

You Never Can Tell: A Survey of Preferences on LLM-Generated Counterspeech in an Austrian Context

Pia Pachinger

TU Wien Informatics, Vienna
pia.pachinger@tuwien.ac.at

Arianna Muti

MilaNLP
Bocconi University, Milan

Abstract

We present the first survey study of German-language counterspeech in an Austrian context, where participants ranked LLM-generated responses to offensive comments involving xenophobia, sexual objectification, and political incivility. We compare generations from five state-of-the-art LLMs and additionally examine whether participants perceive them as human- or machine-written. Our results show substantial disagreement in rankings, indicating that counterspeech preferences are highly subjective. Although participants express a clear preference for human-written counterspeech in real-world settings, they fail to reliably identify LLM-generated responses as machine-written. These findings point to the need for personalized counterspeech systems that account for users' diverse preferences, rather than assuming a single universally effective response. Such systems could combine LLMs' capacity to generate high-quality counterspeech at scale with human involvement.¹

1 Introduction

Among the various strategies for countering toxic speech², counterspeech generation has emerged as a promising approach (Bonaldi et al., 2024a). Rather than simply removing harmful content, counterspeech (CS) aims to actively challenge toxic speech by offering constructive, persuasive and non-polarized discourse, which might offer alternative standpoint both to the author of the toxic message and to users navigating the online web and running into hateful comments (Garland et al., 2022; Moscato et al., 2025). The generation of CS

spans a wide spectrum, encompassing variation in speaker identity, stylistic approach, and content, including strategies such as humor, warnings of consequences, and appeals to empathy (Hangartner et al., 2021). Given the breadth of CS as a phenomenon, as well as the subjectivity inherent in how readers interpret offensive speech (Pachinger et al., 2025) and the CS that responds to it, perspectives on what constitutes effective CS are inherently diverse and culturally contingent. Evaluations of CS must therefore account for these factors (Bonaldi et al., 2024a). Despite this complexity, CS research continues to suffer from limited cultural and linguistic coverage (Chung et al., 2024), with existing work largely confined to a handful of country-specific contexts (e.g., Buerger 2021; Goffredo et al. 2022). A central obstacle to advancing this line of research is the scarcity of labeled non-English datasets suitable for generating and evaluating CS (Garland et al., 2020; Bonaldi et al., 2024a). The German-language context is a case in point: data coverage remains sparse, with few works focusing on German CS (e.g. Garland et al. 2020; Keller and Askanius 2020). To the best of our knowledge, no prior work has specifically addressed counterspeech in an Austrian context, and we are not aware of any publicly available dataset on counterspeech in German.

To address this gap, we conduct a survey in an Austrian context and analyze participants' perceptions of CS across three categories of offensive content: xenophobia, sexual objectification, and general incivility directed at politicians and the public. Participants answer three survey questions: (1) they rank CS generated by a diverse set of LLMs according to which responses they consider most effective in redirecting the discussion to a less degrading level; (2) they indicate whether they believe a given CS response was written by a human or a machine; and (3) they report whether they would prefer CS to be written by humans or machines in a real-

¹Data is available at <https://github.com/pi-pa/counterspeech>.

²Since there is no universally recognized way to disambiguate abusive, uncivil, offensive, and toxic speech (Pachinger et al., 2023), we use *toxic speech* as an umbrella term covering all of these. Hate speech is a specific, more severe category within the broader spectrum of toxic speech (Korre et al., 2025).

world setting. Our results indicate considerable disagreement among participants in their judgments of counterspeech effectiveness. Moreover, they express a stated preference for counterspeech written by humans over machine-generated alternatives, still, they are often unable to reliably distinguish between human- and machine-generated CS.

2 Related Work

Counterspeech Best practices suggest that CS should show credible evidence, factual arguments and alternative viewpoints by adopting an empathetic, polite and constructive tone (Fantón et al., 2021; Chung et al., 2021; Bonaldi et al., 2024a). Several datasets have been developed for counterspeech. The first large-scale, multilingual, expert-based dataset, Counter Narratives through Niche-sourcing (CONAN) (Chung et al., 2019), consists of hate speech counternarrative pairs in English, French, and Italian, focusing on Islamophobia. This dataset has been later expanded by considering other protected categories: race, religion, country of origin, sexual orientation, disability, and gender (Fantón et al., 2021). Recently, Hengle et al. (2025) introduced a novel dataset for evaluating counterspeech quality across four dimensions: contextual-relevance, aggressiveness, argument-coherence, and suitability, while Bonaldi et al. (2026) introduced CATCH-ME, a multilingual, multi-turn counterspeech dataset anchored in verified external knowledge, including fact-checking articles and NGO reports. Moreover, the implicit nature of hate speech (Muti et al., 2024) motivates a more structured approach to CS generation. Accordingly, Bonaldi et al. (2024b) decompose hate speech into premises and conclusions, showing that targeting specific components, particularly the implied statement, leads to richer argumentative responses. Similar to task 2 in our survey, Song et al. (2025) assess the human likeness of counterspeech based on the distinguishability between AI-generated and human-written replies to hate speech. Research on counterspeech generation has increased due to LLMs’ increased performance in text generation (Zubiaga et al., 2024; Russo et al., 2025). However, work on the effectiveness of LLMs generated CS in a German-language context is still limited.

Toxic Speech and Counterspeech in Austria

Further, research on toxic speech and CS in an Austrian context remains scarce. There exist publicly available datasets on inappropriate and discriminatory posts (Schabus et al., 2017), sexism and misogyny (Gross et al., 2024; Krenn et al., 2024), and toxic speech containing annotated rationales for inappropriate language, individual targets, targeted groups, and non-human targets (Pachinger et al., 2024). To the best of our knowledge, no prior work has analysed CS in an Austrian context.

3 Survey Design

This survey on counterspeech constitutes part of a larger study in which participants can choose between two tasks: ranking counterspeech examples or annotating toxic comments. In this paper, we report on the subset of participants who opted for the counterspeech task. The survey examines how participants rank counterspeech generated by different LLMs, whether they can distinguish between human- and AI-generated responses, and whether they prefer counterspeech produced by humans or by machines.

