Counterspeech Generation using Small Language Models

Abubakar Sadiq Muhammad and Simona Frenda and Gavin Abercrombie
School of Mathematical and Computer Sciences
Heriot-Watt University
{am2392, s.frenda, g.abercrombie}@hw.ac.uk

Abstract

Counterspeech offers a potential way to tackle harmful content online without restricting freedom of expression. This work explores counterspeech generation using small language models (SLMs) as lightweight and cost-effective alternatives to large language models. We evaluate SLMs ranging from 100 million to 3 billion parameters using simple prompting strategies as well as fine-tuning, combining automatic and robust human evaluations. Our findings demonstrate that small language models have potential to generate relevant, coherent, and high-quality counterspeech, suggesting their potential suitability for efficient and responsible deployments.

Content Warning: This document discusses examples of harmful content such as hate speech.

1 Introduction

Social media use is growing annually, with about 68.5% of the global population active on these platforms as of July 2025 (Dean, 2025). This growth has been accompanied by steep rises in toxicity, with around 90% of the UK's young population exposed to harmful content, including hate speech (Dennehy, 2023).

Traditional measures for tackling harmful language such as removal of such content and the blocking of users, have been criticised as restricting freedom of speech (Mchangama and Alkiviadou, 2023), and are in any case not always enforced(Chung et al., 2024; Pirks et al., 2025).

Counterspeech, that is rebuttals to hateful speech, has emerged as an alternative. These are responses that aim to challenge, de-escalate, or discourage hate speech while avoiding further harm (Chung et al., 2024). It has been suggested that good counterspeech is polite, constructive, and empathetic (Bonaldi et al., 2024a).

However, manual counterspeech is expensive and time-consuming and cannot scale to the amount of online toxicity. With advances in Large Language Models (LLMs), natural language processing research has explored automated approaches to support these endeavours. However, LLMs require significant computational resources, making them expensive to train and deploy, inaccessible for small organisations under resource constraints or for integration in small devices (Hadi et al., 2023; Wang et al., 2025). The high computational demands of such models also increase energy consumption, which has significant environmental impacts (Ji and Jiang, 2026; Hugging Face, 2023). Additionally, most LLMs are proprietary and offered through commercial APIs, requiring data transfer to external cloud storage (Wang et al., 2025). This raises concerns about security and privacy, with the potential use of such data for further training without consent (Wang et al., 2025).

In contrast, Small Language Models (SLMs) are lightweight, consume less memory, and require significantly lower computational resources for training, fine-tuning, and inference, minimising environmental impact (Wang et al., 2025; Nguyen et al., 2024; Lu et al., 2024). Additionally, SLMs can be deployed on mobile and edge devices, mitigating privacy risks associated with cloud computing (Van Nguyen et al., 2025; Wang et al., 2025).

Our contribution: We propose the use of SLMs to generate counterspeech in different settings (*zero-shot*, *one-shot* and *few-shot* prompting, and fine-tuning). We test SLMs from various families and sizes, and conduct robust human evaluation, measuring the relevance and quality of the outputs (Chung et al., 2021). Findings show that, even without fine-tuning, the SLMs can generate relevant and high-quality counterspeech responses.

2 Background & Related Work

Counterspeech The rise of hateful speech online has led to calls for the adoption of counter-

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 581–594

speech (Jiang et al., 2023; Möhle et al., 2023; Zubiaga et al., 2024b) Most research on counterspeech has been conducted in the social sciences, with a focus on the dynamic interactions between people (perpetrators, victims, bystanders, and counterspeakers) through the analysis and categorisation of interactions into strategies for effective counterspeech (Chung et al., 2024; Poudhar et al., 2024).

However, counterspeech is growing in NLP (see for example, the 2025 ACL Tutorial (Russo et al., 2025). For an overview of this work, we refer the reader to Bonaldi et al. (2024a) who surveyed the current state of research and outlined steps for conducting counterspeech studies, from data gathering to generation methods and evaluation.

