LANGER RESEARCH ASSOCIATES

SURVEY RESEARCH DESIGN · MANAGEMENT · ANALYSIS

White Paper

A Pilot Test of AI Coding of Open-Ended Survey Responses

This paper reports on a pilot test using artificial intelligence to code open-ended survey responses into quantitative categories. Compared with human coding, the test performed poorly. While our investigation was limited in scope, results suggest the need for caution in using AI for open-end coding when data quality is a priority.

Background

Open-ended questions in attitudinal surveys are of use in both qualitative and quantitative analysis. Qualitatively, open-end responses provide for deeper understanding of complex attitudes and experiences. Quantitatively, open-ends can be systematically coded into thematic categories without the priming that is inherent in presentation of pre-coded response options.

While quantitative content analysis is informative, human coding is highly labor intensive. In our customary process, two independent coders separately review and categorize each response based on a codebook that describes each category, then compare and reconcile their work, ensuring agreement on all codes assigned. (We recognize that some researchers may use less rigorous standards.) Categories are constructed inductively as warranted for new questions; for questions asked and coded in previous surveys, coders begin with the previously established codebook and add new categories as necessary.

The present research tested whether AI can reliably automate this time-consuming work without sacrificing data quality. We used a leading AI tool, selected for its assurances of data privacy and its promoted proficiency at complex reasoning and analysis, to code open-ended responses from the 2024 Texas Education Poll we produced for the Charles Butt Foundation.[1, 2]

As detailed below, the test produced poor results, including misalignment with human coding, misclassification of responses and inability to perceive tone or directionality. The need for close human oversight ran counter to the goal of increased productivity. These findings align with

[1] The tool we used, current at the time of this experiment (summer 2025), was Claude Opus 4 from Anthropic. We recognize that other uses of this or any other AI tool may have different results, based on the user's proficiency as well as the tool's user options and capabilities.

[2] We are grateful to the Foundation for its permission to use its survey data in this investigation.

several presentations at the 2025 Annual Conference of the American Association for Public Opinion Research, summarized (with other references) in Appendix A.

Our analysis admittedly represented a challenging test, given the complexity of the question we analyzed: "What do you think is going well with the public schools in your community?" Our analysts hand-coded responses into more than 40 categories. A further challenge emerged in analysis: While the question has a positive valence, some respondents instead cited what was *not* going well in the public schools. Such responses were misclassified in our AI test.

Coding OEs without category guidance

We asked the AI tool to code responses in two ways.[3] First, we provided it with all open-ended responses but not the coding categories we developed by previous hand-coding, informed by our knowledge of the topic area. We asked it to create its own categories so we could assess how well it might perform on occasions in which alignment with past trend is not a consideration and when it might be advantageous for the AI tool to draft the codebook. As with our own coding, we instructed it to assign multiple codes to a response if appropriate.

Reviewing the first 500 of 1,166 responses coded, we observed these results:

- No coding or obvious miscoding of substantive data in 173 cases (35 percent).
- No coding in 115 cases (23 percent) in which respondents said they had no opinion, skipped the question or gave a non-substantive response. (Human coders classified these as: No opinion, Refused, Not applicable, Other.) (Note that even when prompted with these categories the AI test performed poorly, as reported in the next section.)
- Acceptable coding in 212 cases (42 percent), albeit some with suboptimal category creation.

Details of these results follow.

- **Poor coding decisions.** Codes were plainly incorrect for 97 of these 500 responses. At times, a word in a response matched a word in the AI-created category name, but the coding decision otherwise did not make sense. For example:

    - "Our school does a good job with education. It's a small community so we have a good district" was coded as both "small class size/ratio" and "community involvement" (in addition to a more appropriate code, "academic quality").
    - "Parents seem to care" was coded as "teacher care and dedication."
    - "In my community I would say the kids are taught well" was coded as "community involvement."

    We also saw instances of AI's inability to detect nuance, tone or sarcasm, missing the respondent's intended meaning. This challenge was compounded by the directionality of the question; as noted, it asked explicitly what was going well, while some responses

[3] See Appendix B for the prompts that we provided to the AI tool.

expressed what was going poorly. In both cases below, the AI test did not detect the negative valence of the response.

  - "So far, all the schools in my community have too many people on the school boards that are in it for themselves and don't care much about what the teachers and kids need" was coded as "leadership/administration," among others. (We manually coded this as "general negative attitude toward public schools.")
  - "Doing a great job educating despite lack of funding or support by the Governor" was coded as "funding/budget," among others. (We manually coded this as "education/learning (general)" and "schools trying their best.")

In still other cases, assigned codes were simply nonsensical. One example:

  - "I have no children in the system and haven't for many years" was coded as "STEM/technology."

- **Uncoded responses.** In an additional 76 of the 500 responses, the AI test did not assign any code to a substantive response. Many such omissions include straightforward responses. To name a few – all uncoded:

  - "There is no violence"
  - "Managing to maintain a good teaching standard"
  - "Relatively low drop-out rate"
  - "Good quality teaching"
  - "The schools I see are modern and have all the latest equipment for the children to learn"

- **Poor category creation.** While the AI test's coding decisions for the remaining 212 cases were generally correct, some categories were either overlapping or overly broad.

