Identifying bots through LLM-generated text in open narrative questions on LGBTQ+-related topics: A proof-of-concept study

Höhne[1], Claassen[1], Bach[2], & Haensch[3]

[1] DZHW, Leibniz University Hannover

[2] University of Mannheim

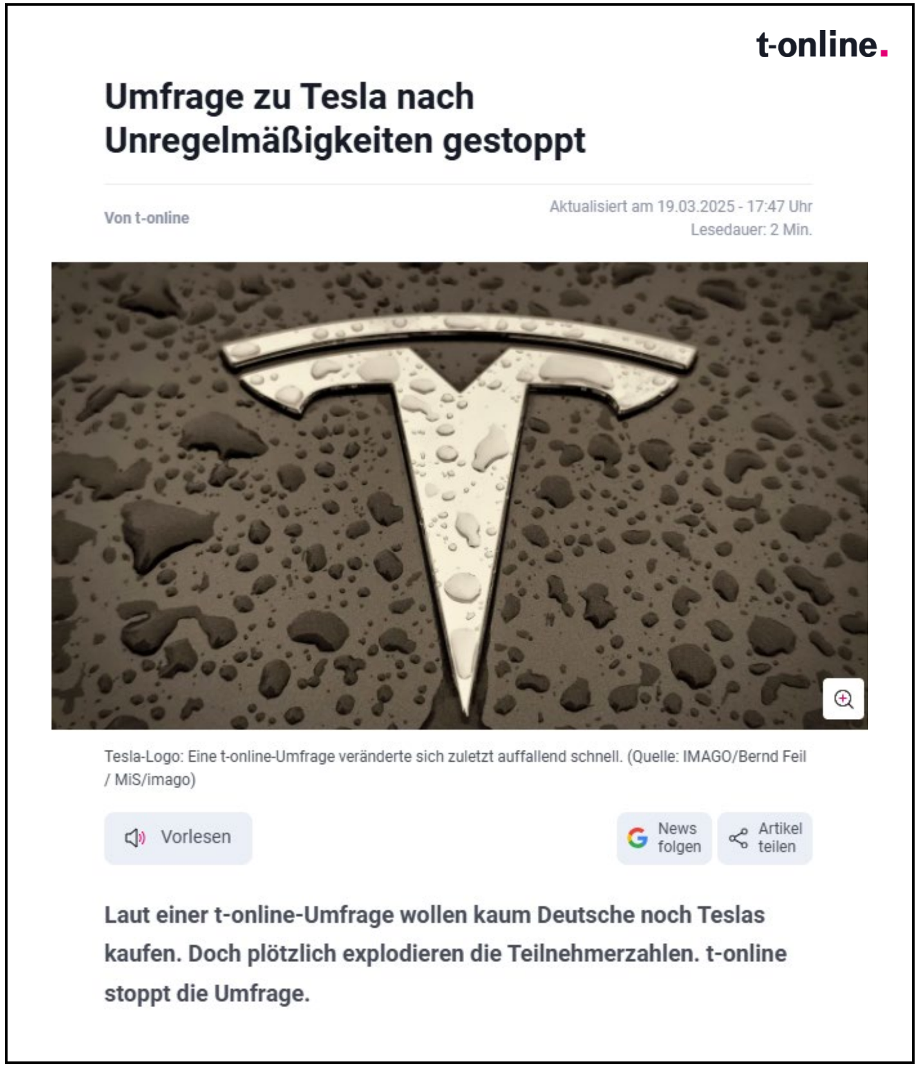
[3] LMU Munich

Frühjahrstagung der DGS-Sektion "Methoden der empirischen Sozialforschung"

Bonn (Germany) – March 6, 2026

This research is funded by the
German Society for Online Research

DGOF



t-online.