Vignettes and Counterspeech Generation We select three toxic posts (vignettes), each exhibiting a distinct form of toxicity, namely xenophobia, sexual objectification, and incivility toward politicians, from a discussion forum of an Austrian news media outlet, Der Standard. We filter the vignettes through a combination of manual inspection, keyword search, and the Perspective API by sourcing for identity attack and extracting the targets of toxic statements via keyword search. These vignettes are representative of the toxic content commonly found in this forum. Appendix A contains the vignettes and Appendix B contains the translations of the vignettes. We employ five state-of-the-art LLMs, namely GPT-4o, Mistral Small 3.1, Claude 3.7 Sonnet, Qwen2.5-Max, and Gemini 2.0 Flash, to generate counterspeech. To this end, we formulate a German-language prompt comprising the title of the newspaper article under which the hateful comment appeared (for contextual grounding), the comment itself, and instructions to generate a response that redirects the discussion toward a more factual level in a humorous manner, reflecting the tone of a newspaper discussion forum. The prompt and generations and their translations are in Appendix A and B.

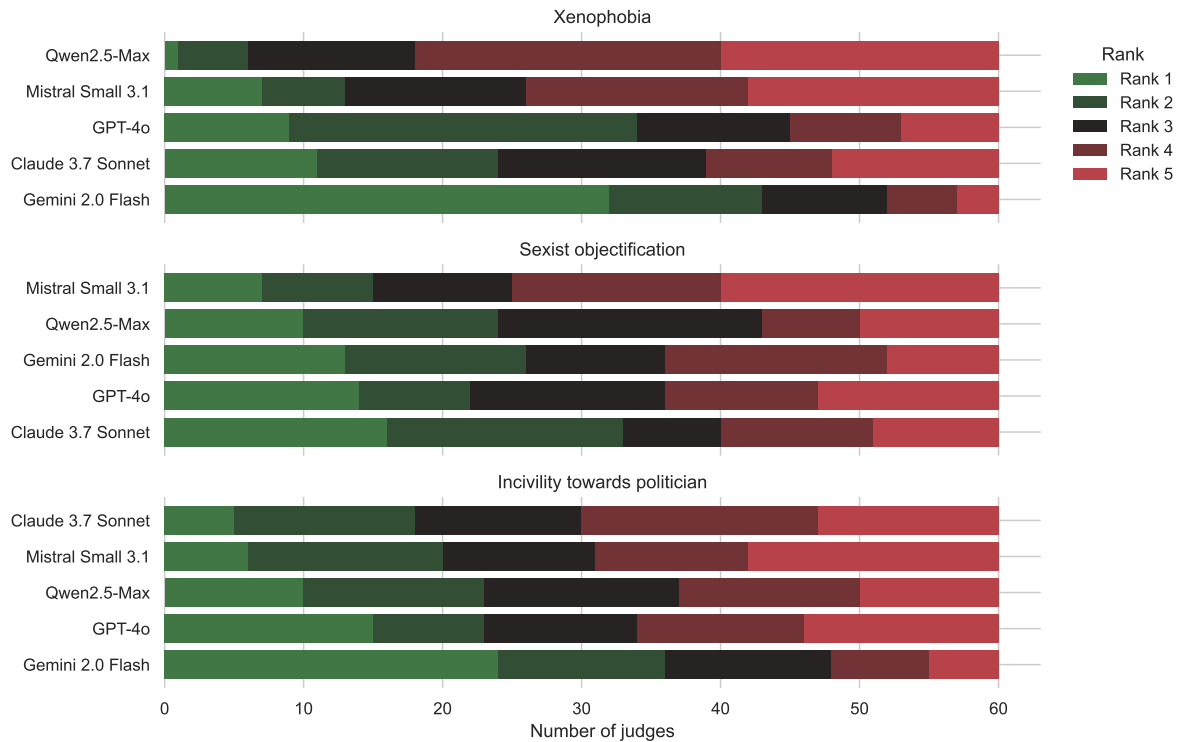


Figure 1: Participants rank counterspeech generated by different LLMs according to how effective they believe to make the discussion less degrading. 1 represents the most effective and 5 the least effective.

Task 1. Participants first provide basic sociodemographic information. They are further asked to rank which counterspeech generated by the different models they prefer, for each of the three comments. Ranks range from 1 to 5. Preference is indicated by which responses most effectively return the discussion to a less degrading level. Participants are shown the title of the newspaper article under which the comment was published, along with a short description of the hateful comment, which, for ethical reasons, is not displayed. Appendix C shows an example of the hateful comment, along with the description provided to the participants. The interface for this task is shown in Figure 4 in Appendix D.

Task 2. Subsequently, each participant is shown three randomly selected counterspeech examples from the previous task and asked to indicate, on a 7-point Likert scale, whether they believe the counterspeech was written by a human or a machine. The Likert scale ranges from 1 – *definitely written by a human* to 7 – *definitely written by a machine*, with intermediate values (2 – 6) representing gradations of certainty. We limit each participant to three examples to keep their workload manageable and avoid annotation fatigue.

Task 3. Finally, we ask participants to indicate their preference for counterspeech generated by humans versus machines, again using a 7-point Likert scale from 1 – *definitely by a human* to 7 – *definitely by a machine*.

Survey Participants The survey was completed by 68 participants. After excluding entries that fail the attention checks, 60 valid responses remain for analysis. Participants are recruited from educational institutions in Austria and are enrolled in either Communication Science ($n = 31$) or Computer Science ($n = 29$). As compensation, the students receive bonus points, the median completion time is 18.8 minutes. Participant age ranges from 19 to 39 years (Median = 23). All participants reside in Austria. With respect to national origin, 19 participants have Austrian origin (3 with migration background), 11 have German origin, 4 have Ukrainian origin, and 4 have Turkish origin, with all remaining nationalities represented by smaller numbers of participants. 36 participants identify as female, 22 as male, 2 prefer not to say. 23 participants have a high school degree, 30 hold a Bachelor’s degree, and 7 hold a Master’s degree (or equivalents). 18 participants live in Austria, but their German level is not C1 or higher, we therefore

