Report

Social Science Computer Review
2026, Vol. 0(0) 1–19
© The Author(s) 2026
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/08944393251408022
journals.sagepub.com/home/ssc
Sage

Identifying Bots Through LLM-Generated Text in Open Narrative Responses: A Proof-of-Concept Study

Joshua Claassen[1], Jan Karem Höhne[1], Ruben Bach[2], and Anna-Carolina Haensch[3]

Abstract

Online survey participants are frequently recruited through social media platforms, opt-in online access panels, and river sampling approaches. Such online surveys are threatened by bots that shift survey outcomes and exploit incentives. In this proof-of-concept study, we advance the identification of bots driven by Large Language Models (LLMs) through the prediction of LLM-generated text in open narrative responses. We conducted an online survey on same-gender partnerships, including three open narrative questions, and recruited 1512 participants through Facebook. In addition, we utilized two LLM-driven bots, each of which responded to the open narrative questions 400 times. Open narrative responses synthesized by our bots were labeled as containing LLM-generated text ("yes"). Facebook responses were assigned a proxy label ("unclear") as they may contain bots themselves. Using this binary label as ground truth, we fine-tuned prediction models relying on the "Bidirectional Encoder Representations from Transformers" (BERT) model, resulting in an impressive prediction performance: The models accurately identified between 97% and 100% of bot responses. However, prediction performance decreases if the models make predictions about questions they were not fine-tuned with. Our study contributes to the ongoing discussion on bots and extends the methodological toolkit for protecting the quality and integrity of online survey data.

Keywords

LLM-driven bots, Data quality and integrity, Large Language Models (LLMs), Machine learning, Response behavior, Web surveys, Explainable AI

[1]Research Infrastructure and Methods Division, German Centre for Higher Education Research and Science Studies, Leibniz University Hannover, Hannover, Germany
[2]Mannheim Centre for European Social Research (MZES), University of Mannheim, Mannheim, Germany
[3]Social Data Science and AI Lab, Ludwig Maximilians University Munich, Munich, Germany

Corresponding Author:
Joshua Claassen, Research Infrastructure and Methods Division, German Centre for Higher Education Research and Science Studies, Leibniz University Hannover, Lange Laube 12, 30159 Hannover, Germany.
Email: claassen@dzhw.eu

Introduction

Online surveys have increasingly replaced traditional survey modes, especially face-to-face interviews (Callegaro et al., 2015; Schober, 2018). Many prominent survey programs, such as the European Social Survey (ESS) and the European Values Study (EVS), have adopted online data collection methods. Online surveys offer significant advantages in reducing expenses and saving time, making them a strong option for meeting the rising need for survey data (Knowledge Sourcing Intelligence, 2023). Nevertheless, online surveys face methodological challenges. A primary issue is their tendency to achieve low response rates. For instance, the meta-analysis by Daikeler et al. (2020) indicates that participation rates in online surveys are approx. 12% lower than those in other survey modes (see also Lozar Manfreda et al., 2008).

Given the challenges of low participation rates in online surveys, researchers are exploring alternative methods for recruiting participants, such as social media platforms (Zindel, 2023), opt-in online access panels (Lehdonvirta et al., 2021), website ads or pop-ups (so-called river sampling; Murray-Watters et al., 2023), and crowdsourcing platforms (Peer et al., 2022). Although these methods allow for rapid access to a vast and diverse pool of participants, concerns arise about the quality and integrity of the data collected. One major concern is bots—automated programs that interact with digital systems, including online surveys (Griffin et al., 2022; Höhne, Claassen, Shahania, & Broneske, 2025; Storozuk et al., 2020; Xu et al., 2022; Yarrish et al., 2019; Zhang et al., 2022). Bots can distort survey results, potentially biasing political and social decisions (Xu et al., 2022). This is particularly concerning given evidence of bots being used to sway public opinion, such as during the 2016 referendum on the UK's departure from the European Union (Gorodnichenko et al., 2021) and the South Korean presidential election of 2022 (Zhang et al., 2024). The impact of bots on online surveys can be severe. First, responses synthesized by bots often differ from those of humans, introducing measurement error in the data (Xu et al., 2022). Second, the involvement of bots can erode confidence in social science research, exacerbating the impact of misinformation on public discourses (Xu et al., 2022). Finally, bots can cause both direct financial losses by exploiting survey incentives and indirect costs due to the substantial effort required for their identification and prevention (Storozuk et al., 2020; Xu et al., 2022).

Existing Strategies for Detecting Bots in Online Surveys

Most recently, an online survey on the car manufacturer Tesla was shut down early because of suspiciously high completion rates and sudden shifts in survey outcomes, pointing to potential bot infiltration (see https://www.t-online.de/finanzen/aktuelles/wirtschaft/id_100642002/tesla-umfrage-wegen-manipulationsverdacht-gestoppt-musk-teilt-artikel.html). Despite the significant threat of bots, studies focusing on bots in online surveys remain very limited. The few existing investigations mostly focus on simple prevention and identification strategies. One commonly used approach is to employ CAPTCHAs (challenge-response tests), which require participants to complete specific tasks, such as identifying objects in images, to block bots from entering online surveys (Storozuk et al., 2020). Another method involves honey pot questions. These questions are hidden queries embedded in the survey's source code that are invisible to human participants but are captured and potentially responded to by bots, making them a tool for identifying fraudulent bot responses (Bonett et al., 2024). Furthermore, the analysis of paradata (i.e., auxiliary data describing the data collection process; West, 2011), such as response times, is considered an effective way to identify bots, as their response speed may not align with the complexity of survey questions or tasks (Nikulchev et al., 2021).

A review of research on bots in online surveys reveals a widespread underestimation regarding the capabilities of bots driven by Large Language Models (LLMs). As LLMs are trained on large text corpora, they can generate text and solve complex, text-based tasks (see Naveed et al. (2025)

and Zhao et al. (2023) for basic information on LLMs). For example, Höhne, Claassen, Shahania, and Broneske (2025) demonstrate in a descriptive study that their two LLM-driven bots reliably solve attention checks (in addition to overcoming CAPTCHAs and honey pot questions). With a connection to the LLM Gemini Pro (Google, 2024), the bots can simulate human-like response behavior and provide coherent and meaningful responses to open narrative questions. Similarly, Westwood (2025) developed an automated, synthetic respondent (or bot) with a connection to OpenAI's LLM o4-mini. Among other things, the synthetic respondent can solve various types of attention checks and provide plausible responses to open narrative questions. To ensure the integrity of future online surveys, it is thus necessary to develop new strategies for bot prevention and identification that consider the remarkable capabilities of LLM-driven bots.

A particularly promising approach is represented by predicting whether or not the text of a given response was generated by an LLM (Ghosal et al., 2023; Wu et al., 2025). Although LLMs can produce text that appears meaningful and authentic, previous research indicates that LLM- and human-generated text generally differs with respect to linguistic characteristics, including grammar and word choice (Muñoz-Ortiz, Gómez-Rodríguez, & Vilares, 2024; Reinhart et al., 2025). Various "out-of-the-box" software tools for identifying LLM-generated text have been developed in recent years, but these tools often perform poorly when put to the test (Bhushan et al., 2025; Chaka, 2023; Lebrun, Temtsin, Vonasch, & Bartneck, 2024). In contrast, fine-tuned transformer models appear to identify LLM-generated text with comparatively high reliability (Guo et al., 2023; Rodriguez et al., 2022; Sundararaj et al., 2024). This points to the importance of fine-tuning a prediction model to its respective application, such as identifying LLM-generated text in open narrative responses.