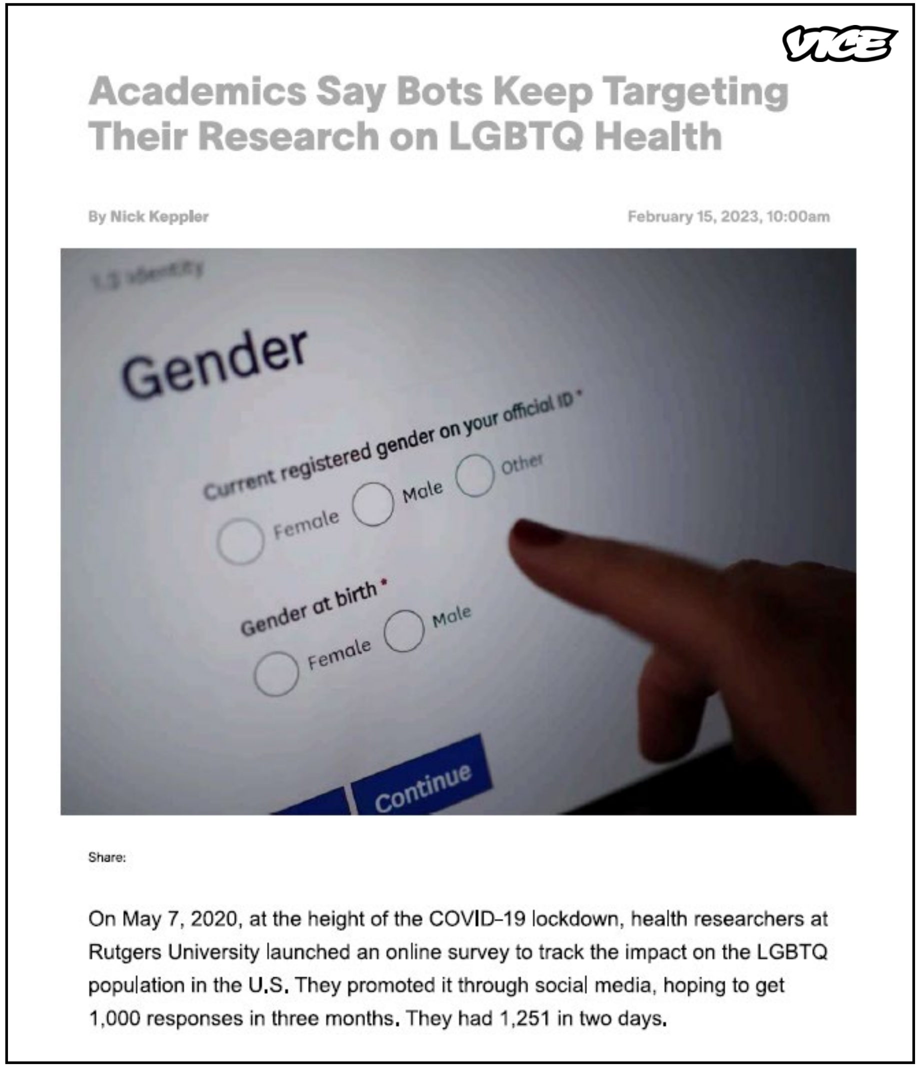
Umfrage zu Tesla nach Unregelmäßigkeiten gestoppt

Von t-online

Aktualisiert am 19.03.2025 - 17:47 Uhr
Lesedauer: 2 Min.

Tesla-Logo: Eine t-online-Umfrage veränderte sich zuletzt auffallend schnell. (Quelle: IMAGO/Bernd Feil / MiS/imago)

Laut einer t-online-Umfrage wollen kaum Deutsche noch Teslas kaufen. Doch plötzlich explodieren die Teilnehmerzahlen. t-online stoppt die Umfrage.



VICE

Academics Say Bots Keep Targeting Their Research on LGBTQ Health

By Nick Keppler

February 15, 2023, 10:00am

On May 7, 2020, at the height of the COVID-19 lockdown, health researchers at Rutgers University launched an online survey to track the impact on the LGBTQ population in the U.S. They promoted it through social media, hoping to get 1,000 responses in three months. They had 1,251 in two days.

Introduction I

- Web surveys struggle with increasingly low response rates (Daikeler et al. 2020)

- Respondents are frequently recruited through nonprobability sampling
  - *Social media platforms, online panels, crowdsourcing platforms, and river sampling*
  - *Quick and easy access to diverse respondent pools* (Lehdonvirta et al. 2021; Zindel 2022)

- However, data integrity are potentially threatened by bots (Griffin et al. 2022; Höhne et al. 2024; Storozuk et al. 2020; Xu et al. 2022; Yarrish et al. 2019; Zhang et al. 2022)
  - *Programs that autonomously interact with digital systems, such as web surveys*
  - *Bots may change survey outcomes and thus political and social decision-making*

- Bots were already used to manipulate public opinion through social media
  - *For example, during Brexit-Referendum in 2016* (Gorodnichenko et al. 2021)

Introduction II

- Research on how to prevent bots from infiltrating web surveys is scarce (Griffin et al. 2022; Höhne et al. 2024; Storozuk et al. 2020; Xu et al. 2022; Yarrish et al. 2019; Zhang et al. 2022)
  - *Methods preventing bots from entering web surveys (e.g., CAPTCHAs)*
  - *Analyzing answer behavior (e.g., open narrative answers)*
  - *Analyzing completion behavior (e.g., response times)*

- Previous studies underestimate the capabilities of advanced LLM-driven bots (Höhne et al. 2024)
  - *LLM-driven bots overcome CAPTCHAs, solve attention checks, and skip honey pots*
  - *Simulate human-like completion behavior and provide coherent open narrative answers*
  - *Established bot detection strategies are not effective anymore*

!! LLM-driven bots require new strategies for bot detection !!

Research Questions (RQs)

- **RQ1**: What are the characteristics of open narrative answers in web surveys provided by LLM-driven bots?

- **RQ2**: Can we detect LLM-driven bots in web surveys by predicting robotic language in open narrative answers?