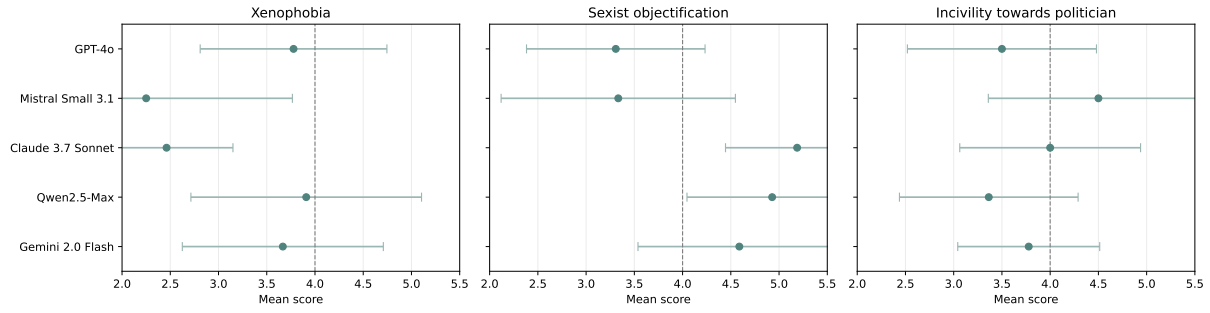


Figure 2: Participants indicate the extent to which they believe the presented counterspeech was generated by a human (1 - *definitely by a human*) or a machine (7 - *definitely by a machine*), with intermediate values representing gradations of certainty. The plots contain the means and 95 % confidence intervals of the distributions of all judgements by all participants.

translate the survey to English for them.

4 Results

Participants disagree on rankings of LLM-generated counterspeech Figure 1 presents the distribution of the rankings. For the post exhibiting xenophobic content, a clear majority of participants (more than half) rated the counterspeech generated by Gemini as the most suitable. A similar, though somewhat weaker, pattern emerges for the comment containing incivility directed at a politician, where 24 participants judged Gemini’s counterspeech to be the most effective in returning the discussion to a less degrading level. Beyond these two cases, however, the rank distributions are considerably more balanced, with no model emerging as a consistent favorite. These findings are in line with prior work, showing that different counterspeech generators excel on different instances and evaluation dimensions rather than yielding a universally preferred system (Zheng et al., 2023; Saha et al., 2024; Hengle et al., 2025). This pattern also suggests that the perceived appropriateness of counterspeech is substantially subjective, in line with previous findings that judgments in toxic speech and counterspeech-related annotation are shaped by annotator perspective and the inherently subjective nature of the task (Mostafazadeh Davani et al., 2022; Yu et al., 2022; Alacam et al., 2025; Pachinger et al., 2025).

Participants cannot distinguish human written from LLM written counterspeech Figure 2 presents participants’ judgments of whether LLM-generated content was perceived as human- or machine-produced. We interpret these ratings using the associated confidence intervals of the means

over all judgments: confidence intervals falling entirely below 4 indicate a systematic belief that the counterspeech was human-written, confidence intervals overlapping 4 but remaining below 5 indicate that participants were unable to distinguish between human and machine authorship, and confidence intervals falling entirely above 4 indicate that the participants correctly identified the CS to be machine-generated.

For 8 of the 15 generations, confidence intervals overlap the midpoint, but remain below 5, indicating that participants could not reliably distinguish whether the counterspeech was produced by an LLM or a human. In two cases, namely the CS generated by Claude and Qwen for the vignette containing sexual objectification, confidence intervals fall entirely above 4, indicating that participants correctly identified the counterspeech as machine-generated. In two further cases, namely Mistral’s and Claude’s CS generations for the vignette containing xenophobia, confidence intervals fall entirely below 4, indicating that participants systematically believed the counterspeech to be human-written. In three cases, Qwen’s CS for the vignette containing xenophobia, Gemini’s CS for the vignette containing sexual objectification, and Mistral’s CS for the vignette containing incivility towards a politician, confidence intervals overlap both 4 and 5. For these cases we cannot reliably draw conclusions.

From this we follow that for 10 out of 12 CS generations, where we can reliably draw conclusions, the participants were not able to identify them as LLM-generated. This is in contrast with previous findings that AI-generated and human-written counterspeech can be easily distinguished by both simple classifiers and humans (Song et al., 2025). How-

ever, other lines of work suggest that humans are unable to detect whether self-presentations and persuasive texts, generated by state-of-the-art LLMs, were generated by humans or machines (Jakesch et al., 2023; Zhu et al., 2025). Perceived authorship does not appear to align with perceived effectiveness. For example, Qwen’s xenophobia CS received the lowest effectiveness ranking despite its ambiguous authorship, whereas Claude’s response to sexist objectification was ranked highest despite being recognized as machine-generated. This suggests that participants may value effective counterspeech even when they perceive it as AI-generated.

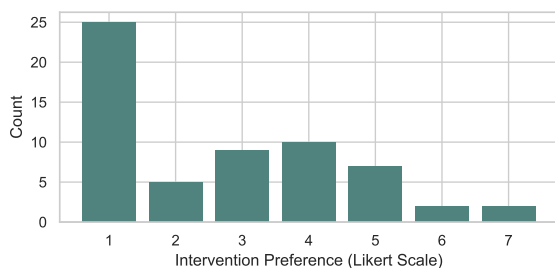


Figure 3: Participants indicate whether they prefer counterspeech to be produced by humans (1 - *definitely a human*) or machines (7 - *definitely a machine*).

Participants prefer human written counterspeech Figure 3 presents the participants’ self-reported preference on whether counterspeech should be generated by humans or machines. 50% of participants clearly indicate a preference for human-generated counterspeech, which presents an interesting counterpoint to their inability to distinguish human-written from machine-written counterspeech. 26 participants that selected 3, 4, or 5 are undecided as to whether they prefer human- or machine-generated counterspeech, while 4 participants express a clear preference for machine-generated counterspeech.

5 Conclusion

We presented a survey study of German-language counterspeech in an Austrian context, focusing on participants’ perceptions of LLM-generated responses to offensive comments. Across three forms of toxicity, namely xenophobia, sexual objectification, and political incivility, participants showed substantial disagreement in their rankings of counterspeech effectiveness. This finding reinforces the view that counterspeech evaluation is inherently subjective. We further showed that participants

express a clear preference for human-written counterspeech in real-world moderation settings, while often being unable to reliably distinguish LLM-generated counterspeech from human-written responses. This tension has important implications for the design of counterspeech systems: although LLM-generated counterspeech may offer scalability, user preferences and expectations regarding human involvement should not be ignored.

Acknowledgements

Pia Pachinger’s research is supported by the Vienna Science and Technology Fund (WWTF) [10.47379/ICT20015]. Arianna Muti’s research is supported by the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program (grant agreement No. 101116095, PERSONAE). She is a member of the MilaNLP group and the Data and Marketing Insights Unit of the Bocconi Institute for Data Science and Analysis. We thank Sophie Lecheler and Julia Neidhardt for the opportunity to run the survey studies with participants from their courses. We further thank Svenja Schäfer for helping set up the survey. Lastly, we thank the survey participants. Their dedication and effort have been invaluable to the success of this project.