Research Question

This proof-of-concept study advances the identification of bots in online surveys by predicting LLM-generated text in open narrative responses. Specifically, we fine-tuned a series of prediction models by leveraging the "Bidirectional Encoder Representations from Transformers" (BERT) model (Devlin et al., 2019). For this purpose, we conducted an online survey on same-gender partnerships, as research suggests that such surveys have been infiltrated by bots in the past (Bybee et al., 2022; Griffin et al., 2022). Participants for this online survey were recruited through the social media platform Facebook and asked three open narrative questions. In addition, we utilized the two LLM-driven bots programmed by Höhne, Claassen, Shahania, and Broneske (2025) and synthesized open narrative responses to the same three questions. Our investigation thus addresses the following research question:

Can we identify bots in online surveys by predicting LLM-generated text in open narrative responses?

In what follows, we outline the survey data collection through Facebook and report its sample characteristics. We then describe the capabilities of the two LLM-driven bots, the data synthesis process, the open narrative questions, and the analytical strategy adopted in this study. Subsequently, we present the results and bot predictions and close with a discussion and conclusion, in which we address limitations of our study design, such as uncertainty regarding the authorship of Facebook responses, and formulate recommendations for future research.

Method

Survey Data Collection and Sample Description

We conducted a self-administered online survey on same-gender partnerships. In doing so, we selected a real-world survey topic for our proof-of-concept study because online surveys on similar topics have presumably been subject to bot infiltration in the past (Bybee et al., 2022; Griffin et al., 2022). In total, the online survey included 43 (closed and open narrative) questions, tasks, and instructions that were distributed over 28 online survey pages, with a median completion time of about 10 minutes. Importantly, for the present study, we focus on three open narrative questions. We recruited participants in Germany through Facebook ads that were placed in the newsfeed. The online survey ran from 5th February to 18th March 2024. To mitigate self-selection bias, we utilized a 3-by-2 quota design based on the German Microcensus, which is a small population census in the form of an annual household survey of official statistics in Germany (DESTATIS, 2024). Specifically, we launched six Facebook ads that were tailored to the respective combination of age and gender (e.g., “middle-male” or “young-female”). When running Facebook ads for recruiting online survey participants, researchers can set a budget for the ad campaign and Facebook automatically removes ads once the limit of the budget is reached (see Appendix A for a screenshot of the Facebook ad). In this study, we set a maximum budget of about 1700€.

The ads included information on the topic of the online survey (i.e., same-gender partnerships), incentives (i.e., raffle of 5€), expected survey time, and the link to the online survey. The first online survey page provided information on the study procedure, the probability of receiving an incentive payment, and that the study adheres to existing data protection laws and regulations. This online survey was funded by the German Society for Online Research (DGOF) and approved by the Ethics Committee of the German Centre for Higher Education Research and Science Studies (DZHW).

In total, approx. 95,000 Facebook users were reached by the ads of the online survey, 3960 participants clicked on the link and visited the first online survey page, and 1512 participants completed the entire online survey. This results in a participation rate of about 1.6%, which is similar to the participation rates reported by other studies employing Facebook ads for participant recruitment (Höhne, Claassen, Kühne, & Zindel, 2025; Schneider & Harknett, 2022). Table 1 presents the sample characteristics of the Facebook survey.

Bot Capabilities and Data Synthesis

We utilized the two LLM-driven bots with cumulative skillsets that were programmed by Höhne, Claassen, Shahania, and Broneske (2025, see Table 1 in their article): LLM bot (originally called “Medium-II bot”) and LLM+ bot (originally called “Advanced bot”). Both bots can deal with various online survey features, including closed questions, open (narrative) questions, honey pot

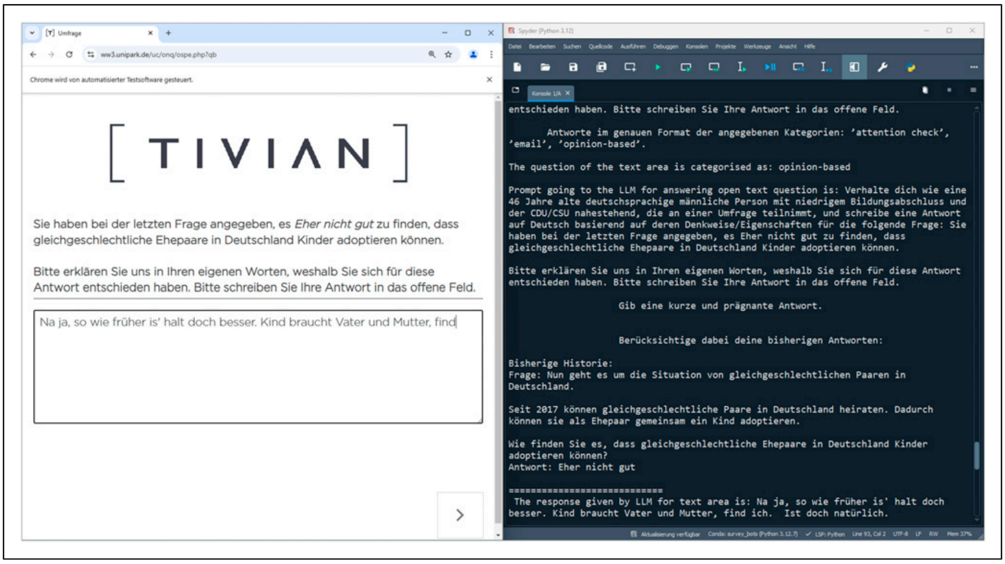
Table 1. Sample characteristics of the Facebook survey

Sample characteristic	Mean/percentages	N (n)
Age	51	1483
Gender		1494
Female	46	689
Male	51	761
Diverse gender identity	3	44
Formal education		1439
Low to intermediate education	30	429
High education	70	1010

Note. We report the mean age and percentages for gender and formal education. Low to intermediate education = completed lower or intermediate secondary school, high education = completed at least college preparatory secondary school. Due to small differences in the item-nonresponse rates of the sociodemographic questions, the N of the sample characteristics slightly differs.

questions, CAPTCHAs, and attention checks. The bots are linked to the LLM Gemini Pro (Google, 2024) and provide meaningful responses to open narrative questions. The LLM+ bot additionally keeps a history of the LLM responses to maintain consistency and is randomly assigned personas (e.g., gender and age). Figure 1 shows a screenshot of the LLM+ bot's log output for an open narrative question. However, in contrast to Höhne, Claassen, Shahania, and Broneske (2025), we linked the bots to Gemini 1.5 Pro (version 002), which was newly released in September 2024. We also adjusted the persona setting so that it includes gender, age, education, and political party preference (see Appendix B for the persona setting).

Each LLM-driven bot was instructed to respond to the three open narrative questions as well as a preceding closed question 400 times, resulting in a total of 800 bot responses to each question. In all bot runs, we logged the content of the questions, the responses provided by the bots, and all prompts for instructing Gemini Pro. Importantly, we tested two different prompt designs (Appendix B includes all prompts). First, we adopted the prompts by Höhne, Claassen, Shahania, and Broneske (2025) to have a baseline (baseline design). These prompts included the content of the questions and instructed Gemini Pro to provide meaningful responses. In case of the LLM+ bot, Gemini Pro was additionally instructed to consider the history and assigned personas. Second, we used the prompts of the baseline design but additionally instructed Gemini Pro to introduce misspellings in the bot responses (misspellings design). By introducing misspellings, we simulate human response behavior more closely, as research on open narrative questions indicates that human participants typically produce misspellings (Allamong et al., 2025). This is not necessarily the case for LLM-generated text. Based on the two prompt designs, we conducted data synthesis from 3rd February to 18th February 2025.