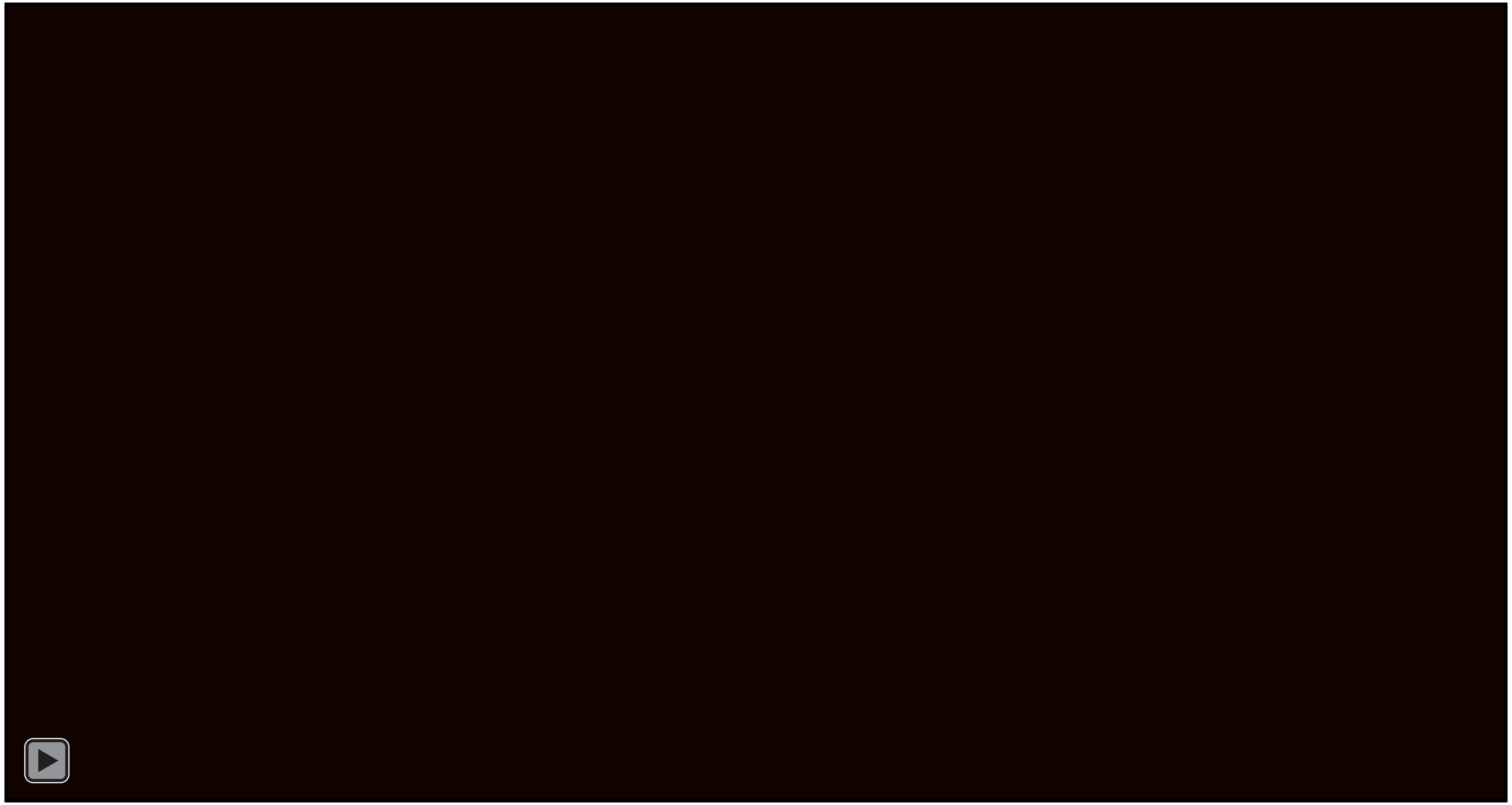
Method: Bot Development

- We utilize two LLM-driven bots: LLM & LLM+ bot (Höhne et al. 2024)
  - *Linked to LLM Gemini 1.5 Pro* (Google 2024)

- Each bot comes with two different prompt designs
  - *Prompts adopted from Höhne et al. (2024): **baseline design***
  - *Baseline design + instruction to introduce misspellings (**misspelling design**)*

LLM bot	**LLM+ bot** (inherits LLM bot skills)
+ Classifies web survey content into opinion-based, emails, and attention checks using LLM + Overcomes CAPTCHAs, attention checks, and honey pots + **Uses LLM to understand and answer questions meaningfully** + Reads questions and mimics human time delay	+ **Remembers previous answers (memory)** + **Answers based on respondent characteristics (personas)** + Handles questions with audio-visual content (speech-to-text) + Simulates paradata (mouse movements and clicks, scrolling, and keystrokes)

Method: Bot Showcase



Method: Data and Analyses

- Web survey on same-gender partnerships programmed with Unipark
  - *Three open narrative questions: Child adoption, discrimination, and final comment*
  - *Each bot took the web survey 400 times (N = 800) in February 2025*
  - *We conducted a web survey through Facebook (N = 1,512) in February/March 2024*

- Each answer was labeled based on whether it was …
  - *… generated by a bot (**LLM-generated text = "yes"**)*
  - *… obtained through the Facebook survey (**LLM-generated text = "unclear"**)*