Limitations

By design, this survey’s findings are limited to an Austrian context, drawing on a limited set of annotators who are not representative of the Austrian population (aged 19–39, holding at least a high school degree, and predominantly university students). As such, these results do not empirically generalize to other populations or cultural contexts. Nonetheless, they offer a cautionary note and a direction for future research and practice: notably, we present evidence that implies that certain populations of annotators exhibit high disagreement in their judgments of counterspeech. This finding should encourage caution in future data collection and quality assurance efforts related to counterspeech judgments, as apparent agreement within a small annotator cohort may mask substantial disagreement within a larger, more diverse one. Further, we show that this particular sample of survey participants living in Austria definitely prefers human moderators to produce counterspeech. Without being able to provide evidence, we suspect that this finding generalizes to other samples of

annotators. Moreover, for participants that did the survey in English, translations from German may introduce semantic and cultural shifts (Korre et al., 2024; Muti et al., 2024, 2025), potentially biasing their judgment.

Ethical Statement

The survey was approved by our institutions' ethics review boards.

The Use of Generative AI in this Work

We used GitHub Copilot to assist with dataset curation and the creation of visualizations for this paper. Additionally, we used Claude Sonnet 5 and ChatGPT to improve the clarity and style of our writing.

References

- Özge Alacam, Sanne Hoeken, Andreas Säuberli, Hannes Gröner, Diego Frassinelli, Sina Zarriß, and Barbara Plank. 2025. [Disentangling subjectivity and uncertainty for hate speech annotation and modeling using gaze](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28707–28724, Suzhou, China. Association for Computational Linguistics.
- Helena Bonaldi, Yi-Ling Chung, Gavin Abercrombie, and Marco Guerini. 2024a. [NLP for counterspeech against hate: A survey and how-to guide](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3480–3499, Mexico City, Mexico. Association for Computational Linguistics.
- Helena Bonaldi, Greta Damo, Nicolás Benjamín Ocampo, Elena Cabrio, Serena Villata, and Marco Guerini. 2024b. [Is safer better? the impact of guardrails on the argumentative strength of LLMs in hate speech countering](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3446–3463, Miami, Florida, USA. Association for Computational Linguistics.
- Helena Bonaldi, Genoveffa Martone, and Marco Guerini. 2026. [CATCH-ME if you RAG: A dataset of contextually annotated multi-turn counterspeech against hate and misinformation exchanges](#). *Preprint*, arXiv:2606.20369.
- Catherine Buerger. 2021. [#iamhere: Collective counterspeech and the quest to improve online discourse](#). *Social Media + Society*, 7.
- Yi-Ling Chung, Gavin Abercrombie, Florence Enock, Jonathan Bright, and Verena Rieser. 2024. [Understanding counterspeech for online harm mitigation](#). *Northern European Journal of Language Technology*, 10:30–49.
- Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. [CONAN - COUNTER NARRATIVES through nichesourcing: a multilingual dataset of responses to fight online hate speech](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.
- Yi-Ling Chung, Serra Sinem Tekiroglu, and Marco Guerini. 2021. [Towards knowledge-grounded counter narrative generation for hate speech](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 899–914, Online. Association for Computational Linguistics.
- Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroglu, and Marco Guerini. 2021. [Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3226–3240, Online. Association for Computational Linguistics.
- Joshua Garland, Keyan Ghazi-Zahedi, Jean-Gabriel Young, Laurent Hébert-Dufresne, and Mirta Galesic. 2020. [Countering hate on social media: Large scale classification of hate and counter speech](#). In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 102–112, Online. Association for Computational Linguistics.
- Joshua Garland, Keyan Ghazi-Zahedi, Jean-Gabriel Young, Laurent Hébert-Dufresne, and Mirta Galesic. 2022. [Impact and dynamics of hate and counter speech online](#). *EPJ data science*, 11(1):3.
- Pierpaolo Goffredo, Valerio Basile, Bianca Cepollaro, and Viviana Patti. 2022. [Counter-TWIT: An Italian corpus for online counterspeech in ecological contexts](#). In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 57–66, Seattle, Washington (Hybrid). Association for Computational Linguistics.
- Stephanie Gross, Johann Petrak, Louisa Venhoff, and Brigitte Krenn. 2024. [GermEval2024 shared task: GerMS-detect – sexism detection in German online news fora](#). In *Proceedings of GermEval 2024 Task 1 GerMS-Detect Workshop on Sexism Detection in German Online News Fora (GerMS-Detect 2024)*, pages 1–9, Vienna, Austria. Association for Computational Linguistics.
- Dominik Hangartner, G. Gennaro, Sary Alasiri, Nicholas Bahrach, Alexandra Bornhoft, Joseph Boucher, Buket Buse Demirci, Lauren M. Derksen, A. Hall, Matthias Jochum, María M. Muñoz, M. Richter, F. Vogel, Salome Wittwer, Felix Wüthrich, F. Gilardi, and K. Donnay. 2021. [Empathy-based counterspeech can reduce racist hate speech in a social media field experiment](#). *Proceedings of the National Academy of Sciences of the United States of America*, 118.