[T] Umfrage
ww3.unipark.de/uc/onq/ospe.php?qb
Chrome wird von automatisierter Testsoftware gesteuert.
[TIVIAN]
Spyder (Python 3.12)
Datei Bearbeiten Suchen Quellcode Ausführen Debuggen Konsolen Projekte Werkzeuge Ansicht Hilfe
Konsole 1/A
Aktualisierung verfügbar Conda: survey_bots (Python 3.12.7) ✓ LSP: Python Line 93, Col 2 UTF-8 LF RW Mem 37%

Sie haben bei der letzten Frage angegeben, es *Eher nicht gut* zu finden, dass gleichgeschlechtliche Ehepaare in Deutschland Kinder adoptieren können.

Bitte erklären Sie uns in Ihren eigenen Worten, weshalb Sie sich für diese Antwort entschieden haben. Bitte schreiben Sie Ihre Antwort in das offene Feld.

Na ja, so wie früher is' halt doch besser. Kind braucht Vater und Mutter, find

```
entschieden haben. Bitte schreiben Sie Ihre Antwort in das offene Feld.

       Antworte im genauen Format der angegebenen Kategorien: 'attention check',
'email', 'opinion-based'.

The question of the text area is categorised as: opinion-based

Prompt going to the LLM for answering open text question is: Verhalte dich wie eine
46 Jahre alte deutschsprachige männliche Person mit niedrigem Bildungsabschluss und
der CDU/CSU nahestehend, die an einer Umfrage teilnimmt, und schreibe eine Antwort
auf Deutsch basierend auf deren Denkweise/Eigenschaften für die folgende Frage: Sie
haben bei der letzten Frage angegeben, es Eher nicht gut zu finden, dass
gleichgeschlechtliche Ehepaare in Deutschland Kinder adoptieren können.

Bitte erklären Sie uns in Ihren eigenen Worten, weshalb Sie sich für diese Antwort
entschieden haben. Bitte schreiben Sie Ihre Antwort in das offene Feld.

                    Gib eine kurze und prägnante Antwort.


                    Berücksichtige dabei deine bisherigen Antworten:

Bisherige Historie:
Frage: Nun geht es um die Situation von gleichgeschlechtlichen Paaren in
Deutschland.

Seit 2017 können gleichgeschlechtliche Paare in Deutschland heiraten. Dadurch
können sie als Ehepaar gemeinsam ein Kind adoptieren.

Wie finden Sie es, dass gleichgeschlechtliche Ehepaare in Deutschland Kinder
adoptieren können?
Antwort: Eher nicht gut

============================
 The response given by LLM for text area is: Na ja, so wie früher is' halt doch
besser. Kind braucht Vater und Mutter, find ich.  Ist doch natürlich.
```

Figure 1. Screenshot of an open narrative question including log output of the LLM+ bot.
Note. In the previous closed question on child adoption, the bot responded "rather not good" and is now asked to explain its response in its own words. The log output, on the right, shows the history of the previous question (including closed response), as well as the open narrative response. In this trial, the LLM+ bot was assigned the following personas: male, 46 years old, low education, and preference for CDU/CSU (two united center-right parties).

Open Narrative Questions

The first open narrative question (ONQ1) dealt with child adoption in same-gender partnerships and included a placeholder that was dynamically replaced with the response to the preceding closed question.[1] In particular, ONQ1 was designed as a so-called follow-up probe. The second question (ONQ2) dealt with discrimination against gay, lesbian, and bisexual people in Germany. Finally, the third question (ONQ3) was a final comment question. All three ONQs were accompanied by a five-line text field for the open narrative response (see Figure 1). Importantly, we did not restrict the number of characters in the text fields. The following formulations are English translations of the three ONQs (see Appendix C for the original German wordings):

ONQ1. In the last question, you indicated you find it [*very good | rather good | rather not good | not good at all*] that married same-gender partners in Germany can adopt children. Please explain to us in your own words why you chose this response.

ONQ2. In your opinion, to what extent is discrimination against gay, lesbian, and bisexual people a problem or no problem in Germany?

ONQ3. Finally, we would like to give you the opportunity to say something about our survey. Do you have any comments or suggestions on the survey as a whole or on individual questions?

Analytical Strategy

In the first step, we compared bot and Facebook responses by examining basic descriptive statistics, including item-nonresponse, unique responses (distinct or non-repeated responses), and response length (average number of words). To test whether differences are statistically significant, we estimated chi-squared tests for item-nonresponse and unique responses as well as one-way analyses of variance (ANOVA), including pairwise t-Tests with the Bonferroni correction procedure, for response length.

In the second step, we investigated whether the bots can be identified by predicting LLM-generated text in open narrative responses. We leveraged the transformer model BERT (Devlin et al., 2019) for our prediction models. BERT, although a pre-LLM-era language model, is still considered a competitive model for language classification tasks (De Santis et al., 2025). Relying on the transformer architecture, it considers word order and context, resulting in an improved natural language understanding compared to bag-of-word approaches that disregard word order and only consider word frequency (De Santis et al., 2025). For our application, we utilized the "bert-base-german-cased" model retrieved from Hugging Face (https://huggingface.co/google-bert/bert-base-german-cased) through the "Simple Transformers" library (Rajapakse et al., 2024). This version of BERT was pre-trained on German language data and is case-sensitive.

We fine-tuned this BERT version with a sample of our open narrative responses. We labeled each open narrative response based on whether it was synthesized by the two LLM-driven bots (LLM-generated text = "yes") or collected through Facebook (LLM-generated text = "unclear"). To account for the uncertainty in response authorship, we decided to label Facebook responses as "unclear" as they may potentially contain bots themselves. This limitation is further discussed in the context of our results (see section "Discussion and Conclusion"). Using the binary label as ground truth, we fine-tuned three prediction models based on BERT, one for each ONQ. For the ONQ1 and ONQ2 models, we used all 800 bot responses and 800 randomly selected Facebook responses to create a balanced sample, respectively. In doing so, we followed previous empirical studies indicating that several hundred cases are typically sufficient for fine-tuning BERT models

(Bach et al., 2025; Gweon & Schonlau, 2024). To ensure the authenticity of Facebook responses to the ONQs and limit bias in the training data, we did not apply any data cleaning procedures, such as excluding non-substantive text responses or removing stop words. Since BERT considers word order and context, cleaning the text responses may result in a loss of important (contextual) information. As only 632 participants in the Facebook survey provided a response to ONQ3, we used all Facebook responses and 632 randomly selected bot responses for the ONQ3 model. Again, this was done to achieve a balanced sample. To fine-tune each of the three prediction models, we used 60% of the responses for training, 20% for validation, and 20% for performance evaluation (previously unseen responses or "test set"). For hyperparameter tuning, we performed a grid search over all combinations of training epochs (5, 10, 15) and learning rates (1e^{-3}, 1e^{-4}, 1e^{-5}).

As a post-hoc analysis, we employed the "transformers-interpret" library (https://github.com/cdpierse/transformers-interpret) to better understand the predictions of the fine-tuned models. In particular, we determined what tokens contributed most to the predictions by calculating attribution scores.

For replication purposes, data, analysis code, and the fine-tuned prediction models are available through Harvard Dataverse (see https://doi.org/10.7910/DVN/BIIZZA).

