Identifying bots through LLM-generated text in open narrative responses: A proof-of-concept study

Joshua Claassen
German Centre for Higher Education Research and Science Studies (DZHW)
Leibniz University Hannover

Jan Karem Höhne
German Centre for Higher Education Research and Science Studies (DZHW)
Leibniz University Hannover

Ruben Bach
University of Mannheim

Anna-Carolina Haensch
Ludwig Maximilians University Munich

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 equal gender partnership, including three open narrative questions, and recruited 1,512 participants through Facebook. In addition, we utilized two LLM-driven bots that each ran through our online survey 400 times. Each open narrative response is labeled based on whether it was synthesized by our bots (LLM-generated text = "yes") or collected through Facebook (LLM-generated text = "unclear"). Using this binary label as ground truth, we fine-tuned prediction models relying on the transformer model BERT, 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 on which they were not fine-tuned. Our study significantly contributes to the ongoing discussion on bots in online surveys 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

Introduction and research question

Online surveys have increasingly replaced traditional survey modes, especially face-to-face interviews (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 response rates in online surveys are approximately 12% lower than those in other survey modes (see also Lozar Manfreda et al., 2008).

Given the challenges of low response rates in online surveys, researchers are exploring alternative methods for recruiting participants, such as social media platforms, opt-in online access panels, crowdsourcing platforms, and river sampling approaches (Lehdonvirta et al., 2021; Zindel, 2023). While 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 et al., 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 Brexit referendum (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).

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 bot infiltration (t-online, 2025). 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 (Bonnet et al., 2024). Furthermore, the analysis of paradata, 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). In their descriptive study, Höhne et al. (2025) demonstrate that modern LLM-driven bots can perform tasks of great complexity when interacting with online surveys. For instance, their two LLM-driven bots reliably solve CAPTCHAs and automatically skip 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. 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.

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 transformer model BERT (Devlin et al., 2019). For this purpose, we conducted an online survey on equal 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 et al. (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 that is accompanied by recommendations for future research.

Method

Survey data collection and sample description

We conducted a self-administered online survey[1] on equal gender partnerships that included 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. We utilized a 3-by-2 quota design based on the German Microcensus (DESTATIS, 2024) by launching six Facebook ads that were tailored to the respective combination of age and gender (e.g., "middle-male" or "young-female").

The ads included information on the topic of the online survey (i.e., equal gender partnerships), expected survey time (i.e., approximately 5 minutes), incentives (i.e., raffle of 5€), and the link to the online survey. The first online survey page provided information on the study procedure, the likelihood 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, approximately 95,000 Facebook users saw the ads of the online survey, 3,960 participants clicked on the link and visited the first online survey page, and 1,512 participants completed the entire online survey. These participants were between 19 and 95 years old, with a mean age of 51 years, and 48% of them were female. Regarding formal education, 30% had completed lower or intermediate secondary school (low to intermediate education), and 70% had completed at least college preparatory secondary school (high education).

Bot capabilities and data synthesis

We utilized the two LLM-driven bots with cumulative skillsets that were programmed by Höhne et al. (2025, see Table 1): LLM bot (originally called "Medium-II bot") and LLM+ bot

[1] See Höhne et al. (2025) for detailed information on the programming and technical set-up of the online survey.

(originally called "Advanced bot"). Both bots can deal with various online survey features, including closed questions, open (narrative) questions, honey pot 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 et al. (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 A for the persona setting).

Each LLM-driven bot was instructed to respond to the three open narrative questions 400 times, resulting in a total of 800 bot responses to each open narrative 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 A includes all prompts). First, we adopted the prompts by Höhne et al. (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 respondents 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.

Open narrative questions

The first open narrative question (ONQ1) dealt with child adoption in equal gender partnerships and included a placeholder that was dynamically replaced with the response to the preceding closed question[2]. 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 placed at the end of the online survey. 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 B 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?

[2] All LLM-driven bots successfully answered 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-sex married couples can adopt children in Germany? [1 "Very good", 2 "Rather good", 3 "Rather not good", 4 "Not good at all"]. Appendix B includes the original German wording and response distribution.

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?