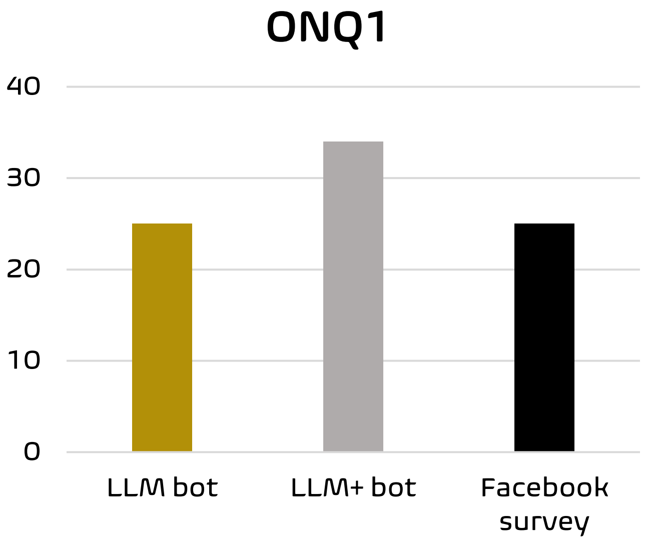
- RQ1: Text-as-data methods in the form of answer length and word choice

- RQ2: Predicting robotic language
  - *Fine-tuning BERT for each ONQ, using the dichotomous label as ground truth*
  - *Performance evaluation: Precision, recall, and F1 score*

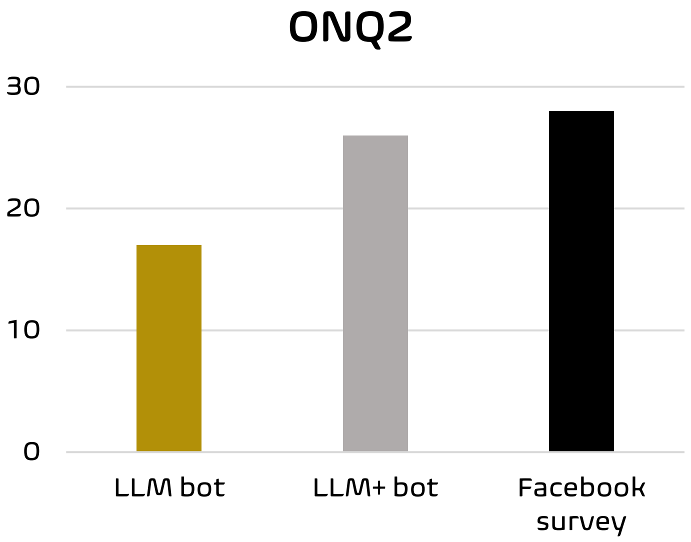
Results: Exemplary Answers

LLM bot	LLM+ bot	Facebook survey
Jeder sollte die gleichen Chancen haben, eine Familie zu gründen. Liebe ist Liebe. ***Translation:*** *Everyone should have the same opportunities to start a family. Love is love.*	Ein Kind braucht ’ne Mutter und ’nen Vater. So is das nun mal vorgesehen. ***Translation:*** *A child needs a mother and a father. That's how it's meant to be.*	Hauptsache es wird sich gut um das Kind gekümmert. ***Translation:*** *The most important thing is that the child is well taken care of.*

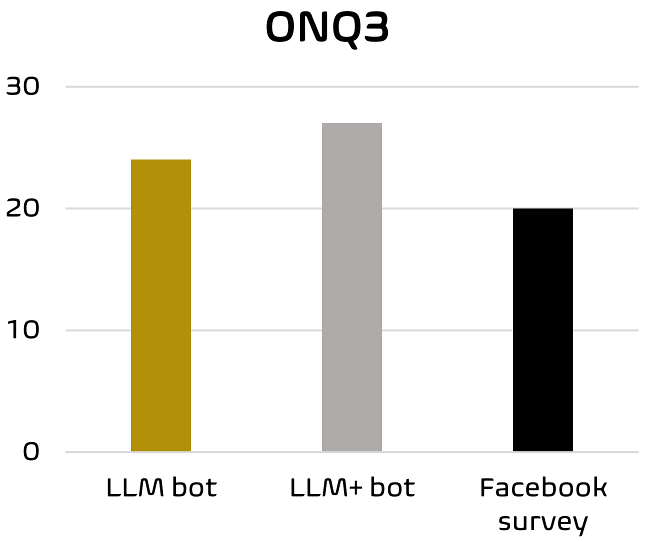
Results: Answer Length (RQ1)



Note. Average number of words.
One-way ANOVA: p < 0.001.