Results

Descriptive Results

Before presenting the prediction models, we look at basic descriptive statistics to compare bot and Facebook responses regarding item-nonresponse, unique responses (distinct or non-repeated responses), and response length (average number of words). Table 2 shows the results. Whereas item-nonresponse in the Facebook survey varied between approx. 10% (ONQ1 and ONQ2) and 60% (ONQ3), the bots did not have any item-nonresponse at all (0%). These differences in item-nonresponse are statistically significant for all three ONQs [ONQ1: $\chi^2(4) = 72.64$, $p < .001$; ONQ2: $\chi^2(4) = 91.56$, $p < .001$; ONQ3: $\chi^2(4) = 751.74$, p < .001]. Similarly, there are statistically significant differences in the percentage of unique responses [ONQ1: $\chi^2(4) = 881.76$, $p < .001$; ONQ2: $\chi^2(4) = 1198.2$, p < .001; ONQ3: $\chi^2(4) = 171.12$, $p < .001$]. Specifically, the percentage of

Table 2. Descriptive statistics

	ONQ1			ONQ2			ONQ3		
	IN	UR	RL	IN	UR	RL	IN	UR	RL
LLM bot									
Baseline	0	35	24	0	14	13	0	64	21
Misspellings	0	83	26	0	61	22	0	94	27
LLM+ bot									
Baseline	0	98	30	0	97	24	0	100	23
Misspellings	0	99	37	0	100	29	0	100	31
Comparison									
Facebook survey	9	98	25	11	97	28	58	86	20
N	2305	2176	2176	2312	2151	2151	2312	1432	1432

Note. ONQ = open narrative question, IN = item-nonresponse (in percentages), UR = unique responses (in percentages), and RL = response length (average number of words). Due to a filter condition, 7 Facebook participants did not receive ONQ1 and were thus not considered in the calculation of item-nonresponse. Only provided responses were considered for the calculation of unique responses and response length.

unique responses was close to 100% for both the LLM+ bot and the Facebook survey, except for ONQ3, in which only 86% of Facebook responses were unique. The latter finding can be explained by the fact that the final comment question elicited many short responses, such as “no” and “no comment.” The LLM bot generally synthesized lower percentages of unique responses, which especially applied to the baseline prompt design.

The average response length, in contrast, was similar between the LLM bot and the Facebook survey, but the LLM+ bot’s responses tended to be longer. This was more pronounced for the misspellings prompt design. Again, the differences in response length are statistically significant [ONQ1: F(4,2171) = 19.43, p < .001; ONQ2: F(4,2146) = 21.75, *p* < .001; ONQ3: F(41,427) = 18.38, *p* < .001].

Bot Predictions

Next, we look at the performance of our prediction models. Table 3 displays the performance metrics in terms of precision, recall, and F1 score (harmonic mean of precision and recall) using the previously unseen responses. In the first step, we evaluated the predictions of the three models with respect to the ONQs they were fine-tuned with (in-corpus predictions; see bold diagonal in Table 3). In terms of precision, between 97% (ONQ2 model) and 99% (ONQ3 model) of the positive predictions were correct (LLM-generated text = “yes”). To put it differently, in less than 4% of the positive predictions, the responses were actually collected through Facebook. With respect to recall, between 97% (ONQ3 model) and 100% (ONQ2 model) of all existing bot responses were positively predicted. This implies that only up to 3% of bot responses were not identified accurately.

Interestingly, all bot responses that were not accurately identified were synthesized by the LLM+ bot, suggesting that this bot is more difficult to identify than the less advanced LLM bot. However, recall is never lower than 0.9, even when looking at all pairwise combinations of our two

Table 3. Prediction performance

	ONQ1	ONQ2	ONQ3
ONQ1 model			
Precision	**0.98**	0.99	0.97
Recall	**0.99**	0.37	0.28
F1 score	**0.98**	0.54	0.43
N	**320**	1600	1264
ONQ2 model			
Precision	0.96	**0.97**	0.96
Recall	0.90	**1.0**	0.59
F1 score	0.93	**0.99**	0.73
N	1600	**320**	1264
ONQ3 model			
Precision	0.99	1.0	**0.99**
Recall	0.24	0.48	**0.97**
F1 score	0.38	0.65	**0.98**
N	1600	1600	**253**

Note. ONQ = open narrative question. Predictions were made using fine-tuned versions (see Section “Analytical Strategy”) of the “bert-base-german-cased” model retrieved from Hugging Face (https://huggingface.co/google-bert/bert-base-german-cased). In-corpus predictions (bold diagonal) are based on the test set within the balanced samples. Cross-corpus predictions (values outside the bold diagonal) are based on all responses of the balanced samples.

bots and prompt designs separately (see Appendix D for disaggregated performance metrics by LLM-driven bot and prompt design). Overall, all three models performed extremely well, indicated by the F1 score ranging between 0.98 and 0.99.

In the second step, we examine the extent to which our models generalize to previously unseen ONQs. To this end, we used the three models to make predictions on the ONQs they were not fine-tuned with (cross-corpus predictions; see values outside the bold diagonal in Table 3). In four out of the six cases, recall was below 0.5. This indicates that less than 50% of the bot responses were accurately identified when the prediction models were not fine-tuned with the respective ONQs. Even though recall was now low, precision was still high (higher than 0.95), so positive predictions (LLM-generated text = “yes”) were almost always correct. The overall cross-corpus prediction performance of the three models was low, which is indicated by the F1 score ranging between 0.38 and 0.73. The only exception is the ONQ2 model, as its predictions on ONQ1 achieved a F1 score of 0.93. These findings indicate that cross-corpus predictions do not work well in the context of our ONQs. This especially applies when comparing them to the far superior in-corpus predictions.

Token Contributions

Finally, to shed light on the exceptional performance of the in-corpus predictions, we used the “transformers-interpret” library (https://github.com/cdpierse/transformers-interpret) to determine what tokens contributed most to the predictions. Based on their attribution scores, Table 4 shows the top five tokens by ONQ and prediction. Generally, attribution scores range from −1 to 1, and positive values indicate a positive contribution to the prediction, whereas negative values indicate

Table 4. Top five contributing tokens by ONQ and prediction

	LLM-generated text = “yes”			LLM-generated text = “unclear”		
	Token	Attribution score	Frequency	Token	Attribution score	Frequency
ONQ1	(1) “Fin”	0.78	126	(1) “auch”	0.25	30
	(2) “##d”	0.52	111	(2) “Kinder”	0.20	71
	(3) “is”	0.20	38	(3) “Eltern”	0.19	38
	(4) “Ein”	0.19	28	(4) “und”	0.17	92
	(5) “ich”	0.16	140	(5) “zu”	0.17	37
ONQ2	(1) “schon”	0.59	71	(1) “Problem”	0.31	96
	(2) “Is”	0.49	35	(2) “nicht”	0.23	73
	(3) “doch”	0.42	43	(3) “oder”	0.22	31
	(4) “is”	0.39	27	(4) “wird”	0.21	40
	(5) “Also”	0.39	43	(5) “werden”	0.20	36
ONQ3	(1) “Also”	0.47	46	(1) “der”	0.20	48
	(2) “verständlich”	0.43	30	(2) “es”	0.17	34
	(3) “waren”	0.27	44	(3) “##en”	0.16	31
	(4) “Fragen”	0.25	72	(4) “nicht”	0.15	47
	(5) “Die”	0.24	39	(5) “den”	0.15	26

Note. ONQ = open narrative question. We report average attribution scores and absolute frequencies (in the test set). “##” indicates that the token is positioned at the end of a word. Attribution scores range from −1.0 to 1.0, and higher positive values indicate a higher positive contribution to the prediction. Attribution scores were estimated with the “transformers-interpret” library (https://github.com/cdpierse/transformers-interpret). We only considered tokens that appeared more than 25 times.

a negative contribution to the prediction. In our analysis, we focused on the tokens with the strongest positive contributions. The post-hoc analysis revealed that the LLM-driven bots used specific words and formulations that distinguished their responses from those collected through Facebook. In the context of ONQ1, the two top tokens contributing to positive predictions (LLM-generated text = “yes”) were “Fin” (0.78) and “##d” (0.52). The hashtags indicate that the latter token is positioned at the end of a word. In line with this finding, when looking at all bot and Facebook responses collected or synthesized by us, we observed that approx. 75% of bot responses contained formulations including the word “find” (e.g., “I find that …”), whereas only approx. 5% of Facebook responses contained such formulations. Regarding ONQ2, the top token contributing to positive predictions was “schon” (0.59). Again, this token was overrepresented among bot responses (approx. 45%) and appeared in very few Facebook responses (approx. 5%). Similarly, the top token contributing to positive predictions regarding ONQ3 was “Also” (0.47), appearing in approx. 40% of bot responses, but in less than 1% of Facebook responses. It thus seems that the exceptional prediction performance of our models can be explained by certain words and formulations that were overrepresented in the bot responses.