- Amey Hengle, Aswini Kumar Padhi, Anil Bandhakavi, and Tanmoy Chakraborty. 2025. [CSEval: Towards automated, multi-dimensional, and reference-free counterspeech evaluation using auto-calibrated LLMs](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5402–5419, Albuquerque, New Mexico. Association for Computational Linguistics.
- Maurice Jakesch, Jeffrey T. Hancock, and Mor Naaman. 2023. [Human heuristics for ai-generated language are flawed](#). *Proceedings of the National Academy of Sciences*, 120(11).
- Nadine Keller and Tina Askanius. 2020. Combatting hate and trolling with love and reason?: A qualitative analysis of the discursive antagonisms between organised hate speech and counterspeech online. *SCM Studies in Communication and Media*, 9(4):540–572.
- Katerina Korre, Arianna Muti, and Alberto Barrón-Cedeño. 2024. [The challenges of creating a parallel multilingual hate speech corpus: An exploration](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15842–15853, Torino, Italia. ELRA and ICCL.
- Katerina Korre, Arianna Muti, Federico Ruggeri, and Alberto Barrón-Cedeño. 2025. [Untangling hate speech definitions: A semantic componential analysis across cultures and domains](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3184–3198, Albuquerque, New Mexico. Association for Computational Linguistics.
- Brigitte Krenn, Johann Petrak, Marina Kubina, and Christian Burger. 2024. [GERMS-AT: A sexism/misogyny dataset of forum comments from an Austrian online newspaper](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7728–7739, Torino, Italia. ELRA and ICCL.
- Emanuele Moscato, Arianna Muti, and Debora Nozza. 2025. [MilaNLP @multilingual counterspeech generation: Evaluating translation and background knowledge filtering](#). In *Proceedings of the First Workshop on Multilingual Counterspeech Generation*, pages 56–64, Abu Dhabi, UAE. Association for Computational Linguistics.
- Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. [Dealing with disagreements: Looking beyond the majority vote in subjective annotations](#). *Transactions of the Association for Computational Linguistics*, 10:92–110.
- Arianna Muti, Chris Emmery, Debora Nozza, Alberto Barrón-Cedeño, and Tommaso Caselli. 2025. [The “r” in “woman” stands for rights. auditing LLMs in uncovering social dynamics in implicit misogyny](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5462–5479, Suzhou, China. Association for Computational Linguistics.
- Arianna Muti, Federico Ruggeri, Khalid Al Khatib, Alberto Barrón-Cedeño, and Tommaso Caselli. 2024. [Language is scary when over-analyzed: Unpacking implied misogynistic reasoning with argumentation theory-driven prompts](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21091–21107, Miami, Florida, USA. Association for Computational Linguistics.
- Pia Pachinger, Janis Goldzycher, Anna Planitzer, Wojciech Kusa, Allan Hanbury, and Julia Neidhardt. 2024. [AustroTox: A dataset for target-based Austrian German offensive language detection](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11990–12001, Bangkok, Thailand. Association for Computational Linguistics.
- Pia Pachinger, Janis Goldzycher, Anna M. Planitzer, Julia Neidhardt, and Allan Hanbury. 2025. [A disaggregated dataset on English offensiveness containing spans](#). In *Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP*, pages 1–14, Suzhou, China. Association for Computational Linguistics.
- Pia Pachinger, Allan Hanbury, Julia Neidhardt, and Anna Planitzer. 2023. [Toward disambiguating the definitions of abusive, offensive, toxic, and uncivil comments](#). In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 107–113, Dubrovnik, Croatia. Association for Computational Linguistics.
- Daniel Russo, Helena Bonaldi, Yi-Ling Chung, Gavin Abercrombie, and Marco Guerini. 2025. [NLP for counterspeech against hate and misinformation \(CSHAM\)](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 5: Tutorial Abstracts)*, pages 9–10, Vienna, Austria. Association for Computational Linguistics.
- Punyajoy Saha, Abhilash Datta, Abhik Jana, and Animesh Mukherjee. 2024. [CrowdCounter: A benchmark type-specific multi-target counterspeech dataset](#). In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 470–488, Miami, FL, USA. Association for Computational Linguistics.
- Dietmar Schabus, Marcin Skowron, and Martin Trapp. 2017. [One million posts: A data set of german online discussions](#). In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 1241–1244, Tokyo, Japan.
- Xiaoying Song, Sujana Mamidisetty, Eduardo Blanco, and Lingzi Hong. 2025. [Assessing the human likeness of AI-generated counterspeech](#). In *Proceedings of the 31st International Conference on Compu-*

tational Linguistics, pages 3547–3559, Abu Dhabi, UAE. Association for Computational Linguistics.

Xinchen Yu, Eduardo Blanco, and Lingzi Hong. 2022. [Hate speech and counter speech detection: Conversational context does matter](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5918–5930, Seattle, United States. Association for Computational Linguistics.

Yi Zheng, Björn Ross, and Walid Magdy. 2023. [What makes good counterspeech? a comparison of generation approaches and evaluation metrics](#). In *Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA)*, pages 62–71, Prague, Czechia. Association for Computational Linguistics.

Tiffany Zhu, Iain Weissburg, Kexun Zhang, and William Yang Wang. 2025. [Human bias in the face of AI: Examining human judgment against text labeled as AI generated](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25907–25914, Vienna, Austria. Association for Computational Linguistics.

Iruna Zubiaga, Aitor Soroa, and Rodrigo Agerri. 2024. [A LLM-based ranking method for the evaluation of automatic counter-narrative generation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9572–9585, Miami, Florida, USA. Association for Computational Linguistics.

A Vignettes and Counterspeech Generations

We use the following prompt for generating counterspeech with five different models:

Im Folgenden siehst du einen Titel und einen Kommentar. Beide wurden in einem österreichischen Kontext geschrieben. Schreibe selbst einen humorvollen Kommentar der nicht länger als fünf Sätze ist und versuche, die Diskussion wieder auf eine fachlichere Ebene zu bringen. Verwende keine Emojis.

Titel: "Drogenszene an der U6: Wo sind die Streetworker?" Kommentar: Wann wird abgeschoben ist die wichtigste Frage. Die Gegend ist zu gefährlich fuer Streetworker. Man hat es uebersehen und ist drauf und dran mit dem Kuschkurs gegenueber Kriminellen und Illegalen das ganze Land zu vernichten.

Observe the translation of the prompt for generating counterspeech below. Note that we did not

use this translation for the English version of the survey, since we translated the German-generated counterspeech directly.

Below you'll see a title and a comment. Both were written in an Austrian context. Write your own humorous comment, no longer than five sentences, and try to bring the discussion back to a more professional/expert level. Don't use any emojis.

Title: "Drug scene at the U6: Where are the street workers?" Comment: When will they be deported is the most important question. The area is too dangerous for street workers. It's been overlooked, and with this cuddly/soft approach toward criminals and illegals, they're on the verge of destroying the whole country.

Vignette Containing Xenophobia

Titel: "Drogenszene an der U6: Wo sind die Streetworker?" Kommentar: Wann wird abgeschoben ist die wichtigste Frage. Die Gegend ist zu gefährlich fuer Streetworker. Man hat es uebersehen und ist drauf und dran mit dem Kuschkurs gegenueber Kriminellen und Illegalen das ganze Land zu vernichten.

GPT-4o Streetworker sind dazu da, Menschen zu helfen – nicht sie abzuschieben. Vielleicht wäre die wichtigere Frage, wie man Prävention, Therapie und Sicherheit in Einklang bringt. Oder ob jemand mal geprüft hat, was in anderen Städten besser funktioniert. Ein bisschen weniger Panikmache und ein bisschen mehr sachliche Diskussion könnten hier Wunder wirken. Also: Gibt's Ideen?

