceptable use are evolving along with the tools and as their use becomes more widespread—your paper will contribute to that, and human validation on a random sample is a good approach. It will also be important to be very explicit about exactly which tool was used, in which configuration, and on which dates. It will be interesting to see how these standards develop. The tools are too powerful to rule out entirely.

18. My question relates to the use of LLMs to write/revise/review drafts of papers. Journals are increasingly adopting policies about the use of AI in preparing manuscripts for publication. I worry that using LLMs in the manner described by Mr. Couzens may run afoul, whether real or perceived, of these policies. Thoughts?

    RESPONSE: We do not recommend using the method described by Lance to draft a manuscript, as that is an area where the document itself is the product, and it is one where one's personal voice and phrasing really do matter. That said, we would still populate a workspace, hook one or more agents up to it, and use them as research assistants and expert librarians. For example, in preparing the manuscript one might find that they want to see the results of a new simulation. The agent can write the code, execute the simulation, and prepare an html file for you to review.

19. Does the agent correct the revisions via pandoc? And then propose new revisions in the next iteration?

    RESPONSE: Once there are one or more Word documents with comments/edits from reviewers (including yourself), the agent ingests those documents via Pandoc and revises markdown version of the document (or creates a revision plan which you can review and revise/approve before action, depending on the nature of the document). Then, the agent creates a fresh Word document by converting the revised markdown document (the markdown document is always the source of

truth in this workflow and it's what the agent works on). It is often helpful to see tracked changes in the revised Word document relative to the prior version. To do this you can have your agent use a command line utility to create a redline version by comparing the before and after versions. At this point, the cycle continues, if necessary. Claude Code can be used for this workflow and have saved as a 'skill' (just a markdown file that lives in a specific folder) that you can invoke without having to explain anything to the agent.

20. No changes to any of this if using Google Docs (more easily sharable) rather than Mircrosoft Word, one assumes?

    RESPONSE: While we have not tried this ourselves, some preliminary research points toward using Pandoc to convert to an intermediate format (like .docx or .odt) and then uploading that to Google Drive with conversion from the command line with conversion flags. So, it seems like there might be an extra step in there, but it seems doable.

21. Do any of the speakers have insights into skip pattern analyses with LLM?

    RESPONSE: If the instrument specifications (or a codebook that details the skip pattern logic) are available to the analyst, then AI can create the code to verify adherence in a given data frame. This code can be applied to the survey data and identify any cases where the skip patterns are not adhered to.

22. Besides time costs, how does this approach impact monetary costs of a project?

    RESPONSE: The cost of the additional human time needed when not using AI usually outweighs the cost of the AI agent. For example, in the open-ended recoding example the total costs were less than 100 Euros for the whole process.

23. Can the agents, particularly for documentation, be configured to identify & protect/omit nonpublic information/PII, for example if product is for a government client?

    RESPONSE: Yes, we believe so. But, it should be noted that we have not done this yet. We have not yet found the need to expose PII to an agent to support documentation, even if the core statistical product acts on or processes PII. In most cases, disclosure can be avoided by simply keeping PII out of the workspace.