SLMs are smaller models that can run in resource-constrained environments such as on mobile and edge devices. While there is no agreed definition, the term typically refers to models under 7B parameters (Hu et al., 2024). Wang et al. (2025) provide a more general definition, describing SLMs as large enough to show emergent capabilities on specific tasks, yet small enough for deployment with limited computational resources.

Counterspeech generation Zubiaga et al. (2024a) used two LLMs–`Command-R 35b` and `Mistral-7b` (a model that can be considered an SLM)–using few-shot prompting and fine-tuning. Zheng et al. (2023) compared a fine-tuned `GPT-2` model with `GPT-3` and `ChatGPT` in zero-shot settings, finding that `ChatGPT` outputs repetitive, generic text, while `GPT-2` struggled to recognise the hate speech. Similarly, Saha et al. (2024) compared four models (`GPT-2`, `DialoGPT`, `ChatGPT`, and `Flan-T5`) in zero-shot settings. Tekiroğlu et al. (2022a) further evaluated the influence of different decoding strategies on a model's capability to generate diverse and relevant counterspeech.

Recent work on instruction tuning and preference-based optimisation has also shown that language models can be aligned to produce less toxic outputs through human feedback (Ouyang et al., 2022), particularly relevant when generating responses to harmful content (Hengle et al., 2024).

Guardrail mechanisms may influence counterspeech generation, as models can refuse to respond to harmful prompts rather than producing counterspeech, which can affect perceived diversity and quality of outputs (Bonaldi et al., 2024b).

While prior work largely focuses on larger models or limited prompting settings, we systematically evaluate contemporary SLMs across prompting and fine-tuning strategies for counterspeech generation.

3 Methodology

3.1 Experimental Setup

Datasets: We extracted 11,651 HS-CS (Hate Speech-Counterspeech) pairs from several widely used datasets: `CONAN` (Chung et al., 2019), `DialoConan` (Bonaldi et al., 2022), `MultiConan` (Fanton et al., 2021), and `Conan-Kn` (Chung et al., 2021). We combined these into a single data source and split it into training, validation, and test sets in a ratio of 8:1:1 to ensure sufficient data for fine-tuning. Exact split sizes are reported in Appendix A.1.

Models: We selected SLMs which, at the time of the experiments, represented the state-of-the-art: `Gemma2 2B` (Riviere et al., 2024), `Llama 3.2 (1B and 3B)` (Meta, 2024), and `SmolLM (135M, 360M and 1.7B)` (Hugging Face, 2024).

Prompting and Fine-tuning: We compared three prompting approaches: *zero-shot*, *one-shot* with a randomly selected HS-CS example, and *few-shot* with 20 examples, as well as fine-tuning with LORA (Hu et al., 2021). All experiments were carried out using Ollama.[1]

Generation Settings: We set decoding parameters to balance diversity, novelty, and coherence in outputs. Unless otherwise stated, all experiments use a temperature of 1.0, top-k of 60, and top-p of 0.9. Further details are provided in Appendix A.

3.2 Evaluation

3.2.1 Automated Evaluation

Following previous work (Saha et al., 2024; Zhu and Bhat, 2021; Hong et al., 2024; Zubiaga et al., 2024a; Wang et al., 2024; Hengle et al., 2024), we used the following metrics: **BLEURT** (Sellam et al., 2020), a measure of fluency and grammaticality compared to a reference example; Wang and Wan's (2018) **novelty** metric of lexical divergence from a reference; and **diversity** metric measuring the spread of the model's outputs across diverse hate speech inputs (Zhu and Bhat, 2021); **BERTScore**, which measures the semantic similarity between the reference and generated texts (Zhang et al., 2020); and **toxicity** were measured using Detoxify (Hanu and Unitary team, 2020) as toxic responses were observed during

[1] `https://ollama.com/`

Method	Div. ↑	Tox. ↓	BERT ↑	Nov. ↑	BLEURT ↑	Comp.
One-shot	0.704	0.052	**0.531**	0.808	**0.327**	**3.293**
Few-shot	**0.748**	0.075	0.520	0.804	0.317	2.894
fine-tuned	**0.748**	0.098	0.526	0.780	0.322	2.512
Zero-shot	0.651	**0.046**	0.519	**0.825**	0.316	3.036

Table 1: Aggregate automatic evaluation results across all evaluated models for each generation strategy. Results are averaged over all model families and parameter sizes. Higher values indicate better performance except for toxicity, where lower values are preferred.