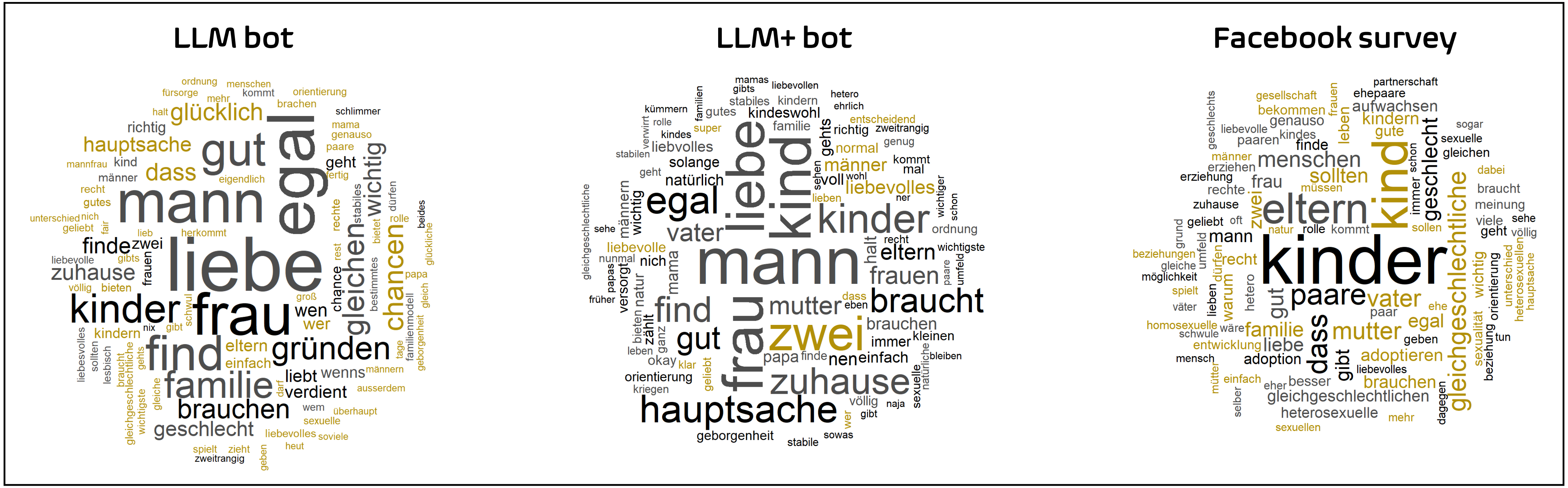


Note. Average number of words.
One-way ANOVA: p < 0.001.



Note. Average number of words.
One-way ANOVA: p < 0.001.

Results: Word Choice (RQ1)



LLM bot

LLM+ bot

Facebook survey

Note. Each word cloud contains the 100 most frequently mentioned words (ONQ1) among the LLM bot, LLM+ bot, and Facebook survey, respectively. The size of a word is proportional to its frequency.

Results: Predicting Robotic Language (RQ2)

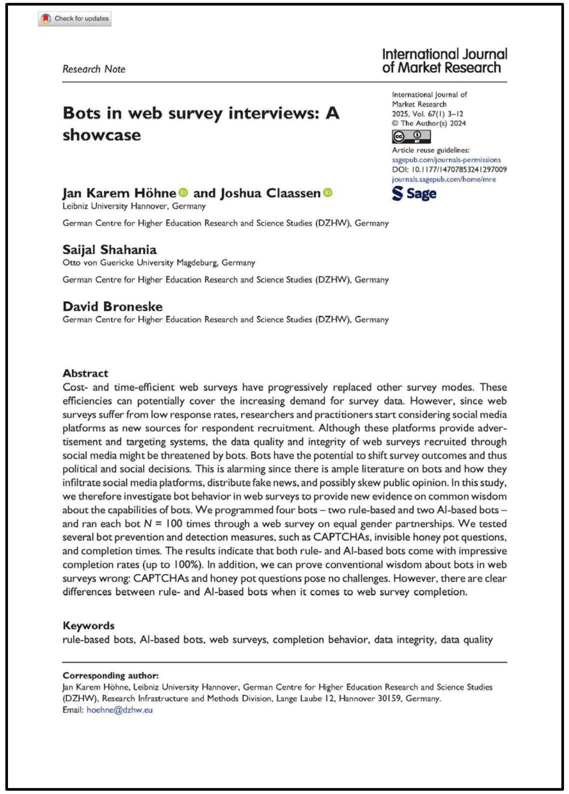
Table 1. Prediction performance aggregated for bots and prompt designs

	ONQ1	ONQ2	ONQ3
Training set size (60%)	960	960	758
Validation set size (20%)	320	320	253
Test set size (20%)	320	320	253
Precision	0.98	0.97	0.99
Recall	0.99	1.0	0.97
F1 score	0.98	0.99	0.98

Note. We used the "bert-base-german-cased" model via the "Simple Transformers" library in Python. For ONQ1 and ONQ2, we used all 800 bot answers as well as 800 randomly selected Facebook survey answers, respectively, to create a balanced sample. For ONQ3, in contrast, we used all 632 Facebook survey answers as well as 632 randomly selected bot answers.