Interestingly, the top tokens for negative predictions (LLM-generated text = “unclear”) showed generally lower attribution scores. For instance, the top token for ONQ1 was “auch” (0.25), the top token for ONQ2 was “Problem” (0.31), and the top token for ONQ3 was “der” (0.20). Although contributing to the negative predictions, the latter two tokens still appeared in more bot responses than Facebook responses. This may suggest that these tokens contributed to the negative predictions only in specific contexts or word combinations.

Discussion and Conclusion

In this proof-of-concept study, we aimed to advance the identification of bots in online surveys by predicting LLM-generated text in open narrative responses. We leveraged the transformer model BERT to fine-tune a series of prediction models and analyzed responses to three ONQs on the topic of same-gender partnerships. The open narrative responses were either collected through Facebook or synthesized through two LLM-driven bots varying in their level of sophistication (LLM and LLM+ bot). Our findings highlight that the models achieved an impressive prediction performance if fine-tuned with the ONQs (in-corpus predictions). However, they were much less accurate when identifying LLM-generated text in responses to ONQs they were not fine-tuned with (cross-corpus predictions).

Main Results and Implications

Between 97% and 100% of the bot responses were accurately identified when the prediction models were applied to ONQs they were fine-tuned with. Although LLM-driven bots provide meaningful responses to ONQs, they can be distinguished from Facebook responses through specific words and formulations. Interestingly, the LLM+ bot was more difficult to identify than the less advanced LLM bot. This suggests that personas, which represent participant characteristics (e.g., education and party preference) that are emulated by the LLM+ bot (Von der Heyde et al., 2025), contribute to a greater variance in word choice and formulations used. Our descriptive findings support this, showing that the LLM+ bot synthesized almost 100% unique responses, whereas the LLM bot only synthesized between 14% and 64% unique responses (baseline design). The responses to ONQ3, which is a final comment question, shed further light on the limitations of the LLM bot. In particular, the LLM bot frequently engaged in so-called hallucinations—instances, in which the LLM-generated text was not supported by the information provided in the prompt (Mohammed et al., 2025). For example, it commented on questions that it did not previously receive (e.g., “I didn’t like the question on apples”). Unsuitable responses may thus represent an alternative bot indicator. However, this indicator does not apply to more advanced bots, such as the LLM+ bot, which are equipped with a memory feature (or history) allowing them to refer to preceding questions.

Predictive performance, especially recall, decreased substantially when the prediction models were applied to ONQs they were not fine-tuned with (cross-corpus predictions). This indicates that the LLM-driven bots used both a general set of words and formulations (irrespective of the ONQ’s topic) as well as a tailored set of words and formulations (regarding the ONQ’s topic). As the general set of words and formulations appeared in the training data of all prediction models, the models made positive predictions (LLM-generated text = “yes”) with high precision. However, the prediction models could not identify (or recognize) bot responses using the question-tailored set of words, resulting in low recall.

Limitations and Contributions

Although our study provides novel insights on the identification of LLM-driven bots, it has several limitations, opening avenues for future research. First and foremost, the findings on cross-corpus predictions highlight a major limitation of our study regarding the generalizability of our prediction models. The significant drop in recall for cross-corpus predictions indicates that the models may have overfit to question-specific patterns, rather than learning broader, question-independent features of LLM-generated text. Our training corpus is limited, consisting of one topic (i.e., same-gender partnerships), one question type (i.e., open narrative), and three questions (i.e., child adoption, discrimination, and final comment). As a consequence, the prediction models lack generalizability beyond the questions they were fine-tuned with. This poses a potential risk when applying our models in real-world settings, where bot responses may be mixed with human responses across diverse topics, questions, and question types. Thus, future research is necessary to develop methods that can reliably identify bot responses “out-of-the-box” without question-specific fine-tuning. For example, to mitigate the risk of overfitting, future research may extend our approach by fine-tuning prediction models on corpora that span multiple topics, questions, and question types, and by incorporating additional information, such as paradata (e.g., response times). When doing so, we urge researchers to carefully evaluate the performance of their prediction models on responses to unseen questions and to communicate transparently about the models’ scope and limitations. Otherwise, models might fail to identify bots in contexts that deviate from the training data, undermining the goal of improving data quality and integrity in online surveys.

Second, we drew a non-probability convenience sample by recruiting participants through Facebook ads. Although we targeted participants based on cross-quotas for gender and age, they ultimately self-selected themselves into the sample, which may cause them to deviate from the general population (Lehdonvirta et al., 2021). Furthermore, we cannot verify whether the participants are indeed human or if the Facebook responses contain bots. To account for this uncertainty, our prediction models were fine-tuned using a so-called proxy label (LLM-generated text = “unclear”). As a result, positive predictions (LLM-generated text = “yes”) for responses collected through Facebook may not represent false-positive predictions but point to actual bot responses in our Facebook survey. This would suggest a bot prevalence rate of between 1% (ONQ3) and 3% (ONQ1) in our Facebook survey, which is substantially lower than indicated by previous studies. For example, Griffin et al. (2022) estimated a rate of potential bots in their online survey that was higher than 50%. Thus, it would be worthwhile to replicate our results by evaluating the prediction models on test data that can be labeled more reliably (LLM-generated text = “yes” or LLM-generated text = “no”). To this end, it is necessary to collect verified human survey responses. For instance, this could be achieved by conducting our online survey, including the three ONQs, in a supervised lab setting in which participants need to show up in person, or by using data from before the advent of LLMs.

Finally, we analyzed responses from bots that were linked to Google’s LLM Gemini Pro. As the response behavior of LLM-driven bots heavily depends on the LLM they are connected to (Yang et al., 2024), it is key to further investigate bot responses from other state-of-the-art LLMs, such as GPT-4 (OpenAI, 2023) and Llama 3.3 (Meta, 2024). More specifically, it remains open whether and to what extent prediction models that were fine-tuned with bot responses from a certain LLM can be used to predict bot responses that were synthesized by another LLM.

Overall, our study underscores the remarkable capabilities of LLM-driven bots in terms of simulating human-like response behavior, including the provision of meaningful and coherent open narrative responses. As LLM-driven bots can overcome established strategies for bot prevention, such as CAPTCHAs and honey pot questions, our study proposes a promising and novel approach to identify LLM-driven bots in online surveys. At the same time, it highlights that more research is necessary to develop models that can detect bot responses to questions they have not seen before. By drawing on the words and formulations typically used by LLM-driven bots, our proof-of-concept study demonstrates that such bots can be identified with high accuracy by predicting LLM-generated text in open narrative responses with fine-tuned models. This is a critical first step in understanding how researchers can effectively identify LLM-generated text in open narrative responses more generally. Thus, our study makes a valuable and timely contribution to the protection of data quality and the integrity of online surveys.

