Down the LLM Rabbit Hole: Evaluating the Performance of Different Large Language Models in Coding Open-Ended Questions

Michael Link and Nick Bertoni

Ipsos - US Public Affairs

4/17/24

Federal Computer Assisted Survey Information Collection (FedCASIC) Workshops 2024

Study Background

Today's Question: What are the strengths & Limits of using LLMs to code open-ended survey responses?

What Are Large Language Models (LLMs)?

- Large Language Models (LLMs) are like digital brains that read vast amounts of text, helping them understand and generate human language.
- They predict what comes next in a text, similar to guessing a story's outcome.
- LLMs can analyze survey text, identifying patterns and main ideas, akin to a speed reader summarizing a stack of books.
- They are not perfect and may miss or misconstrue things, so human understanding of the process and review is essential.
- Examples: OpenAI Chat GPT; Google Gemini; Meta Llama 2; Falcon

Current Uses of LLMs in Coding Open-Ended Responses

- ✓ Text classification – theme identification
- ✓ Text classification – predefined categories
- ✓ Sentiment Analysis

Methods

1. Following a “Panelist First” approach, we use short 2 question surveys to assess panel satisfaction in the Ipsos KnowledgePanel – are they satisfied with their experience as a panelist (yes/no) and why/why not (open-ended).

2. Combined responses from three rounds: new panelists, Hispanic and African-Americans, and young adults aged 18-29, yielding 5,526 responses.

3. Selected valid open-ended responses: 3,056 affirmatives & 135 negatives.

4. From the 'Yes' respondents, we randomly chose 135 to match the number of 'No' respondents, resulting in a total of 270 responses for analysis.

5. Ipsos uses internal platform (“Ipsos Facto”) to provide access to the latest versions of Large Language Models (LLMs).

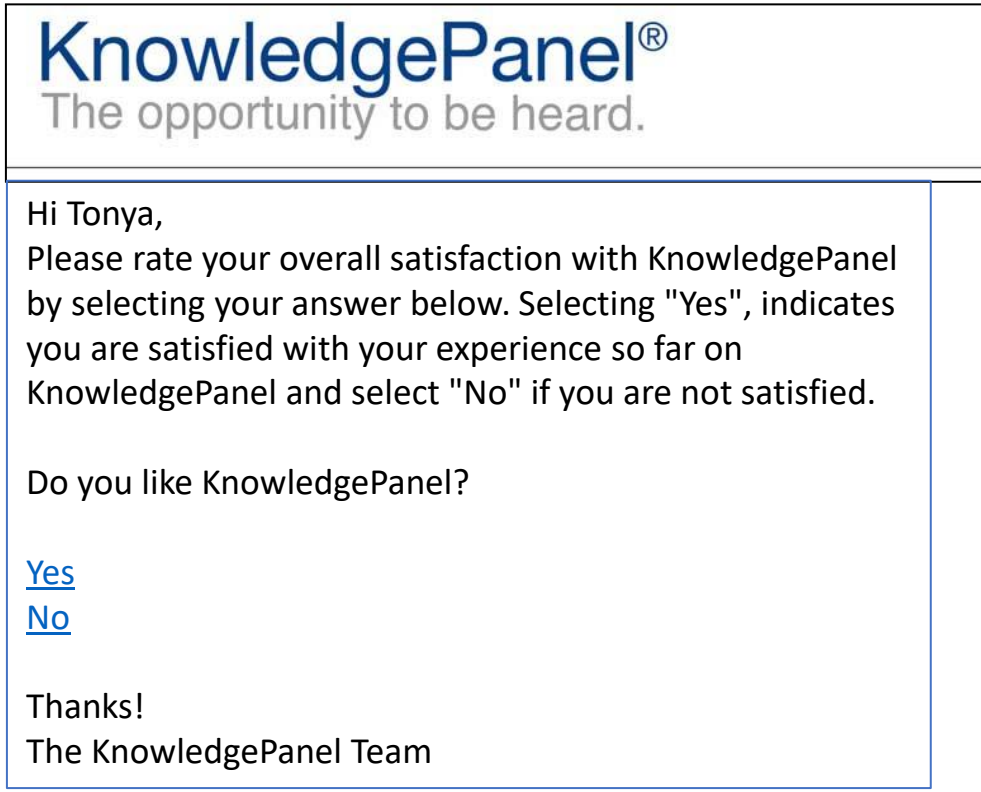
- Focused on use of Open AI GPT Ver 3.5 Turbo, 4 & 4 32k

6. Tested:

(a) theme identification at pooled response level varying LLM version;

(b) same model, but varying time (when runs were submitted); and

(c) used Ipsos tool TACTIC to examine respondent-level coding w same LLM model, but varying time



Test 1: Same Version of ChatGPT – Different Days (Vary Time of Submission)

Question to Ponder: What do you get when you load all responses into an LLM and ask it to provide the top themes across all responses – **using the same model, but varying the days/times** you submit the request?