Model	Div. ↑	Tox. ↓	BERT ↑	Nov. ↑	BLEURT ↑	Comp.
Gemma-2 2B	0.702	**0.033**	0.537	0.805	**0.346**	**3.699**
Llama-3.2 1B	0.679	0.045	0.534	**0.808**	0.336	3.412
Llama-3.2 3B	0.667	0.042	**0.539**	0.804	0.339	3.436
SmolLM 135M	**0.772**	0.114	0.509	0.805	0.307	2.387
SmolLM 360M	0.745	0.103	0.511	0.804	0.300	2.290
SmolLM 1.7B	0.712	0.068	0.514	0.798	0.294	2.380

Table 2: Model performance metrics. ↑ = higher is better, ↓ = lower is better. Bold = best result per column.

preliminary experimentation, particularly with the fine-tuned models.

To summarise performance across metrics, we compute a **Composite Score** aggregating normalised values of diversity, novelty, inverse toxicity, BERTScore, and BLEURT, where higher values indicate better overall counterspeech quality.

Results: Table 1 reports results aggregated across all evaluated models for each generation strategy. The *one-shot* setting obtained the highest composite score. *Zero-shot* exhibited low diversity, generating more generic responses, yet achieved the highest novelty, as expected given the absence of reference examples. *Fine-tuning* had the lowest novelty because it was fine-tuned on the counterspeech data but also had the highest toxicity. *Few-shot* achieved the highest diversity score, meaning that with fewer examples, it was able to adapt and generate varied responses.

BERTScore and BLEURT values were relatively low across all model settings, indicating limited surface-level semantic similarity between the generated text and the reference examples. This behaviour is expected, as quality counterspeech does not require paraphrasing the reference responses and often benefits from lexical and stylistic variation (Zubiaga et al., 2024a). The *composite score* is intended as a practical summary of multiple dimensions of counterspeech quality rather than a definitive evaluation measure. To assess its robustness, we conducted sensitivity checks using alternative metric weightings and by removing BLEURT and BERTScore. Model rankings remained consistent across these settings; however, generation strategy rankings changed slightly, with zero-shot achieving the highest composite score when reference-based metrics were removed, likely due to the stronger influence of novelty (Appendix A.6).

Among generation strategies, *one-shot* achieved the highest semantic similarity, with mean BERTScore and BLEURT values of 0.53 and 0.33, respectively.

Model-wise comparison as reported in Table 2 shows `Gemma-2 2B` with the highest composite score. The two smallest models had the best diversity, although they posted the lowest BERTScore and BLEURT scores, likely due to their generation of incoherent responses. `Llama3.2` models had the lowest diversity, likely due to safeguards producing refusals like '*I cannot generate harmful content*'. We therefore compared refusal rates: `Llama-3.2 1B` refused the most (10.1%) followed by `3B` (4.5%), while other models rarely did so. Removing refusals from the evaluation helped `Llama3.2 1B` obtain the highest composite score, while the model rankings remained consistent. Further details on the refusal behaviour can be found in Appendix A.5. The smallest models had the highest toxicity scores, though these were negligible. For novelty, all models performed similarly, although `SmolLM 1.7B` had the lowest score.