Discussion and Conclusion

- There are similarities between LLM-driven bots and the Facebook survey
  - *LLM-driven bots provide meaningful open narrative answers*
  - *No systematic differences regarding answer length*
  - *Word choice may offer clues when it comes to detecting LLM-driven bots*
- BERT reliably predicts robotic language in open narrative answers
  - *Between 97 and 100 percent of LLM-driven bots are correctly detected*
  - *Applies to both bots (LLM and LLM+) and prompt designs (baseline and misspelling)*
- In a next step, we explore further possibilities regarding bot detection
  - *Predicting robotic language in ONQs that BERT was not fine-tuned with*
  - *Using BERT to predict prevalence of LLM-driven bots in web survey data*
  - *Making predictions based on closed questions*
  - *Examining bots that are connected to other LLMs, such as GPT-4 and Llama 3.3*



Check for updates

Research Note

International Journal of Market Research

International Journal of
Market Research
2025, Vol. 67(1) 3–12
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/14707853241297009
journals.sagepub.com/home/mre
Sage

Bots in web survey interviews: A showcase

Jan Karem Höhne and Joshua Claassen
Leibniz University Hannover, Germany

German Centre for Higher Education Research and Science Studies (DZHW), Germany

Saijal Shahania
Otto von Guericke University Magdeburg, Germany

German Centre for Higher Education Research and Science Studies (DZHW), Germany

David Broneske
German Centre for Higher Education Research and Science Studies (DZHW), Germany

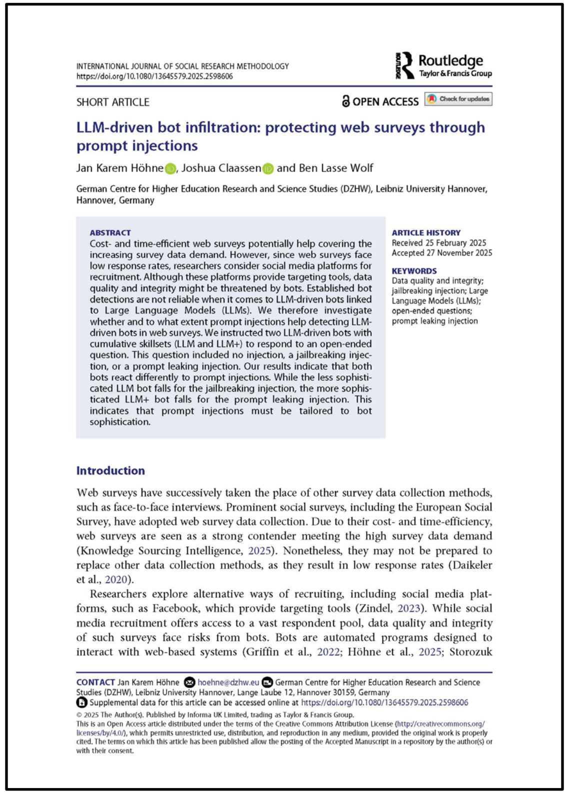
Abstract

Cost- and time-efficient web surveys have progressively replaced other survey modes. These efficiencies can potentially cover the increasing demand for survey data. However, since web surveys suffer from low response rates, researchers and practitioners start considering social media platforms as new sources for respondent recruitment. Although these platforms provide advertisement and targeting systems, the data quality and integrity of web surveys recruited through social media might be threatened by bots. Bots have the potential to shift survey outcomes and thus political and social decisions. This is alarming since there is ample literature on bots and how they infiltrate social media platforms, distribute fake news, and possibly skew public opinion. In this study, we therefore investigate bot behavior in web surveys to provide new evidence on common wisdom about the capabilities of bots. We programmed four bots – two rule-based and two AI-based bots – and ran each bot $N = 100$ times through a web survey on equal gender partnerships. We tested several bot prevention and detection measures, such as CAPTCHAs, invisible honey pot questions, and completion times. The results indicate that both rule- and AI-based bots come with impressive completion rates (up to 100%). In addition, we can prove conventional wisdom about bots in web surveys wrong: CAPTCHAs and honey pot questions pose no challenges. However, there are clear differences between rule- and AI-based bots when it comes to web survey completion.

Keywords

rule-based bots, AI-based bots, web surveys, completion behavior, data integrity, data quality

Corresponding author:
Jan Karem Höhne, Leibniz University Hannover, German Centre for Higher Education Research and Science Studies (DZHW), Research Infrastructure and Methods Division, Lange Laube 12, Hannover 30159, Germany.
Email: hoehne@dzhw.eu



INTERNATIONAL JOURNAL OF SOCIAL RESEARCH METHODOLOGY
https://doi.org/10.1080/13645579.2025.2598606

Routledge
Taylor & Francis Group

SHORT ARTICLE

OPEN ACCESS Check for updates

LLM-driven bot infiltration: protecting web surveys through prompt injections

Jan Karem Höhne, Joshua Claassen and Ben Lasse Wolf

German Centre for Higher Education Research and Science Studies (DZHW), Leibniz University Hannover, Hannover, Germany

ABSTRACT

Cost- and time-efficient web surveys potentially help covering the increasing survey data demand. However, since web surveys face low response rates, researchers consider social media platforms for recruitment. Although these platforms provide targeting tools, data quality and integrity might be threatened by bots. Established bot detections are not reliable when it comes to LLM-driven bots linked to Large Language Models (LLMs). We therefore investigate whether and to what extent prompt injections help detecting LLM-driven bots in web surveys. We instructed two LLM-driven bots with cumulative skillsets (LLM and LLM+) to respond to an open-ended question. This question included no injection, a jailbreaking injection, or a prompt leaking injection. Our results indicate that both bots react differently to prompt injections. While the less sophisticated LLM bot falls for the jailbreaking injection, the more sophisticated LLM+ bot falls for the prompt leaking injection. This indicates that prompt injections must be tailored to bot sophistication.