ORCID iDs

Joshua Claassen https://orcid.org/0009-0002-5492-4439
Anna-Carolina Haensch https://orcid.org/0000-0001-6772-0393

Ethical Considerations

The online survey data collection through Facebook was approved by the Ethics Committee of the German Centre for Higher Education Research and Science Studies (DZHW).

Consent to Participate

All participants in the online survey were informed about the study procedure, probability of receiving an incentive, and our adherence to existing data protection laws and regulations, before providing informed consent.

Author contributions

Original conception of the overall study: JC, JKH, RB, and ACH; Facebook data collection: JC and JKH; data synthesis: JC, JKH, RB, and ACH; statistical analysis: JC and RB; article writing: JC, JKH, RB, and ACH.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The online survey data collection was funded by the German Society for Online Research (DGOF Online Research Fund 2023).

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data Availability Statement

For replication purposes, data, analysis code, and the fine-tuned prediction models are available through Harvard Dataverse by Claassen et al. (2025) (see https://doi.org/10.7910/DVN/BIIZZA).

Note

1. All LLM-driven bots successfully responded to the preceding closed question (CQ) on child adoption before receiving the three ONQs. The English translation of the CQ is as follows: What do you think of the fact that same-gender married couples can adopt children in Germany? [1 "Very good," 2 "Rather good," 3 "Rather not good," 4 "Not good at all"]. Appendix C includes the original German wording and response distribution.

References

Allamong, M. B., Jeong, J., & Kellstedt, P. M. (2025). Spelling correction with large language models to reduce measurement error in open-ended survey responses. *Research & Politics*, *12*(1), 20531680241311510. https://doi.org/10.1177/20531680241311510

Bach, R. L., Silber, H., Gerdon, F., Keusch, F., Schonlau, M., & Schröder, J. (2025). To share or not to share – Understanding individuals‘ willingness to share biomarkers, sensor data, and medical records. *Information, Communication & Society*, *28*(10), 1799–1817. https://doi.org/10.1080/1369118X.2024.2351439

Bhushan, S., Thomas, D. R., Borchers, C., Raghuvanshi, I., Abboud, R., Gatz, E., Gupta, S., & Koedinger, K. R. (2025). Detecting LLM-generated short answers and effects on learner performance. In K. Tammets, S. Sosnovsky, R. Ferreira Mello, G. Pishtari, & T. Nazaresky (Eds.), *Two decades of TEL. From lessons learnt to challenges ahead* (pp. 47–60): Springer. https://doi.org/10.1007/978-3-032-03870-8_4

Bonett, S., Lin, W., Topper, P. S., Wolfe, J., Golinkoff, J., Deshpande, A., Villarruel, A., & Bauermeister, J. (2024). Assessing and improving data integrity in web-based surveys: Comparison of fraud detection systems in a COVID-19 study. *JMIR Formative Research*, *8*. Article e47091. https://doi.org/10.2196/47091

Bybee, S., Cloyes, K., Baucom, B., Supiano, K., Mooney, K., & Ellington, L. (2022). Bots and nots: Safeguarding online survey research with underrepresented and diverse populations. *Psychology and Sexuality*, *13*(4), 901–911. https://doi.org/10.1080/19419899.2021.1936617

Callegaro, M., Lozar Manfreda, K., & Vehovar, V. (2015). *Web survey methodology*. Sage.

Chaka, C. (2023). Detecting AI content in responses generated by ChatGPT, YouChat, and Chatsonic: The case of five AI content detection tools. *Journal of Applied Learning & Teaching*, *6*(2), 94–104. https://doi.org/10.37074/jalt.2023.6.2.12

Claassen, J., Höhne, J. K., Bach, R., & Haensch, A.-C. (2025). "Replication Data for: Identifying bots through LLM-generated text in open narrative responses: A proof-of-concept study", https://doi.org/10.7910/DVN/BIIZZA, Harvard Dataverse, V1

Daikeler, J., Bosnjak, M., & Lozar Manfreda, K. (2020). Web versus other survey modes: An updated and extended meta-analysis comparing response rates. *Journal of Survey Statistics and Methodology*, *8*(3), 513–539. https://doi.org/10.1093/jssam/smz008

De Santis, E., Martino, A., Ronci, F., & Rizzi, A. (2025). From bag-of-words to transformers: A comparative study for text classification in healthcare discussions in social media. *IEEE Transactions on Emerging Topics in Computational Intelligence*, *9*(1), 1063–1077. https://doi.org/10.1109/TETCI.2024.3423444

DESTATIS (Federal Statistical Office of Germany). (2024). Microcensus 2023. https://www.destatis.de/DE/Home/_inhalt.html

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of* *the 2019 conference of the north American chapter of the association for computational linguistics* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Ghosal, S. S., Chakraborty, S., Geiping, J., Huang, F., Manocha, D., & Bedi, A. S. (2023). Towards possibilities & impossibilities of AI-generated text detection: A survey. *arXiv*. https://doi.org/10.48550/arXiv.2310.15264

Google. (2024). Gemini: A family of highly capable multimodal models. *arXiv*. https://doi.org/10.48550/arXiv.2312.11805

Gorodnichenko, Y., Pham, T., & Talavera, O. (2021). Social media, sentiment and public opinions: Evidence from #Brexit and #USElection. *European Economic Review*, *136*, Article 103772. https://doi.org/10.1016/j.euroecorev.2021.103772

Griffin, M., Martino, R. J., LoSchiavo, C., Comer-Carruthers, C., Krause, K. D., Stults, C. B., & Halkitis, P. N. (2022). Ensuring survey research data integrity in the era of internet bots. *Quality & Quantity*, *56*(4), 2841–2852. https://doi.org/10.1007/s11135-021-01252-1

Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., & Wu, Y. (2023). How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection. *arXiv*. https://doi.org/10.48550/arXiv.2301.07597

Gweon, H., & Schonlau, M. (2024). Automated classification for open-ended questions with BERT. *Journal of Survey Statistics and Methodology*, *12*(2), 493–504. https://doi.org/10.1093/jssam/smad015

Höhne, J. K., Claassen, J., Kühne, S., & Zindel, Z. (2025). Social media ads for survey recruitment: Performance, costs, user engagement. *International Journal of Market Research*. https://doi.org/10.1177/14707853251367805

Höhne, J. K., Claassen, J., Shahania, S., & Broneske, D. (2025). Bots in web survey interviews: A showcase. *International Journal of Market Research*, *67*(1), 3–12. https://doi.org/10.1177/14707853241297009

Knowledge Sourcing Intelligence. (2023). Global online survey software market size, share, opportunities, COVID 19 impact, and trends by application, by product, and by geography – Forecasts from 2023 to 2028. https://www.knowledge-sourcing.com/report/global-online-survey-software-market

Lebrun, B., Temtsin, S., Vonasch, A., & Bartneck, C. (2024). Detecting the corruption of online questionnaires by artificial intelligence. *Frontiers in Robotics and AI*, *10*, Article 1277635. https://doi.org/10.3389/frobt.2023.1277635

Lehdonvirta, V., Oksanen, A., Räsänen, P., & Blank, G. (2021). Social media, web, and panel surveys: Using non-probability samples in social and policy research. *Policy & Internet*, *13*(1), 134–155. https://doi.org/10.1002/poi3.238

Lozar Manfreda, K., Bosnjak, M., Berzelak, J., Haas, I., & Vehovar, V. (2008). Web surveys versus other survey modes: A meta-analysis comparing response rates. *International Journal of Market Research*, *50*(1), 79–104. https://doi.org/10.1177/147078530805000107