  The AI test created 21 coding categories, compared with the 45 categories our analysts created. In some cases, this pushed distinct ideas into a single category. For example, our human coders included a category for "parent involvement" and another for "community involvement." By contrast, the test produced a single category called "community involvement" and assigned parent-related comments to this category, losing the distinction between parent and broader community-level involvement.

  In other cases, the test produced overlapping or duplicate categories. In one example, the AI test created one category for "teacher care and dedication" and another for "teachers – positive." Our 2024 coding combined these categories into a single category called "teachers/teacher quality." In another example, it created distinct categories for "sports," "extracurricular activities" and "programs (general)," often assigning multiple of these codes to a response or misclassifying them to the wrong category (e.g., identifying athletic-related responses as one or both of the other two categories, rather than "sports").

Responses being assigned multiple overlapping codes limits clarity in the resulting data and analysis. For example:

- "Hopefully sports programs" was coded as both "sports" and "programs (general)."
- "Lots of STEM programs" was coded as both "STEM/technology" and "programs (general)."

While it created fewer than half as many coding categories as our human coders, a related question was whether the AI test nonetheless created any unique categories that our human coders missed. It did not, with one minor exception: The test created a category for "Special Education" for three responses – "Inclusion, diversity, collaborative learning," "Inclusion of all students, diversity," and "Nice facilities and special education programs." We coded the first two as both "Diverse student body" and "Efforts for multicultural inclusion" and the last as "Programs/resources (unspecified) and "school infrastructure."

Coding OEs with categories provided

In a second test, we asked the AI tool to forget its coding decisions and re-code the same responses using the 45 coding categories we had developed by hand. Alignment between its coding and our own was poor. (Again, this was a challenging test, given a complex open-ended question with nearly four dozen coding categories and the possibility of multiple codes assigned for each response.)

In all, we found at least one mismatched code for 823 of the 1,166 responses (71 percent), including 153 in which no code was assigned even though nonresponse categories were provided. The remaining 29 percent were fully aligned. As before, puzzling and nonsensical coding decisions were common. In a few examples:

- A response of "sports" was coded as "food and nourishment." (We coded this as "sports and extracurricular activities.")
- Conversely, a response of "access to meals" was coded as both "sports and extracurricular activities" and "graduation rates." (We coded this as "food and nourishment.")
- "I think Texas public schools are declining" was coded as "general positive attitude of public education." (We coded this as "general negative attitude of public education.")
- "Not sure, as I don't have a child in school (and intend to send my future children to a charter school)" was coded as "actions from policymakers." (We coded this as "not sure.")
- "Jesus Christ" was coded as "Administrators/political leadership." (We coded this as "Other.")
- "Class size, quality of teachers, quality of electives" was coded as "fine arts/electives," "class size/student-teacher ratio" and "efforts for multicultural inclusion." (Our coders agreed on the first two codes but chose "teachers/teacher quality" in place of the third.)

Beyond issues with misclassification, another factor driving poor alignment between our coding and the AI test was its tendency to apply more codes to a given response than we did. Ignoring cases in which it did not assign any code to a response (in this analysis, most such cases were no opinion, not applicable or refusals), the AI test assigned an average of 1.5 codes per response compared with our 1.2 codes.

Limitations and conclusions

We produced a pilot test of AI's capability in coding an open-ended survey question. Our test has several limitations. We tested a single AI product at its then-current stage of development, used results to a single complex open-ended question, used our own independently human-coded results for comparison and used AI prompts that we designed. We used a tool that was readily available and user-friendly, but the affordable access option we selected included limits on processing volume and speed. We used a web interface rather than an API, meaning model output was less customizable, but spared us the time and expense of building an API interface.

Further investigation with other AI products, other open-ended questions, other human coding and other prompts would test the generalizability of our results. Asking the tool to explain its codes, for example, may have improved performance in coding the question used in this analysis, albeit again at greater time and/or expense, posing a risk of sunk-cost fallacy. Prompting the tool with category descriptions and examples, rather than just category names, may have improved the test performance, but would work only in cases in which a verified, human-produced codebook was available.

There are potential challenges in the use of AI that are not addressed in our test. These include the role of system prompts, data security, the risk of bias in LLM training and the question of energy consumption. Our simple aim was to see if our use of AI could simplify the time-consuming task of coding open-ended survey responses without sacrificing data quality.

The test's mismatches with gold-standard human coding are discouraging. Further advances and additional testing may produce better results. Our question is not only whether AI can code open-ends efficiently, but whether it can do so with the level of accuracy that we require. We continue to monitor and test this and other potential applications of AI in our research practice.

Appendix A: Conference notes and other references

Notes on AI coding presentations at the 2025 conference of the American Association for Public Opinion Research

Sessions covered:

The Promise and Pitfalls of AI in Qualitative Social Science Research
Wednesday, 5/14/25, 10:45am

AI as OpenText Coder & Classifier
Wednesday, 5/14/25, 3:45pm

Several presenters discussed the use of AI to code, analyze or otherwise complement their qualitative research. Presentations underscored significant and concerning limitations that called into question the ability of AI to replace or complement human coding at its then-current state of development. Key challenges included poor alignment with human coders, replicability challenges and misidentification of key topics.