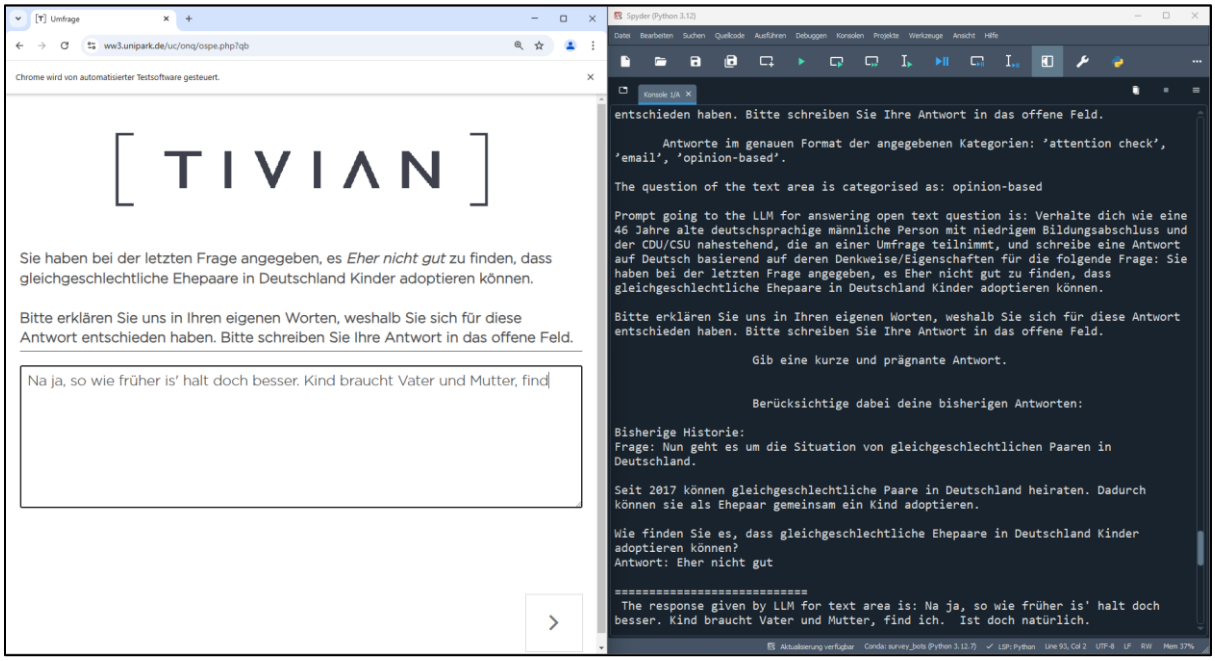

Figure 1. Screenshot of an open narrative question including log output of the LLM+ bot

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).

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).

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 (bidirectional encoder representations from transformers; 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 (Gweon & Schonlau, 2024). For our application, we utilized the "bert-base-german-cased" model retrieved from Hugging Face (https://huggingface.co/google-bert/bert-base-german-cased). This version of BERT was pre-trained on German language data and is case-sensitive.

We fine-tuned this BERT version on 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"). Using this binary label as ground truth, we trained 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. 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.

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 1 shows the results. While item-nonresponse in the Facebook survey varied between approximately 10% (ONQ1 and ONQ2) and 60% (ONQ3), the bots did not have any item-nonresponse at all (0%). This finding aligns with the findings reported by Höhne et al. (2025). The percentage of unique responses was close to 100% for both the LLM+ bot and the Facebook survey, except for ONQ3, in which only 89% of Facebook responses were unique. The LLM bot 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, while the LLM+ bot's responses tended to be longer. This was more pronounced for the misspellings prompt design.

Table 1. Descriptive statistics

	ONQ1			ONQ2			ONQ3		
	IN	UR	RL	IN	UR	RL	IN	UR	RL
LLM bot									
Baseline	0	45	24	0	25	13	0	80	21
Misspellings	0	91	26	0	76	22	0	97	27
LLM+ bot									
Baseline	0	99	30	0	99	24	0	100	23
Misspellings	0	100	37	0	100	29	0	100	31
Comparison									
Facebook survey	9	99	25	11	98	28	58	89	20

Note. IN = Item-nonresponse (in percentages), UR = unique responses (in percentages), RL = response length (average number of words).

Bot predictions

Next, we look at the performance of our prediction models. Table 2 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 on which they were fine-tuned (in-corpus predictions; see

bold diagonal in Table 2). 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 bots and prompt designs separately (see Appendix C 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.

Table 2. 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
ONQ2 model			
Precision	0.96	**0.97**	0.96
Recall	0.90	**1.0**	0.59
F1 score	0.93	**0.99**	0.73
ONQ3 model			
Precision	0.99	1.0	**0.99**
Recall	0.24	0.48	**0.97**
F1 score	0.38	0.65	**0.98**

Note. 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). Within-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.

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 on which they were not fine-tuned (cross-corpus predictions; see values outside the bold diagonal in Table 2). 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 on 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 3 shows the top five tokens by ONQ and prediction. Attribution scores range from -1 to 1, and higher values indicate a higher contribution to the predictions. 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, we observed that 75% of bot responses contained formulations including the word "find" (e.g., "I find that …"), while only 6% 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 (47%) and appeared in very few Facebook responses (6%). Similarly, the top token contributing to positive predictions regarding ONQ3 was "Also" (0.47), appearing in 43% of bot responses, but only in 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 three 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 in an online survey on equal gender partnerships. The open narrative responses were either collected through Facebook or synthesized via two LLM-driven bots varying in their level of sophistication (LLM and LLM+ bot). Our findings highlight that the models achieve an impressive prediction performance if they are fine-tuned on the ONQs (in-corpus predictions).