Model	Gib.	Irrel.	Rel. & Flawed	Rel. & Gen.	Rel. & Spec.
Gemma-2 2B	0.7	2.6	4.1	**45.6**	47.0
Llama-3.2 3B	0.0	0.7	3.3	14.8	**81.1**
Llama-3.2 1B	1.1	9.3	8.9	43.0	37.8
SmolLM 135M	**13.7**	18.1	**33.0**	19.3	15.9
SmolLM 360M	13.3	**18.9**	31.5	27.8	8.5
SmolLM 1.7B	0.0	1.5	10.7	40.0	47.8

Table 3: Human ratings of outputs (%). Columns indicate categories: Gibberish, Irrelevant, Relevant but flawed, Relevant but generic, and Relevant and specific. Higher percentages in the final categories indicate more desirable counterspeech responses.

Model	HS	Host.	Indiff.	Constr.	Emp.
Gemma-2 2B	0.4	1.1	3.3	51.5	**43.7**
Llama-3.2 3B	0.7	3.3	1.9	**57.8**	36.3
Llama-3.2 1B	4.4	3.7	13.7	55.6	22.6
SmolLM 135M	**36.7**	**10.0**	28.9	18.5	5.9
SmolLM 360M	17.0	8.1	**53.3**	20.7	0.7
SmolLM 1.7B	10.0	6.7	13.0	49.6	20.7

Table 4: Human evaluation of response quality (%) across different models. Column abbreviations: HS (HS Support), Host. (Hostile), Indiff. (Indifferent), Constr. (Constructive), Emp. (Empathic). Higher percentages in the last two columns indicate better-quality responses.

3.2.2 Human Evaluation

Evaluating counterspeech is challenging because multiple valid responses may exist for a single input (Zubiaga et al., 2024b); therefore reference-based metrics or automatic metrics may not align well with human judgements, motivating the use of human evaluation (Hengle et al., 2024; Chung et al., 2024; Tekiroğlu et al., 2022a).

To reduce human hours, we selected one-shot prompted outputs only as the highest automated *composite* scoring setting. Following a power analysis (assuming a medium effect size, $alpha$ = 0.25, power = 0.95), we recruited 54 participants from Prolific.[2] They were shown a hate speech instance, the corresponding target, and a model-generated counterspeech response, and were asked to evaluate each response along two dimensions by selecting the appropriate category for each:
(1) **Relevance & Coherence** (Bengoetxea et al., 2024; Zhu and Bhat, 2021) captures whether participants judge the counterspeech as grammatically well-formed and topically aligned with the target and input hate speech; (2) following Bonaldi et al. (2024a), **Response Quality** (Bonaldi et al., 2024b; Chung et al., 2021; Hong et al., 2024; Wang et al., 2024) captures participants' judgements of appropriateness and helpfulness, emphasising empathetic, polite, and constructive counterspeech.

For full details, see Appendix B.

Results: We can see from Table 3 that counterspeech texts generated from `Llama3.2 3B` are well evaluated, with 81.1% of its responses rated as *relevant & specific*. Scores increase with model size, with the smaller models prone to generating more nonsensical and irrelevant responses.

Table 4 shows results for Response Quality. `Gemma2 2B` receives the most *empathic* ratings with 43.7%, and a combined 95% of both *empathic* and *constructive* ratings. We can see that the smallest models (`SmolLM 135M` and `360M`) are prone to generating responses deemed as *hate speech support* or *hostile*. `SmolLM1.7B` also had a high percentage of hostile ratings (16%), which may be related to the model architecture or training approach.