Poor alignment with human coders

- Samantha Collins (ICF) and her team used AI to code OE responses in an annual health surveillance program administered by the CDC, finding that AI-coded data had a 76 percent alignment with their own manual coding. In another presentation, Jake Nelson (Harris Poll) shared strikingly similar results in which AI coding of OE responses aligned with manual coding in 75 percent of cases.

- Jessica Eckard (Alan Newman Research) likewise compared human and AI coding of responses to OE questions, finding that AI performed better on simple questions but aligned poorly on more complex questions.

- Collins (ICF) further shared that AI-generated coding of OE responses was prone to misclassification, produced overly generalized coding schemes and missed tone, sarcasm and cultural references. Her team found that misclassification was common and "close human oversight is needed."

- Nelson (Harris Poll) and Eckard (Alan Newman Research) both noted that AI coding produced categories that completely overlapped one other. Nelson concluded that AI was ready to code OEs in a "strongly supervised capacity" that requires humans to edit, validate and refine the coding it produces.

Spanish translation

- Using AI for a different purpose but with similar results, Ana Gonzalez-Barrera (KFF) presented on the use of AI for instrument translation, comparing several LLMs against human translations and conventional web-based translation tools.

While AI performed at least as well as conventional internet-based translation tools, key limitations include its tendency to misinterpret tone, poorly detect contextual meaning and miss cultural nuance.

Lack of replicability

- Masahiko Aida (Applecart) shared results in which, problematically, AI tools generated notably different coding of OE data depending on which LLM was used. Results also varied within a single AI tool when the same data were tested at two time points.

- Similarly, Michael Schober (Census Bureau) presented results on the use of AI to analyze and code large datasets of tweets. He and his colleagues also found that different LLMs generated different coding schemes and summaries of the same data. They concluded that "constant and intense validation of output is essential."

- Eckard (Alan Newman Research) emphasized that AI tools often use external data to inform its coding, a key challenge for replicability.

Missing the mark topically

- Bob Torongo (Benenson Strategy Group) shared work in which he uploaded transcripts as training data and used them to generate synthetic AI personas as a tool to pretest a qualitative discussion guide with the aim of shortening the timeline needed to field the instrument.

  The model performed poorly: While the overwhelming theme from the training data was that 2024 votes were driven by economic factors, few AI responses mentioned the economy. He concluded that it "didn't work to test the instrument" and that it was "very dubious as a tool, for now." He also noted that "AI did not reduce time constraints."

Other references:

Morgan, D. L. (2023). Exploring the Use of Artificial Intelligence for Qualitative Data Analysis: The Case of ChatGPT. International Journal of Qualitative Methods, 22. https://doi.org/10.1177/16094069231211248 (Original work published 2023)

Tai, R. H., Bentley, L. R., Xia, X., Sitt, J. M., Fankhauser, S. C., Chicas-Mosier, A. M., & Monteith, B. G. (2024). An Examination of the Use of Large Language Models to Aid Analysis of Textual Data. International Journal of Qualitative Methods, 23. https://doi.org/10.1177/16094069241231168 (Original work published 2024)

Dunivin, Z. O. (2024). Scalable qualitative coding with LLMs: Chain-of-thought reasoning matches human performance in some hermeneutic tasks. *arXiv preprint arXiv:2401.15170*.

Williams, R.T. (2024) Paradigm shifts: exploring AI's influence on qualitative inquiry and analysis. Frontiers in Research Metrics and Analytics. 9:1331589. doi: 10.3389/frma.2024.1331589

Appendix B: Prompts

Prompt 1

You are tasked with coding open-ended responses collected from a survey about public education in Texas into qualitative themes. The data you will be coding is in the provided spreadsheet, which includes the following columns:

Column a: Case ID
Column b: Responses to the open-ended question, "What do you think is going well with the public schools in your community?"

Your job is to code responses in column B into distinct topical categories. To do this, produce a new spreadsheet that contains the same columns and information as the original dataset, but adds additional columns for each coding category that you create. Do not change the order of the rows.

If a response falls under a category, code the cell with the number 1. A response may be coded as 1 under multiple categories if it is appropriate. If a response does not fall under the category, leave the cell empty. When you create the spreadsheet, organize the columns that contain coding categories by broader general topic so that similar categories are together.

Prompt 2

Forget the codes that you just assigned to each response.

Code the responses again but instead of creating your own categories, use the categories that I am providing in the column headers in the new version of the spreadsheet that I am providing. Your job is to code responses in column B into the provided categories in column C through column AW. To do this, produce a new spreadsheet that contains the same columns and information as the original dataset, but adds your coding for each response in column C through column AW. Do not change the order of the rows.

If a response falls under a category, code the cell with the number 1. A response may be coded as 1 under multiple categories if it is appropriate. If a response does not fall under the category, leave the cell empty.