More specifically, between 97% and 100% of the bot responses were accurately identified. 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, suggesting that personas, such as education and party preference, 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, while the LLM bot only synthesized between 25%

Table 3. 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"	0.16	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. 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 values indicate a higher contribution to the prediction. Attribution scores were estimated with the "transformers-interpret" library retrieved from GitHub (https://github.com/cdpierse/transformers-interpret). We only considered tokens that appeared more than 25 times.

.

and 80% unique responses. The responses to ONQ3, which is a final comment question positioned at the end of the online survey, shed further light on the limitations of the LLM bot. In particular, the LLM bot frequently engaged in so-called hallucinations (Mohammed et al., 2025) and commented on questions that were not part of the online survey (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.

While still achieving high precision, recall decreased substantially when the prediction models were applied to ONQs on which they were not fine-tuned (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, bot responses using the question-tailored set of words and formulations could not be identified (or recognized) by the prediction models, resulting in low recall. These findings highlight the importance of fine-tuning bot prediction models on a broad set of questions and topics, so that they can identify bots robustly and irrespective of the survey (questionnaire).

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, we analyzed a comparatively small corpus consisting of one topic (i.e., equal gender partnerships), one question type (i.e., open narrative), and three questions (i.e., child adoption, discrimination, and final comment). We therefore encourage future studies to go beyond our proof-of-concept study by, for instance, investigating the performance of bot prediction models that are based on more topics and questions. In doing so, these studies can build on our fine-tuned prediction models. In addition, future studies may incorporate paradata, such as mouse movements and keystrokes, in prediction models. Second, our prediction models were fine-tuned on a so-called proxy label (LLM-generated text = "unclear") as the Facebook responses themselves may have contained bots. 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, 2023). More specifically, it remains open whether and to what extent

prediction models that were fine-tuned on 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. 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. Thus, our study makes a valuable and timely contribution to the protection of data quality and integrity of online surveys.

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 and Politics*, 12(1). https://doi.org/10.1177/20531680241311510

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 & Sexuality*, 13(4), 901-911. https://doi.org/10.1080/19419899.2021.1936617

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

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 and Quantity*, 56(4), 2841–2852. https://doi.org/10.1007/s11135-021-01252-1

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

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

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, 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. *Journal of the Market Research Society*, 50, 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 et al. (Eds.), *Business Sustainability with Artificial Intelligence* (pp. 197-203). Springer. https://doi.org/10.1007/978-3-031-71318-7_18

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

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

t-online (2025). Umfrage zu Tesla nach Unregelmäßigkeiten gestoppt. https://www.t-online.de/finanzen/aktuelles/wirtschaft/id_100642002/tesla-umfrage-wegen-manipulationsverdacht-gestoppt-musk-teilt-artikel.html

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.

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. M´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

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

Appendix A

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: {question}
Gib eine kurze und prägnante Antwort."

Prompt by LLM+ bot:

"Verhalte dich wie eine {age} Jahre alte deutschsprachige {gender} Person mit {education} und {party preference} nahestehend, die an einer Umfrage teilnimmt, und schreibe eine Antwort auf Deutsch basierend auf deren Denkweise/Eigenschaften für die folgende Frage: {question}
Gib eine kurze und prägnante Antwort.
Berücksichtige dabei deine bisherigen Antworten: {history}"

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: {question}
Gib eine kurze und prägnante Antwort, die typische Tipp-, Rechtschreib-, und/oder Grammatikfehler enthalten kann. "

Prompt by LLM+ bot:

"Verhalte dich wie eine {age} Jahre alte deutschsprachige {gender} Person mit {education} und {party preference} nahestehend, die an einer Umfrage teilnimmt, und schreibe eine Antwort auf Deutsch basierend auf deren Denkweise/Eigenschaften für die folgende Frage: {question}
Gib eine kurze und prägnante Antwort, die typische Tipp-, Rechtschreib-, und/oder Grammatikfehler enthalten kann.
Berücksichtige dabei deine bisherigen Antworten: {history}"

Personas – LLM+ bot only

Age: 18 to 89 years
Gender: female or male
Education: low education, medium education, or high education
Party preference: SPD, CDU/CSU, Greens, FDP, AfD, or Left

Gemini parameters

generation_config = {"temperature": 1.0, "max_output_tokens": 2048,}.

Appendix B

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 B1. 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 percent.

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 C

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

Table C1. 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 on a binary label (LLM-generated text = "yes" or LLM-generated text = "unclear") and therefore did not differentiate 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.