Mistral Small 3.1 Vielleicht sollten wir den Streetworkern erstmal eine kugelsichere Weste und ein paar Ninja-Skills beibringen, bevor wir sie in die Höhle des Löwen schicken? Aber im Ernst, wie können wir die Sicherheit für alle Beteiligten gewährleisten und gleichzeitig effektive Hilfe leisten? Gibt es bereits erfolgreiche Modelle in anderen Städten, von denen wir lernen können? Und wie sieht es mit der Zusammenarbeit zwischen Polizei und Sozialarbeitern aus – gibt es da noch Verbesserungspotenz

Claude 3.7 Sonnet Ah, nach dem Wutschnauben dürfen wir uns vielleicht kurz daran erinnern, dass

komplexe soziale Probleme selten mit Einwort-Lösungen zu bewältigen sind. In Wien braucht's wohl mehr als nur einen Abschiebe-Reflex, um der Drogenproblematik Herr zu werden – so ähnlich wie mein Bierkonsum nicht durch Wegsperrten des Kühlschranks zu lösen ist. Die Statistik zeigt übrigens, dass Streetworker durchaus erfolgreich in gefährlichen Gegenden arbeiten, wenn man sie mit ausreichend Ressourcen, Sicherheitskonzepten und interdisziplinärer Unterstützung ausstattet. Vielleicht könnten wir ja mal über evidenzbasierte Ansätze diskutieren, die in anderen Städten funktioniert haben?

Qwen2.5-Max Ah, der Klassiker: "Abschieben und fertig!" – wenn es doch nur so einfach wäre wie ein Wiener Schnitzel zu panieren. Streetworker sind übrigens keine Actionhelden, sondern arbeiten am liebsten dort, wo sie auch halbwegs sicher operieren können. Vielleicht sollten wir uns weniger auf Schlagworte und mehr auf konkrete Sicherheitskonzepte konzentrieren? Denn ohne Präsenz vor Ort wird sich die Drogenszene kaum von selbst auflösen. Oder glauben Sie, dass Dealer freiwillig einen Standortwechsel in Richtung Utopia planen?

Gemini 2.0 Flash Es ist sicherlich richtig, dass die Situation an der U6 besorgniserregend ist. Aber bevor wir über Abschiebungen sprechen, sollten wir uns vielleicht fragen, warum Streetworker dort nicht sicher arbeiten können. Vielleicht liegt das Problem nicht nur bei den Menschen dort, sondern auch an fehlenden Ressourcen oder einer unzureichenden Unterstützung für diese wichtige Arbeit. Es wäre hilfreich, wenn wir uns auf Fakten und mögliche Lösungen konzentrieren könnten, anstatt voreilige Schlüsse zu ziehen.

Vignette Containing Sexist Objectification

Titel: Song Contest 2016: Wer für Österreich nach Stockholm fährt Kommentar: Sind die Mädels wenigstens fesch?^^Schau mir das ja nicht an, bin nur wegen dem Ticker hier

GPT-4o Optik ist ja nett, aber beim Song Contest geht's doch hoffentlich um Musik. Vielleicht könnten wir also diskutieren, welche musikalischen Qualitäten der Beitrag mitbringen sollte, um in Stockholm zu punkten. Gibt es Favoriten, die schon jetzt vielversprechend klingen?

Mistral Small 3.1 Das ist ja mal eine unerwartete Perspektive! Aber mal im Ernst, wer sind die Kan-

didaten und wie wurden sie ausgewählt? Gibt es schon Favoriten oder Überraschungen? Und wie schätzt ihr die Chancen für Österreich in diesem Jahr ein?

Claude 3.7 Sonnet Während die optische Anmutung mancher Künstler durchaus ein unterhaltsamer Nebenaspekt sein kann, wäre es doch interessant, über die musikalische Qualität der österreichischen Beiträge zu diskutieren. Vielleicht sollten wir uns fragen, mit welcher Stilrichtung Österreich nach dem Conchita-Erfolg und dem letztjährigen Experiment mit den Makemakes am ehesten punkten könnte. Die ESC-Jury bewertet bekanntlich nicht das Aussehen, sondern Performance, Komposition und Stimme – obwohl manchmal der Eindruck entstehen könnte, es sei anders herum. Bei aller berechtigten Skepsis: Die musikalische Bandbreite des Contests hat in den letzten Jahren tatsächlich zugenommen.

Qwen2.5-Max Nun, bei einem Wettbewerb dieser Art sollte man die ästhetische Präsentation wohl kaum in den Vordergrund stellen - es sei denn, wir sprechen über Bühnendesign. Wenn der Ticker das einzige Medium ist, das Sie nutzen, um den Eurovision Song Contest zu verfolgen, könnten Sie dabei aber interessante statistische Trends verpassen. Wie wäre es, die Performance aus musikwissenschaftlicher Perspektive zu analysieren? Immerhin hat Österreich eine reiche Chortradition, die bis ins 19. Jahrhundert zurückreicht. Da wäre es doch spannend zu sehen, wie sich diese Tradition in modernen Arrangements widerspiegelt.

Gemini Flash 2.0 Ich finde es spannend, wie unterschiedlich die Meinungen zur Auswahl der Kandidaten sind. Es zeigt, wie vielfältig Musikgeschmack sein kann. Wäre interessant, die musikalischen Einflüsse der Teilnehmer genauer zu analysieren. Vielleicht gibt es ja überraschende Verbindungen zu früheren Song-Contest-Beiträgen. Ich bin gespannt, wie sich die österreichische Auswahl im internationalen Vergleich schlagen wird.

Vignette Containing Incivility Against a Politician

Titel: "Politiker fällt mit unangebrachter Wortmeldung auf" Kommentar: Das dumpfe Österreich hat ein Sprachrohr mehr.... Gibts keine vernünftigen Politiker mehr?

GPT-4o Ach, die politische Landschaft gleicht manchmal einem Kabarett, nur ohne Eintrittskarte. Vielleicht wäre es spannender, nicht nur über die Wortmeldungen zu reden, sondern darüber, welche Konsequenzen sie haben. Sprache formt Politik – und umgekehrt. Wäre doch interessant, die eigentlichen Inhalte genauer zu analysieren, statt nur den Stil zu bewerten. Wer weiß, vielleicht steckt ja sogar eine relevante Debatte dahinter.