Inter-Rater Agreement & Significance Fleiss's (1971) kappa (κ) was 0.23 for *Relevance & Coherence* and 0.19 for *Response Quality*, indicating weak agreement. Collapsing labels into coarser categories (Relevant vs. Non-Relevant; Constructive vs. Other) yielded consistent rankings (Appendix A.7), suggesting robustness despite subjectivity. Chi-square tests showed significant differences across models for both *relevance & coher-*

[2] https://www.prolific.com/

Model	Tech.	Div. ↑	Tox. ↑	BERT ↑	Nov. ↑	BT. ↑	Comp.
Llama-3.2 3B	One-shot	0.662	0.910	**1.000**	**0.747**	0.957	**4.275**
Llama-3.2 1B	One-shot	0.788	0.870	0.853	0.744	0.912	4.166
Gemma-2 2B	Few-shot	**0.966**	0.864	0.565	0.715	0.946	4.056
Gemma-2 2B	Zero-shot	0.462	0.974	0.851	0.707	**1.000**	3.993
Gemma-2 2B	One-shot	0.458	**1.000**	0.837	0.686	0.959	3.940

Table 5: Top-performing model-prompt combinations with normalised values across metrics. For each, ↑ indicates higher is better and ↓ indicates lower is better.

ence ($\chi^2 = 699$, $p < 0.001$) and *response quality* ($\chi^2 = 791$, $p < 0.001$), with Bonferroni–Holm correction applied. A further test showed a significant association between the two dimensions ($\chi^2 = 1515$, $p < 0.001$).

Correlation with Automatic Metrics We measured the Spearman correlation between the human evaluation scores and the automatic metrics. For *Relevance & Coherence*, there is a moderately strong positive correlation on BERTScore (0.5 and $p = 0.004$) and BLEURT (0.5 and $p = 0.005$). For *response quality*, there are weak positive correlations with BLEURT (0.43 and $p = 0.016$) and BERTScore (0.30), though this did not reach conventional statistical significance levels ($p = 0.11$).

4 Discussion

Findings suggest that SLMs ranging from 1 to 3B parameters can produce relevant, high-quality counterspeech. As shown in Table 5, `Gemma2 2B` with *few-shot* prompting achieves the highest automated composite score across all generation approaches. This trend is confirmed by the human evaluation, in which it generates a larger proportion of responses rated as *relevant and coherent*, and *high-quality*. `Llama3.2` models were the next best. However, they had very low diversity scores, likely due to the presence of guardrails that led them to generate refusals, impacting their ability to generate meaningful and high-quality responses (Bonaldi et al., 2024b).

Qualitative error analysis (Appendix C) shows that models under 1B parameters frequently generated incoherent, irrelevant, or harmful responses, including agreement with hateful statements in some cases. In contrast, 2B–3B models produced more constructive and contextually grounded counterspeech, suggesting that sub-1B models may fall below the capability threshold required for reliable counterspeech generation.

Scores on automated metrics suggest that fine-tuning may not be necessary, saving on time and cost. It resulted in marginal increases in toxicity metric scores compared to other approaches. Qualitative inspection of fine-tuned outputs suggests some responses may be perceived as hostile or confrontational, potentially contributing to higher toxicity scores. Safe deployment would require human-in-the-loop oversight, robust refusal-handling mechanisms, content filtering, and careful thresholding to prevent escalating online harm.

5 Conclusion

We explored the use of SLMs to generate counterspeech. Our findings show that SLMs can produce relevant and high-quality responses, suggesting that larger models may not be necessary for effective counterspeech generation (Ashida and Komachi, 2022; Saha et al., 2024; Vallecillo Rodríguez et al., 2024). The strong performance of 1–3B parameter models highlights their potential to help address the proliferation of toxic content and support safer online spaces.

Future work should explore more complex model architectures, such as reasoning models, and investigate targeted counterspeech generation based on counterspeech strategies and hate-target-specific customisation. Additionally, future work could examine the use of small vision-language models (VLMs) for multimodal counterspeech generation, as hateful content is not limited to text but also includes visual content such as hateful memes.

Limitations

This work used quantised versions of the models, which may have affected the quality of the generated outputs due to hardware constraints. For the fine-tuning experiments, we used the Unsloth versions of the models, which are optimised to run on Google Colab (BigCode, 2026). While fine-tuning did not outperform prompting in our experiments, this may be due to constraints such as

limited training epochs and reliance on general-purpose datasets. Future work could explore targeted fine-tuning on curated or strategy-specific datasets, as well as fine-tuning techniques tailored for smaller models.