Approach:

- Using the same prompt used before (below) and open-ended responses from the 270 respondents, asked LLM to identify the main themes – this time using the same model (OpenAI GPT 4.0 – but three different runs across three different days

"Provide your Prompt" is the new "Provide your questionnaire or weighting code"

PROMPT: Disregard previous instructions and prompts: Using your proficiency in natural language processing (NLP), analyze and categorize responses from an open-ended survey question: "Why are you (satisfied / not satisfied) with your experience so far on KnowledgePanel"; the analysis should identify key themes in the responses, such as satisfaction with resources, dissatisfaction with user interface, etc. The output should be a readable listing of the top themes and the percentage of responses that fall into each theme category

Test 1: Same Version of ChatGPT 4.0 – Different Days
(Vary Time of Submission)

Rank	4/26/24 4:30Pm	4/27/24 1:42PM	4/28/24 6:13PM
1	Satisfaction w/ rewards & compensation: 26%	Rewards and Compensation: 25%	Rewards and Compensation: 27%
2	Enjoyment of surveys & contribution: 24%	Enjoyment and Interest in Surveys: 20%	Enjoyment and Engagement: 23%
3	Ease and simplicity of surveys: 17%	Feeling Heard and Having a Voice: 15%	Impact and Contribution: 12%
4	Appreciation for being heard & valued: 13%	Survey Length and Frequency: 10%	Survey Length & Frequency: 11%
5	Desire for more surveys: 9%	Rewards Redemption and Availability: 8%	User Interface & Technical Issues: 7%
6	Dissatisfaction with rewards and compensation: 7%	Survey Content and Topics: 8%	Prize Drawings & Sweepstakes: 6%
7	Interest in topics and learning: 6%	Survey Process and User Experience: 7%	Language & Translation: 5%
8	Technical issues and concerns: 5%	Survey Frequency and Lack of Surveys: 6%	Lack of Surveys: 4%
9		Privacy and Security: 4%	Politics and Bias: 3%
10		Survey Relevance and Incentives: 3%	Personal Information & Privacy: 3%
Total %	**107%**	**106%**	**101%**

Test 2: Different Versions of OpenAI ChatGPT (Vary Model Versions)

Question to Ponder: What do you get when you load all responses into an LLM and ask it to provide the top themes across all responses – **using different LLM models each time, within the same timeframe**?

Approach:

- Using the same prompt (below) and open-ended responses from the 270 respondents, asked LLM to identify the main themes
- Tested 3 different versions of OpenAI LLM: GPT-3.5, GPT-4, GPT- 4 32k
- Note that the LLM does not apply coded output at the respondent level, rather it provides the top themes and the percentage of cases in which that theme was identified

Differences in Model Versions:

- Model size / # parameters
- Training data used
- Training algorithms
- Performance
- Scalability

PROMPT: Disregard previous instructions and prompts: Using your proficiency in natural language processing (NLP), analyze and categorize responses from an open-ended survey question: "Why are you (satisfied / not satisfied) with your experience so far on KnowledgePanel“; the analysis should identify key themes in the responses, such as satisfaction with resources, dissatisfaction with user interface, etc. The output should be a readable listing of the top themes and the percentage of responses that fall into each theme category

Test 2: Different Versions of OpenAI ChatGPT
(Vary Model Versions)

Rank	GPT 3.5	GPT 4.0	GPT 4.0 32k
1	Satisfaction with rewards and compensation: 26%	Satisfaction with Surveys and Participation: 40%	Satisfaction with earning potential and rewards: 42%
2	Enjoyment of surveys & contribution: 24%	Rewards and Compensation: 25%	Enjoyment of the surveys: 38%
3	Ease and simplicity of surveys: 17%	Desire for More Surveys: 10%	User experience with the surveys: 35%
4	Appreciation for being heard and valued: 13%	Discontent with Survey Length and Repetition: 10%	Dissatisfaction with rewards: 30%
5	Desire for more surveys: 9%	User Interface and Accessibility: 5%	Dissatisfaction with survey frequency and content: 28%
6	Dissatisfaction with rewards and compensation: 7%	Issues with Rewards System: 5%	Trouble with user interface and technical issues: 15%
7	Interest in topics and learning: 6%	Language and Cultural Consideration: 5%	Desire for more transparency and communication: 12%
8	Technical issues and concerns: 5%		
Total %	**107%**	**100%**	**186%**

Implications

- In each instance, LLMs provided theme identification and a summary in less than 60 seconds
- The themes made sense and matched with visual inspection of the open-ended data (**construct validity**)
- In each run, however, the results varied:
    - Different labeling of themes
    - Different percentage of cases in which themes were recognized
- Variations seen across different versions of LLMs and across time using same versions, prompts and data (**Lack of consistency**)
- Percentages added to >100% in most instances, indicating the models assigned more than one code per case in its calculations
    - ✓ Note that my prompt did **NOT** specify a unique code or theme per case (**Researcher error)**

Why the Variation? The Black Box is Complex!

1. LLMs are probabilistic (not "rule-based"): as such there will always be a degree or randomness or uncertainty involved in generating results.

2. Training Data Variability: LLMs models are trained on diverse datasets -- variations in the data type and volume significantly influence theme interpretation in open-ended text.