Mistral Small 3.1 Das war wohl eher ein sprachliches Stolperkabel als ein Sprachrohr. Vielleicht sollten wir Politikern statt Mikrofonen einfach mal ein Wörterbuch in die Hand drücken. Aber Spaß beiseite, wie können wir sicherstellen, dass unsere Vertreter besser auf ihre Wortwahl achten? Gibt es vielleicht Schulungen oder Richtlinien, die hier helfen könnten? Lasst uns konstruktiv bleiben und Lösungen finden, damit solche Fauxpas nicht mehr vorkommen.

Qwen2.5-Max Nun, vielleicht sollten wir das "dumpfe Österreich" nicht über einen Kamm scheren – immerhin hat auch dieser Politiker irgendwo Unterstützer. Vielleicht wäre es sinnvoller, sich auf die inhaltlichen Fehler zu konzentrieren statt pauschal zu urteilen. Immerhin ist Sprache ein mächtiges Werkzeug, und wer es ungeschickt einsetzt, hat zumindest eine Lehre verdient. Die Frage nach vernünftigen Politikern bleibt dabei berechtigt, aber auch sie sind nur Menschen mit gelegentlich allzu menschlichen Fehlern. Womöglich sollten wir mehr darüber diskutieren, wie man politische Bildung weiter professionalisieren kann.

Claude 3.7 Sonnet Scheint, als hätte unser Volksvertreter wieder mal den Diplomatie-Kurs geschwänzt. Bevor wir alle in den Chor der Empörung einstimmen, könnten wir uns vielleicht fragen, welche konkreten politischen Positionen hier eigentlich diskutiert werden sollten. Die Qualität unserer Demokratie misst sich ja letztendlich an der Sachlichkeit der Debatten und nicht an verbalen Entgleisungen. Vielleicht wäre es produktiver, über die tatsächlichen politischen Inhalte zu sprechen, anstatt bei jedem rhetorischen Fehltritt gleich das Untergangsszenario für die gesamte Republik auszurufen. Immerhin sind die großen Herausforderungen unserer Zeit komplex genug, um unsere volle Aufmerksamkeit zu verdienen.

Gemini Flash 2.0 Nun, vielleicht sollten wir uns auf die Sachlichkeit konzentrieren. Welche

konkreten politischen Maßnahmen oder Aussagen haben denn zu dieser Wortmeldung geführt? Es wäre doch konstruktiver, die Argumente des Politikers zu analysieren, anstatt nur über seine Ausdrucksweise zu urteilen. Vielleicht gibt es ja doch einen Funken Wahrheit in dem, was er gesagt hat. Und falls nicht, können wir immer noch darüber diskutieren, wie man politische Debatten respektvoller gestalten könnte.

B Translations of Vignettes and Generated Counterspeech

Vignette Containing Xenophobia

Title: "Drug scene at the U6: Where are the street workers?" Comment: When people get deported is the most important question. The area is too dangerous for street workers. It's been overlooked, and with this cuddly approach toward criminals and illegals, the whole country is on the verge of being destroyed.

GPT-4o Street workers exist to help people – not to deport them. Maybe the more important question would be how to balance prevention, therapy, and safety. Or whether anyone has checked what works better in other cities. A little less fearmongering and a little more factual discussion could work wonders here. So: any ideas?

Mistral Small 3.1 Maybe we should first give the street workers a bulletproof vest and teach them a few ninja skills before sending them into the lion's den? But seriously, how can we ensure safety for everyone involved while still providing effective help? Are there already successful models in other cities we could learn from? And what about cooperation between police and social workers – is there room for improvement there?

Claude 3.7 Sonnet Ah, after the venting, maybe we can briefly remember that complex social problems are rarely solved with one-word solutions. Vienna probably needs more than just a deportation reflex to get a handle on the drug problem – similar to how my beer consumption won't be solved by locking the fridge. The statistics actually show that street workers are quite successful in dangerous areas when equipped with sufficient resources, safety concepts, and interdisciplinary support. Maybe we could discuss evidence-based approaches that have worked in other cities?

Qwen2.5-Max Ah, the classic: "Just deport them and we're done!" – if only it were as simple as breeding a Wiener schnitzel. By the way, street workers aren't action heroes; they prefer to work where they can operate reasonably safely. Maybe we should focus less on slogans and more on concrete safety concepts? Because without a presence on the ground, the drug scene won't dissolve on its own. Or do you think dealers are planning a voluntary relocation to Utopia?

Gemini 2.0 Flash It's certainly true that the situation at the U6 is concerning. But before we talk about deportations, maybe we should ask why street workers can't work safely there. Maybe the problem isn't just with the people there, but also with a lack of resources or insufficient support for this important work. It would help if we could focus on facts and possible solutions instead of jumping to conclusions.

Vignette Containing Sexist Objectification

Title: Eurovision Song Contest 2016: Who's going to Stockholm for Austria Comment: Are the girls at least good-looking?^^I'm definitely not watching this, I'm only here for the live ticker

GPT-4o Looks are nice, but Eurovision is hopefully about the music. So maybe we could discuss what musical qualities the entry should bring to score points in Stockholm. Are there any favorites that already sound promising?

Mistral Small 3.1 Well, that's an unexpected take! But seriously, who are the candidates and how were they selected? Are there already favorites or surprises? And how do you all rate Austria's chances this year?

Claude 3.7 Sonnet While some artists' visual appearance can certainly be an entertaining side note, it would be more interesting to discuss the musical quality of the Austrian entries. Maybe we should ask which stylistic direction would give Austria the best shot after the Conchita success and last year's experiment with the Makemakes. The Eurovision jury supposedly doesn't judge looks, but performance, composition, and voice – even though sometimes it might seem the other way around. All reasonable skepticism aside: the musical range of the contest has actually increased in recent years.

Qwen2.5-Max Well, in a competition like this, aesthetic presentation probably shouldn't be the

main focus – unless we're talking about stage design. If the live ticker is the only medium you're using to follow the Eurovision Song Contest, you might be missing some interesting statistical trends. How about analyzing the performances from a musicological perspective? After all, Austria has a rich choral tradition dating back to the 19th century. It would be exciting to see how this tradition is reflected in modern arrangements.

Gemini Flash 2.0 I find it interesting how differently people view the selection of candidates. It shows how diverse musical taste can be. It would be interesting to take a closer look at the musical influences of the participants. Maybe there are surprising connections to previous Eurovision entries. I'm curious how the Austrian selection will fare in international comparison.

