Evaluating the Human-Likeness of LLM-Generated Open-Ended Responses

Anonymous authors
Paper under double-blind review

Large-language models (LLMs) are increasingly utilized within survey research for various tasks, including drafting questionnaires, summarizing open-ended responses, and generating synthetic respondents—"silicon samples"—that replicate human response distributions efficiently (Argyle et al., 2023). However, their effectiveness in generating open-ended survey responses, which are crucial for capturing nuanced and spontaneous participant thoughts, remains less explored. Preliminary evidence (Lerner et al., 2024; Huang et al., 2025; Bisbee et al., 2024) indicates notable systematic differences between LLM-generated and authentic human responses, particularly regarding verbosity, lexical diversity, and specificity. Thus, a critical methodological question arises: what precisely constitutes a human-like open-ended response, and how can we effectively evaluate synthetic responses generated by LLMs?

Drawing primarily from natural language processing, artificial intelligence, and survey-methodology literatures, this study proposes a structured framework for assessing the "humanness" of LLM-generated survey responses along four dimensions. We posit that these dimensions reflect the qualities typically observed (Krosnick, 1991; Smyth et al., 2021) among attentive, effortful, high-quality respondents (hereafter "authentic human respondents") as opposed to satisficers:

- **Parsimony**: Authentic human respondents generally provide concise answers, typically only a sentence or two, whereas LLM outputs frequently exceed typical response lengths.
- **Heterogeneity**: Authentic human respondents exhibit substantial lexical diversity and thematic variation, reflecting varied life experiences. Synthetic responses often cluster around common vocabulary and limited themes.
- **Noise**: Authentic human respondents include informalities, grammatical errors, truncated sentences, and pragmatic shortcuts. LLM outputs tend to lack such imperfections, producing uniformly polished text.
- **Contextual Specificity**: Authentic human respondents frequently embed personal or situational references (e.g., occupations, locations), whereas LLMs without targeted prompts default to generalized, abstract language.

To operationalize these dimensions, we propose metrics—including word and sentence counts, readability scores, corrected type–token ratios, embedding or topical variance, grammatical-error frequencies, and frequency of personal references—to develop a composite *humanness rating*. This rating systematically quantifies differences between synthetic and authentic responses, enabling clear diagnostic insights.

Our study uses data from 1,024 AmeriSpeak respondents who answered two open-ended survey questions: one on personal definitions of artificial intelligence and another on the most important national issue. Synthetic responses are generated with GPT-3.5-Turbo, GPT-4-o, Claude-3, and Llama-3-70B-Instruct, plus a GPT-3.5 model fine-tuned on 100 randomly selected authentic responses. Each model generates 1,000 synthetic answers.

The analytical approach is twofold. First, multiple machine-learning classifiers trained on the humanness-rating metrics identify the textual features most effective at distinguishing synthetic from authentic responses – we can classify and predict which features are most deterministically human vs. machine-generated.

Second, a supplementary, smaller-scale Survey Turing Test (STT) provides qualitative validation: blinded human raters and LLMs classify a stratified sample of synthetic and authentic responses. Near-chance accuracy would suggest human-like quality, whereas

higher accuracy would highlight residual artefacts; thus, the STT complements quantitative diagnostics without driving them.

This research anticipates significant improvements in parsimony and noise through targeted prompting and fine-tuning strategies, whereas heterogeneity and contextual specificity are expected to pose greater challenges due to their deeper linguistic complexity and situational grounding. Identifying these residual gaps will inform future methodological enhancements, potentially incorporating retrieval-augmented generation or demographic-specific conditioning.

Overall, this research contributes to survey methodology and NLP by clearly defining and operationalizing "humanness," establishing a transparent evaluation framework, and providing guidance for the responsible integration of LLM-generated open-ended responses into survey research. This will practically enable a better distinction between human and LLM-generated responses to open-ended questions, ultimately paving the way for developing synthetic responses that share the same characteristics that make open-ended responses human. If synthetic open-ended responses are ever going to be useful to survey practitioners, it is necessary that they can approximate human conventions as responses.

References

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. doi: 10.1017/pan.2023.2.

James Bisbee, Joshua D. Clinton, Cassy L. Dorff, Brenton Kenkel, and Jennifer M. Larson. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, 32(4):401–416, 2024. doi: 10.1017/pan.2024.5.

Lilian Huang, Brandon Sepulvado, and Joshua Y. Lerner. Llms don't respond like humans. In *80th Annual Conference of the American Association for Public Opinion Research (AAPOR)*, Philadelphia, PA, 2025. Slide deck.

Jon A. Krosnick. Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5(3):213–236, 1991. doi: 10.1002/acp.2350050305.

Joshua Y. Lerner, Brandon Sepulvado, Ipek Bilgen, Leah Christian, and Lilian Huang. Enhancing survey research data quality with llms: Using ai to optimize open-ended questions. In *79th Annual Conference of the American Association for Public Opinion Research (AAPOR)*, Atlanta, GA, 2024. Slide deck.

Jolene D. Smyth, Don A. Dillman, Leah M. Christian, and Austin C. O'Neill. *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method*. Wiley, Hoboken, NJ, 4th edition, 2021.