2. Architectural Design: structure and complexity of different models can cause disparities, with certain designs better suited to specific themes or contexts.

3. Contextual Interpretation: extent of a model's contextual understanding can vary, influencing how themes are identified and coded.

4. Linguistic Capability: size of a model's vocabulary and its proficiency in understanding language can influence theme coding, as these factors vary across models.

5. Model Evolution: updates and enhancements introduced in newer model versions can create variations.

6. Inference Mechanisms: methods employed for inference -- like beam search (multiple possible initial responses created, then secondary algorithm selects the one chosen) -- can result in variations in theme coding.

Using LLMs for analysis suffers from the **"Perpetual Dynamic Algorithm Problem"** – that is, they are platforms that evolve by design, are probabilistic and over time, therefore, make replicability of results difficult at best

Note: There are parameters on can use to "dial back" the level of random variability in the models (e.g. "Temperature", which will help when the same model is used – but not across different LLM models

Moving from Pooled Theme Generation to Respondent-Level Theme Coding

While use of general LLM models will analyze a corpus of input, then generate a specified formatted output, as researchers we want to apply a theme code at the respondent / case level

Ipsos developed a suite of tools called **TACTIC** -- Allows researchers to use the power of LLMs to code both theme & sentiment at the respondent level

How it works:

1. **Label Discovery:** TACTIC uses OpenAI GPT-3.5 Turbo to generate candidate labels (an initial "code book": by providing a sample of the verbatims to the LLM and asking the LLM to extract themes.

2. **Labeling:** TACTIC iteratively asks the LLM to classify the verbatims, considering all the candidate labels.

3. **Label Selection:** We developed an algorithm that ranks the labels by maximal coverage and minimum intercorrelations. We then select a final optimal label set by scoring the model performance versus the number of features included.

Two items to note:

1. The platform prioritizes understanding the meaning of the labels and the verbatim over direct keyword matches.
2. The LLM is considering each label in the context of the other candidate labels.

Ipsos TACTIC: Text Analytic Comprehensive Toolkit

Test 3: Using LLMs to Theme Code at the Respondent Level

Question to Ponder: How to results vary when LLMs are used to code open-ended response themes at the respondent level (not across the pool of data)? How might these results vary across multiple runs?

Approach:

- Upload the open-ended data file from 270 respondents
- Click “One Step” – which automates the three steps in sequence: labeling discovery; initial labeling; final labeling
- TACTIC currently uses OpenAI GPT-3.5 turbo
- Theme code added into the data set at the respondent / case level for additional analyses – similar to when open-ended codes are applied manually by researchers
- Conducted two data runs with identical parameters – within 15 minutes of each other

Test 3: Respondent Level Coding: GPT 4.0

Top 10 Codes	Run #1 4/27/24 5:15PM	Run #2 4/27/24 5:38PM
1	Survey Rewards: 12.2%	Survey Compensation: 10.0%
2	Survey Frequency: 9.3%	Survey Length: 8.9%
3	Survey Compensation: 9.6%	Survey Frequency: 8.5%
4	Survey Interest Level: 8.1%	Unique Surveys: 8.1%
5	Survey Length: 8.1%	Survey Clarity: 7.8%
6	Survey Topics: 7.0%	Helping thru Surveys: 6.7%
7	Survey Participation: 5.9%	Minimum Compensation: 6.3%
8	Survey Point System: 4.8%	Lack of Surveys: 5.9%
9	Survey Opinion Value: 4.8%	Low Compensation Rate: 5.9%
10	Survey Completion time: 3.3%	Sharing Knowledge: 4.4%
	All Other Codes (<3% ea.): 23.9%	All Other Codes (<4% ea.): 24.5%
Total # Categories	26	24
Overall % Cases Coded	96.7%	97.4%

Findings:

- Major themes are similar
- Greater variation in themes with fewer respondents
- Percentages vary across even similar themes across the two runs

Question: Would these findings change your conclusions?

Takeaways

- Utilization of LLMs in survey research is in early stages, yet it holds significant potential.
- Although useful, they should not be seen as a flawless solution, as factors such as training data, potential bias, inherent randomness, and the research goal's accuracy level must be considered.
- We cannot treat these as simply “1-click, GUI interfaces” with little understanding of what is happening inside the “black box”
- Variability and replicability issues exist due to the way in which LLMs operation and evolve over time
- “Fitness for Use” is a good standard to apply to use of results, along with recognition of limitations and the impact of slight alterations, such as model used, run-to-run variation, and variations in prompts used.
- LLMs, like many other analytical tools, require human guidance and review to achieve desired results; they supplement our experience and quality control, not replace them.
- LLM is advantageous when combined with research expertise, providing a 'win-win' solution.

Human expertise and understanding of the computational abilities – and limits -- of LLMs is essential to fully realizing their potential in any disciple and preventing the production of erroneous results.

Thank you!

Michael Link, Ph.D. & Nick Bertoni

Michael.Link@Ipsos.com

Ipsos U.S. Public Affairs
Celebrating 25 Years of the KnowledgePanel Providing Quality Insights Quickly!