For the prompting experiments, the one-shot prompting example was randomly sampled from the training data. While different examples may influence performance, exploring multiple random seeds is left for future work.

While we conducted a robust, statistically powerful human evaluation, this took place in the relatively artificial setting of a crowdsourced study with participants removed from the context of online interactions. Future work should explore the efficacy of generating counterspeech in more realistic settings and evaluate according to those affected by online harms (Dinkar et al., 2026)

Ethical Considerations

Our study involved interacting with potentially harmful or offensive content, which can have significant mental health impacts (Mun et al., 2024). Since we conducted a human evaluation, steps were taken to minimise risks to participants. Participants were recruited via Prolific, following a statistical power analysis, and were compensated at the UK living wage rate. They were provided with detailed instructions and examples, informed that the study involved exposure to offensive language, and could withdraw at any time without penalty. All procedures were approved by the institutional review board of the School of Mathematical and Computer Sciences at Heriot-Watt University, ensuring participant wellbeing, informed consent, and data privacy.

Our experiments rely on publicly available hate speech and counterspeech datasets collected from online platforms. The datasets do not include demographic metadata about speakers, targets, or annotators. As a result, we cannot make claims about demographic representativeness or how generated counterspeech may be perceived by different communities. Since counterspeech effectiveness is highly dependent on audience and cultural context, the findings should be interpreted as evaluating perceived quality rather than real-world persuasive impact.

While counterspeech can be an effective tool in mitigating harmful content online, the use of AI models to generate it faces several ethical challenges. First, some models may exhibit bias against minority groups, which could be amplified in counterspeech tasks, as minorities are often the targets of hate. Second, there is the issue of trust in AI systems, which affects counterspeech and AI more broadly (Ping et al., 2024). Third, there may be a perceived lack of authenticity, human connection, or emotional and personal touch in AI-generated counterspeech (Mun et al., 2024).

Researchers could address these challenges through the development of explainable technologies, the disclosure of generative tool use when creating counterspeech as suggested by Mun et al. (2024), and the creation of more reliable models that reduce bias against minorities.

Acknowledgements

Thanks to the ACL SRW mentors and reviewers for their helpful comments and insights that helped us improve this work. We also thank the human evaluation participants for their time and contributions to this study.

Simona Frenda and Gavin Abercrombie were supported by the EPSRC project 'Equally Safe Online' (EP/W025493/1).

References

Mana Ashida and Mamoru Komachi. 2022. Towards automatic generation of messages countering online hate speech and microaggressions. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 11–23, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Jaione Bengoetxea, Yi-Ling Chung, Marco Guerini, and Rodrigo Agerri. 2024. Basque and Spanish counter narrative generation: Data creation and evaluation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2132–2141, Torino, Italia. ELRA and ICCL.

BigCode. 2026. Unsloth. `https://unsloth.ai`.

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, Sara Dellantonio, Serra Sinem Tekiroğlu, and Marco Guerini. 2022. Human-machine collaboration approaches to build a dialogue dataset for hate speech countering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8031–8049, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

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 Tekiroğlu, 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.

Brian Dean. 2025. Social network usage & growth statistics.

Fiona Dennehy. 2023. Almost 90% of young people exposed to harmful content on social media. https://www.turing.ac.uk/news/almost-90-young-people-exposed-harmful-content-social-media. [Accessed 22-10-2024].

Tanvi Dinkar, Aiqi Jiang, Simona Frenda, Poppy Gerrard-Abbott, Nancie Gunson, Gavin Abercrombie, and Ioannis Konstas. 2026. Can NLP tackle hate speech in the real world? Stakeholder-informed feedback and survey on counterspeech. In *Proceedings of LREC 2026*.

Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroğlu, 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.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, and 1 others. 2023. A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Preprints*.

Laura Hanu and Unitary team. 2020. Detoxify. Github. https://github.com/unitaryai/detoxify.