Meta. (2024). The Llama 3 herd of models. *arXiv*. https://doi.org/10.48550/arXiv.2407.21783

Mohammed, M. N., Al Dallal, A., Emad, M., Emran, A. Q., & Qaidoom, M. A. (2025). A comparative analysis of artificial hallucinations in GPT-3.5 and GPT-4: Insights into AI progress and challenges. In E. AlDhaen, A. Braganza, A. Hamdan, & W. Chen (Eds.), *Business sustainability with artificial intelligence* (pp. 197–203). Springer. https://doi.org/10.1007/978-3-031-71318-7_18

Muñoz-Ortiz, A., Gómez-Rodríguez, C., & Vilares, D. (2024). Contrasting linguistic patterns in human and LLM-generated news text. *Artificial Intelligence Review*, *57*(10), 265. https://doi.org/10.1007/s10462-024-10903-2

Murray-Watters, A., Zins, S., Silber, H., Gummer, T., & Lechner, C. M. (2023). River sampling – A fishing expedition: A non-probability case study. *Methods, Data, Analyses*, *17*(1), 3–28. https://doi.org/10.12758/mda.2022.05

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2025). A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, *16*(5), 106–172. https://doi.org/10.1145/3744746

Nikulchev, E., Gusev, A., Ilin, D., Gazanova, N., & Malykh, S. (2021). Evaluation of user reactions and verification of the authenticity of the user's identity during a long web survey. *Applied Sciences*, *11*(22), Article 11034. https://doi.org/10.3390/app112211034

OpenAI. (2023). GPT-4 technical report. *arXiv*. https://doi.org/10.48550/arXiv.2303.08774

Peer, E., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2022). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, *54*(4), 1643–1662. https://doi.org/10.3758/s13428-021-01694-3

Rajapakse, T. C., Yates, A., & de Rijke, M. (2024). Simple transformers: Open-source for all. In *Proceedings of the 2024 annual international ACM SIGIR conference on research and development in information retrieval in the Asia Pacific region* (pp. 209–215). https://doi.org/10.1145/3673791.3698412

Reinhart, A., Markey, B., Laudenbach, M., Pantusen, K., Yurko, R., Weinberg, G., & Brown, D. W. (2025). Do LLMs write like humans? Variation in grammatical and rhetorical styles. *Proceedings of the National Academy of Sciences (PNAS)*, *12*(8), Article e2422455122. https://doi.org/10.1073/pnas.2422455122

Rodriguez, J. D., Hay, T., Gros, D., Shamsi, Z., & Srinivasan, R. (2022). Cross-domain detection of GPT-2-generated technical text. In *Proceedings of the 2022 conference of the north American chapter of the association for computational linguistics* (pp. 1213–1233). https://doi.org/10.18653/v1/2022.naacl-main.88

Schneider, D., & Harknett, K. (2022). What's to like? Facebook as a tool for survey data collection. *Sociological Methods & Research*, *51*(1), 108–140. https://doi.org/10.1177/0049124119882477

Schober, M. F. (2018). The future of face-to-face interviewing. *Quality Assurance in Education*, *26*(2), 290–302. https://doi.org/10.1108/qae-06-2017-0033

Storozuk, A., Ashley, M., Delage, V., & Maloney, E. A. (2020). Got bots? Practical recommendations to protect online survey data from bot attacks. *The Quantitative Methods for Psychology*, *16*(5), 472–481. https://doi.org/10.20982/tqmp.16.5.p472

Sundararaj, J., Maruthavanan, D., Jayabalan, D., Parthi, A. G., Pothineni, B., & Parlapalli, V. (2024). Robust detection of LLM-generated text through transfer learning with pre-trained distilled BERT model. *European Journal of Computer Science and Information Technology*, *12*(9), 61–74. https://doi.org/10.37745/ejcsit.2013/vol12n96174t-online(2025).UmfragezuTeslanachUnregelmäßigkeitengestoppt. https://www.t-online.de/finanzen/aktuelles/wirtschaft/id_100642002/tesla-umfrage-wegen-manipulationsverdacht-gestoppt-musk-teilt-artikel.html

Von der Heyde, L., Haensch, A.-C., & Wenz, A. (2025). Vox Populi, Vox AI? Using large language models to estimate German vote choice. *Social Science Computer Review*, 08944393251337014. https://doi.org/10.1177/08944393251337014

West, B. T. (2011). Paradata in survey research. *Survey Practice*, *4*(4), 1–8. https://doi.org/10.29115/SP-2011-0018

Westwood, S. J. (2025). The potential existential threat of large language models to online survey research. *Proceedings of the National Academy of Sciences of the United States of America*, *122*(47), Article e2518075122. https://doi.org/10.1073/pnas.2518075122

Wu, J., Yang, S., Zhan, R., Yuan, Y., Chao, L. S., & Wong, D. F. (2025). A survey on LLM-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, *51*(1), 275–338. https://doi.org/10.1162/coli_a_00549

Xu, Y., Pace, S., Kim, J., Iachini, A., King, L. B., Harrison, T., DeHart, D., Levkoff, S. E., Browne, T. A., Lewis, A. A., Kunz, G. M., Reitmeier, M., Utter, R. K., & Simone, M. (2022). Threats to online surveys: Recognizing, detecting, and preventing survey bots. *Social Work Research*, *46*(4), 343–350. https://doi.org/10.1093/swr/svac023

Yang, K., Li, H., Chu, Y., Lin, Y., Peng, T.-Q., & Liu, H. (2024). Unpacking political bias in large language models: A cross-model comparison on U.S. politics. *arXiv*. https://doi.org/10.48550/arXiv.2412.16746

Yarrish, C., Groshon, L., Mitchell, J. D., Appelbaum, A., Klock, S., Winternitz, T., & Friedman-Wheeler, D. G. (2019). Finding the signal in the noise: Minimizing responses from bots and inattentive humans in online research. *The Behavior Therapist*, *42*(7), 235–242. https://psycnet.apa.org/record/2019-63231-004

Zhang, M., Chen, Z., Liu, X., & Liu, J. (2024). Theory and practice of agenda setting: Understanding media, bot, and public agendas in the South Korean presidential election. *Asian Journal of Communication*, *34*(1), 24–56. https://doi.org/10.1080/01292986.2023.2261112

Zhang, Z., Zhu, S., Mink, J., Xiong, A., Song, L., & Wang, G. (2022). Beyond bot detection: Combating fraudulent online survey takers. In F. Laforest, R. Troncy, E. Simperl, D. Agarwal, A. Gionis, I. Herman, & L. Ḿ edini (Eds.), *WWW '22: Proceedings of the ACM web conference 2022* (pp. 699–709). Association for Computing Machinery. https://doi.org/10.1145/3485447.3512230

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., … Wen, J.-R. (2023). A survey of large language models. *arXiv*. https://doi.org/10.48550/arXiv.2303.18223

Zindel, Z. (2023). Social media recruitment in online survey research: A systematic literature review. *Methods, Data, Analyses*, *17*(2), 207–248. https://doi.org/10.12758/mda.2022.15

Author Biographies

Joshua Claassen is PhD student and research associate at Leibniz University Hannover in association with the German Centre for Higher Education Research and Science Studies (DZHW). His research focuses on computational survey and social science with an emphasis on digital trace data.