ARTICLE HISTORY
Received 25 February 2025
Accepted 27 November 2025

KEYWORDS
Data quality and integrity; jailbreaking injection; Large Language Models (LLMs); open-ended questions; prompt leaking injection

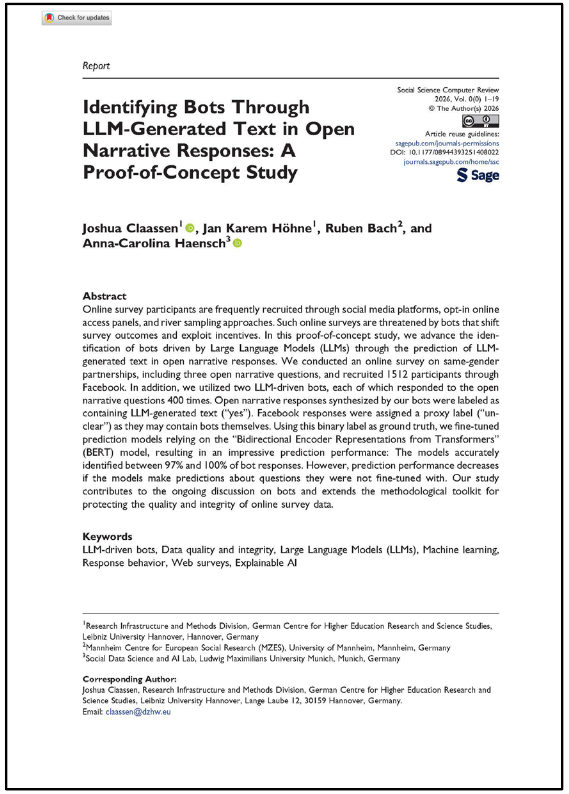
Introduction

Web surveys have successively taken the place of other survey data collection methods, such as face-to-face interviews. Prominent social surveys, including the European Social Survey, have adopted web survey data collection. Due to their cost- and time-efficiency, web surveys are seen as a strong contender meeting the high survey data demand (Knowledge Sourcing Intelligence, 2025). Nonetheless, they may not be prepared to replace other data collection methods, as they result in low response rates (Daikeler et al., 2020).

Researchers explore alternative ways of recruiting, including social media platforms, such as Facebook, which provide targeting tools (Zindel, 2023). While social media recruitment offers access to a vast respondent pool, data quality and integrity of such surveys face risks from bots. Bots are automated programs designed to interact with web-based systems (Griffin et al., 2022; Höhne et al., 2025; Storozuk

CONTACT Jan Karem Höhne ✉ hoehne@dzhw.eu German Centre for Higher Education Research and Science Studies (DZHW), Leibniz University Hannover, Lange Laube 12, Hannover 30159, Germany

Supplemental data for this article can be accessed online at https://doi.org/10.1080/13645579.2025.2598606



Check for updates

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

Many thanks for your attention!

hoehne@dzhw.eu

Literature

- 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, 513-539. https://doi.org/10.1093/jssam/smz008
- 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, 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, 2841-2851. https://doi.org/10.1007/s11135-021-01252-1
- Höhne, J.K., Claassen, J., Shahania, S., & Broneske, D. (2024). Bots in web survey interviews: A showcase. International Journal of Market Research, 67, 3-12. https://doi.org/10.1177/14707853241297009
- 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
- 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, 472-481. https://doi.org/10.20982/tqmp.16.5.p472
- Xu, Y., Pace, S., Kim, J., et al. (2022). Threats to online surveys: Recognizing, detecting, and preventing survey bots. Social Work Research, 46, 343-350. https://doi.org/10.1093/swr/svac023
- 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, 235–242.
- 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. Médini (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. (2022). Social media recruitment in online survey research: A systematic literature review. Methods, Data, Analysis. https://doi.org/10.12758/mda.2022.15