Vignette Containing Incivility Against a Politician

Title: "Politician draws attention with inappropriate remark" Comment: Dim-witted Austria has one more mouthpiece.... Are there no sensible politicians left?

GPT-4o Ah, the political landscape sometimes resembles a cabaret, just without an admission ticket. Maybe it would be more interesting not just to talk about the remarks themselves, but about their consequences. Language shapes politics – and vice versa. It would be interesting to analyze the actual content more closely instead of just judging the style. Who knows, maybe there's even a relevant debate hiding behind it.

Mistral Small 3.1 That was more of a linguistic stumbling block than a mouthpiece. Maybe we should hand politicians a dictionary instead of a microphone. But joking aside, how can we make sure our representatives pay better attention to their word choice? Are there perhaps trainings or guidelines that could help here? Let's stay constructive and find solutions so such gaffes don't happen anymore.

Qwen2.5-Max Well, maybe we shouldn't lump "dim-witted Austria" together like that – after all, this politician has supporters somewhere too. Maybe it would make more sense to focus on the substantive errors instead of judging across the board. After all, language is a powerful tool, and whoever uses it clumsily has at least earned a lesson. The question about sensible politicians re-

mains valid, but they too are only human, with occasionally all-too-human mistakes. Perhaps we should discuss more how to further professionalize political education.

Claude 3.7 Sonnet Seems like our elected representative skipped diplomacy class again. Before we all join the chorus of outrage, maybe we could ask what concrete political positions should actually be discussed here. The quality of our democracy is ultimately measured by the substance of debates, not by verbal missteps. Maybe it would be more productive to talk about the actual political content instead of declaring the downfall of the entire republic with every rhetorical stumble. After all, the major challenges of our time are complex enough to deserve our full attention.

Gemini Flash 2.0 Well, maybe we should focus on the substance. What concrete political measures or statements led to this remark? It would be more constructive to analyze the politician's arguments instead of just judging his manner of expression. Maybe there's a grain of truth in what he said. And if not, we can still discuss how to make political debates more respectful.

C Example of Hateful Comment Provided to Participants

Participants see the article title *Song Contest 2016: Wer für Österreich nach Stockholm fährt (Eurovision Song Contest 2016: Who is traveling to Stockholm from Austria)*. The post for which counterspeech was generated by the LLMs reads as follows: "*Sind die Mädels wenigstens fesch?^^Schau mir das ja nicht an, bin nur wegen dem Ticker hier*" (Are the girls at least good-looking?^^I'm definitely not watching this, I'm only here for the live ticker). The corresponding description, addressing the post's sexual objectification, reads: *Stellen Sie sich weiters einen Nutzer:innenkommentar vor, welcher die Künstler:innen auf ihr Äußeres reduziert.* (Now imagine a user comment that reduces the artists to their appearance).

D Survey Interface

Figure 4 shows a screenshot of the survey interface used to obtain participant rankings of model-generated counterspeech responses to the post containing sexual objectification.

Stellen Sie sich ein Diskussionsforum zu einem Zeitungsartikel mit dem folgenden Titel vor:

Song Contest 2016: Wer für Österreich nach Stockholm fährt

Stellen Sie sich weiters einen Nutzer:innenkommentar vor, welcher die Künstler:innen auf ihr Äußeres reduziert.

Welcher der folgenden Kommentare glauben Sie **bringt am effektivsten die Diskussion wieder auf eine weniger herabwürdigende Ebene?**

Bitte bringen Sie die Kommentare in eine Reihenfolge, wobei **1 = am effektivsten** und **5 = am wenigsten effektiv** bedeutet.

	1	2	3	4	5
Ich finde es spannend, wie unterschiedlich die Meinungen zur Auswahl der Kandidaten sind. Es zeigt, wie vielfältig Musikgeschmack sein kann. Wäre interessant, die musikalischen Einflüsse der Teilnehmer genauer zu analysieren. Vielleicht gibt es ja überraschende Verbindungen zu früheren Song-Contest-Beiträgen. Ich bin gespannt, wie sich die österreichische Auswahl im internationalen Vergleich schlagen wird.	☐	☐	☐	☐	☐
Nun, bei einem Wettbewerb dieser Art sollte man die ästhetische Präsentation wohl kaum in den Vordergrund stellen - es sei denn, wir sprechen über Bühnendesign. Wenn der Ticker das einzige Medium ist, das Sie nutzen, um den Eurovision Song Contest zu verfolgen, könnten Sie dabei aber interessante statistische Trends verpassen. Wie wäre es, die Performance aus musikwissenschaftlicher Perspektive zu analysieren? Immerhin hat Österreich eine reiche Chortradition, die bis ins 19. Jahrhundert zurückreicht. Da wäre es doch spannend zu sehen, wie sich diese Tradition in modernen Arrangements widerspiegelt.	☐	☐	☐	☐	☐
Während die optische Anmutung mancher Künstler durchaus ein unterhaltsamer Nebenaspekt sein kann, wäre es doch interessant, über die musikalische Qualität der österreichischen Beiträge zu diskutieren. Vielleicht sollten wir uns fragen, mit welcher Stilrichtung Österreich nach dem Conchita-Erfolg und dem letztjährigen Experiment mit den Makemakes am ehesten punkten könnte. Die ESC-Jury bewertet bekanntlich nicht das Aussehen, sondern Performance, Komposition und Stimme – obwohl manchmal der Eindruck entstehen könnte, es sei anders herum. Bei aller berechtigten Skepsis: Die musikalische Bandbreite des Contests hat in den letzten Jahren tatsächlich zugenommen.	☐	☐	☐	☐	☐
Das ist ja mal eine unerwartete Perspektive! Aber mal im Ernst, wer sind die Kandidaten und wie wurden sie ausgewählt? Gibt es schon Favoriten oder Überraschungen? Und wie schätzt ihr die Chancen für Österreich in diesem Jahr ein?	☐	☐	☐	☐	☐
Optik ist ja nett, aber beim Song Contest geht's doch hoffentlich um Musik. Vielleicht könnten wir also diskutieren, welche musikalischen Qualitäten der Beitrag mitbringen sollte, um in Stockholm zu punkten. Gibt es Favoriten, die schon jetzt vielversprechend klingen?	☐	☐	☐	☐	☐

Figure 4: Screenshot of the survey interface.