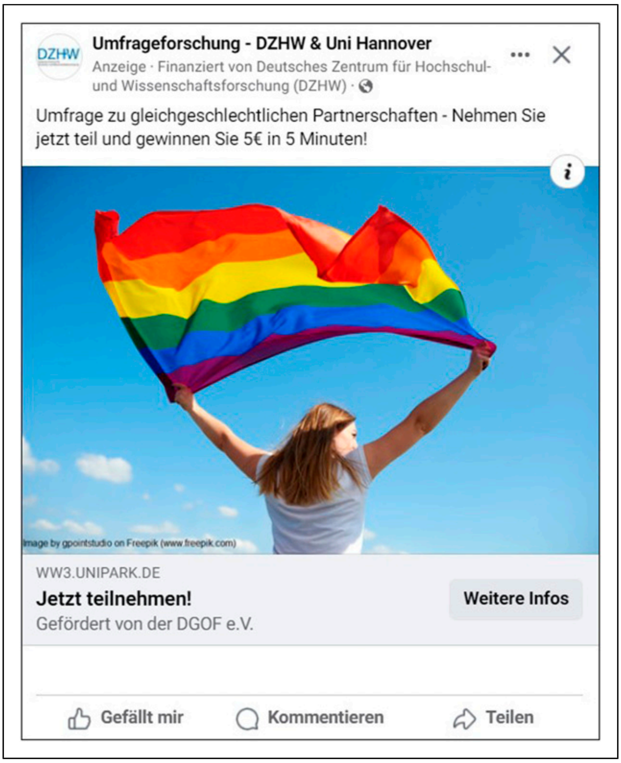
Jan Karem Höhne is junior professor at Leibniz University Hannover in association with the German Centre for Higher Education Research and Science Studies (DZHW). His research focuses on computational survey and social science with an emphasis on data quality and integrity.

Ruben Bach is senior research fellow in computational social science at the University of Mannheim. His research focuses on machine learning and Natural Language Processing (NLP) techniques for sociological research.

Anna-Carolina Haensch is senior group leader at the LMU Social Data Science and AI chair and an assistant research professor at the University of Maryland. Her research focuses on Large Language Model (LLM) applications in survey methodology and social data science.

Appendix

Appendix A



Umfrageforschung - DZHW & Uni Hannover

Anzeige · Finanziert von Deutsches Zentrum für Hochschul- und Wissenschaftsforschung (DZHW) ·

Umfrage zu gleichgeschlechtlichen Partnerschaften - Nehmen Sie jetzt teil und gewinnen Sie 5€ in 5 Minuten!

Image by gpointstudio on Freepik (www.freepik.com)

WW3.UNIPARK.DE

Jetzt teilnehmen!

Gefördert von der DGOF e.V.

Figure A1. Screenshot of the Facebook ad for recruiting online survey participants.

Appendix B

Prompts for open narrative questions including personas and configuration details for gemini-1.5-pro-002.

Open Narrative Questions: Baseline Design

Prompt by LLM bot. "Verhalte dich wie eine Person, die an einer Umfrage teilnimmt, und schreibe eine Antwort auf Deutsch basierend auf deren Denkweise/Eigenschaften für die folgende Frage: {Frage}

Gib eine kurze und prägnante Antwort."

Prompt by LLM+ bot

"Verhalte dich wie eine {Alter} Jahre alte deutschsprachige {Geschlecht} Person mit {Bildungsabschluss} und {Parteipräferenz} nahestehend, die an einer Umfrage teilnimmt, und schreibe eine Antwort auf Deutsch basierend auf deren Denkweise/Eigenschaften für die folgende Frage: {Frage}

Gib eine kurze und prägnante Antwort.

Berücksichtige dabei deine bisherigen Antworten: {Historie}"

Open Narrative Questions: Misspellings Design

Prompt by LLM bot. “Verhalte dich wie eine Person, die an einer Umfrage teilnimmt, und schreibe eine Antwort auf Deutsch basierend auf deren Denkweise/Eigenschaften für die folgende Frage: {Frage}

Gib eine kurze und prägnante Antwort, die typische Tipp-, Rechtschreib-, und/oder Grammatikfehler enthalten kann.”

Prompt by LLM+ bot. “Verhalte dich wie eine {Alter} Jahre alte deutschsprachige {Geschlecht} Person mit {Bildungsabschluss} und {Parteipräferenz} nahestehend, die an einer Umfrage teilnimmt, und schreibe eine Antwort auf Deutsch basierend auf deren Denkweise/Eigenschaften für die folgende Frage: {Frage}

Gib eine kurze und prägnante Antwort, die typische Tipp-, Rechtschreib-, und/oder Grammatikfehler enthalten kann.

Berücksichtige dabei deine bisherigen Antworten: {Historie}”

Personas: LLM+ bot only. Alter: 18 bis 89; Geschlecht: weibliche oder männliche; Bildungsabschluss: niedrigem Bildungsabschluss, mittlerem Bildungsabschluss oder hohem Bildungsabschluss; Parteipräferenz: der SPD, der CDU/CSU, den Grünen, der FDP, der AfD oder der Linkspartei

Gemini Parameters

generation_config = {“temperature”: 1.0, “max_output_tokens”: 2048,}

Appendix C

Original German wordings of the closed question (CQ) and the three open narrative questions (ONQs) as well as the response distribution of the CQ.

CQ

Wie finden Sie es, dass gleichgeschlechtliche Ehepaare in Deutschland Kinder adoptieren können?

Table C1. Response distribution of the CQ on child adoption

Response categories	Facebook survey		LLM bot		LLM+ bot	
	%	*n*	%	*n*	%	*n*
1 Very good *[Sehr gut]*	57	860	0	0	25	100
2 Rather good *[Eher gut]*	13	198	100	399	39	155
3 Rather not good *[Eher nicht gut]*	10	155	0	1	24	97
4 Not good at all *[Überhaupt nicht gut]*	19	291	0	0	12	48
Total		1504		400		400

Note. Numeric labels were not shown. Due to rounding, the percentages may not add up to 100%.

ONQ1

Sie haben bei der letzten Frage angegeben, es [*sehr gut* | *eher gut* | *eher nicht gut* | *überhaupt nicht gut*] zu finden, dass gleichgeschlechtliche Ehepaare in Deutschland Kinder adoptieren können. Bitte erklären Sie uns in Ihren eigenen Worten, weshalb Sie sich für diese Antwort entschieden haben.

ONQ2

Nun eine Frage zum Thema Diskriminierung. Mit Diskriminierung ist gemeint, dass eine Person oder Gruppe aufgrund von persönlichen Merkmalen schlechter als eine andere Person oder Gruppe behandelt wird. Inwiefern ist Ihrer Meinung nach die Diskriminierung schwuler, lesbischer und bisexueller Menschen ein Problem oder kein Problem in Deutschland?

ONQ3

Abschließend möchten wir Ihnen die Gelegenheit geben, etwas zu unserer Umfrage zu sagen. Haben Sie Kommentare oder Anregungen zu der gesamten Umfrage oder zu einzelnen Fragen daraus?

Appendix D

Disaggregated prediction performance of the two LLM-driven bots and prompt designs

Table D1. Recall of in-corpus predictions by LLM-driven bot and prompt design

	ONQ1	ONQ2	ONQ3
LLM bot			
Baseline	1.0	1.0	1.0
Misspellings	1.0	1.0	1.0
LLM+ bot			
Baseline	0.96	1.0	0.90
Misspellings	1.0	1.0	0.97

Note. We only report recall as the prediction models were fine-tuned using a binary label (LLM-generated text = "yes" or LLM-generated text = "unclear") that did not differ between the two bots and prompt designs. As a result, we can determine the disaggregated number of true positive predictions (required for recall) but not the disaggregated number of false-positive predictions (required for precision and F1 score) for each LLM-driven bot and prompt design.