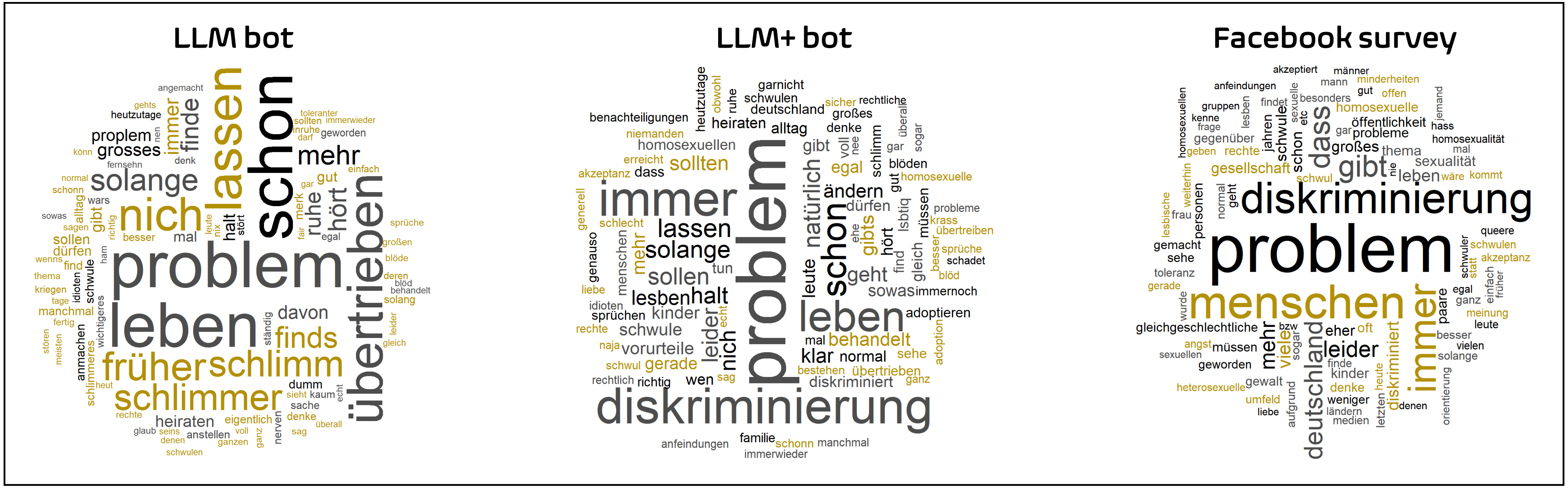
Appendix A: Open Narrative Questions

- **ONQ1**: In the last question, you indicated to find it (very good | rather good | rather not good | not good at all) that married same-sex 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?

Appendix B: Word Choice (ONQ2)



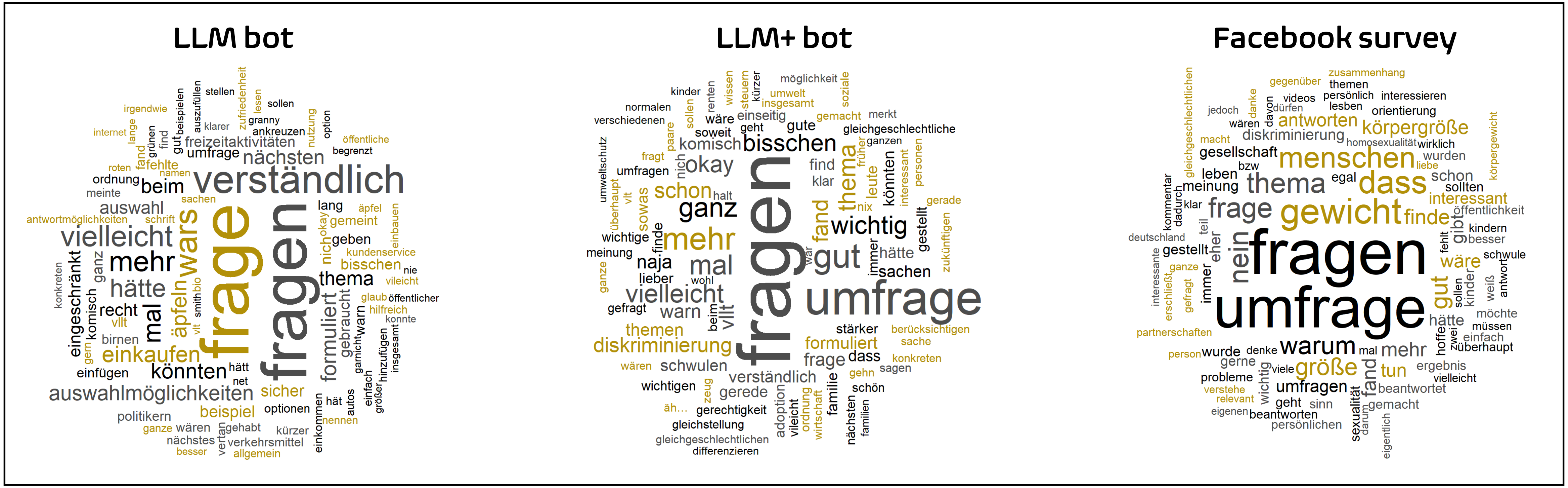
LLM bot

LLM+ bot

Facebook survey

Note. Each word cloud contains the 100 most frequently mentioned words (ONQ1) among the LLM bot, LLM+ bot, and Facebook survey, respectively. The size of a word is proportional to its frequency.

Appendix C: Word Choice (ONQ3)



Note. Each word cloud contains the 100 most frequently mentioned words (ONQ1) among the LLM bot, LLM+ bot, and Facebook survey, respectively. The size of a word is proportional to its frequency.

Appendix D: Prompt (Baseline Design)

LLM version
gemini-1.5-pro-002.

Open narrative questions – Prompt design (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.

Open narrative questions – Prompt design (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}"

Personas (LLM+ bot)
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 E: Token Analysis

	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