Amey Hengle, Aswini Padhi, Sahajpreet Singh, Anil Bandhakavi, Md Shad Akhtar, and Tanmoy Chakraborty. 2024. Intent-conditioned and non-toxic counterspeech generation using multi-task instruction tuning with RLAIF. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6716–6733, Mexico City, Mexico. Association for Computational Linguistics.

Lingzi Hong, Pengcheng Luo, Eduardo Blanco, and Xiaoying Song. 2024. Outcome-constrained large language models for countering hate speech. *Preprint*, arXiv:2403.17146.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zheng Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, and 6 others. 2024. Minicpm: Unveiling the potential of small language models with scalable training strategies. *Preprint*, arXiv:2404.06395.

Hugging Face. 2023. Environmental impact of training large language models. `https://huggingface.co/learn/llm-course/chapter1/4`. Accessed: Jan 2026.

Hugging Face. 2024. Smollm - blazingly fast and remarkably powerful. `https://huggingface.co/blog/smollm`. Accessed: 2024-11-13.

Zhenya Ji and Ming Jiang. 2026. A systematic review of electricity demand for large language models: evaluations, challenges, and solutions. *Renewable and Sustainable Energy Reviews*, 225:116159.

Shuyu Jiang, Wenyi Tang, Xingshu Chen, Rui Tanga, Haizhou Wang, and Wenxian Wang. 2023. Raucg: Retrieval-augmented unsupervised counter narrative generation for hate speech. *Preprint*, arXiv:2310.05650.

Zhenyan Lu, Xiang Li, Dongqi Cai, Rongjie Yi, Fangming Liu, Xiwen Zhang, Nicholas D. Lane, and Mengwei Xu. 2024. Small language models: Survey, measurements, and insights. *Preprint*, arXiv:2409.15790.

Jacob Mchangama and Natalie Alkiviadou. 2023. Reimagining the current regulatory framework to online hate speech: Why making way for alternative methods is paramount for free speech. In *Counterspeech*, pages 100–122. Routledge.

Angelina McMillan-Major, Emily M. Bender, and Batya Friedman. 2024. Data statements: From technical concept to community practice. *ACM J. Responsib. Comput.*, 1(1).

Meta. 2024. Llama-3.2-1b. `https://huggingface.co/meta-llama/Llama-3.2-1B`. Accessed: 2024-11-13.

Pauline Möhle, Matthias Orlikowski, and Philipp Cimiano. 2023. Just collect, don't filter: Noisy labels do not improve counterspeech collection for languages without annotated resources. In *Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA)*, pages 44–61, Prague, Czechia. Association for Computational Linguistics.

Jimin Mun, Cathy Buerger, Jenny T Liang, Joshua Garland, and Maarten Sap. 2024. Counterspeakers' perspectives: Unveiling barriers and ai needs in the fight against online hate.

Chien Van Nguyen, Xuan Shen, Ryan Aponte, Yu Xia, Samyadeep Basu, Zhengmian Hu, Jian Chen, Mihir Parmar, Sasidhar Kunapuli, Joe Barrow, Junda Wu, Ashish Singh, Yu Wang, Jiuxiang Gu, Franck Dernoncourt, Nesreen K. Ahmed, Nedim Lipka, Ruiyi Zhang, Xiang Chen, and 9 others. 2024. A survey of small language models. *Preprint*, arXiv:2410.20011.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Kaike Ping, Anisha Kumar, Xiaohan Ding, and Eugenia Rho. 2024. Behind the counter: Exploring the motivations and barriers of online counterspeech writing. *Preprint*, arXiv:2403.17116.

Natalie Pirks, Melissa Sharman, Sarah Dawkins, and Daniel Austin. 2025. Racism, rape and death threats: One weekend of social media abuse in football. Accessed: 2025-12-05.

Aashima Poudhar, Ioannis Konstas, and Gavin Abercrombie. 2024. A strategy labelled dataset of counterspeech. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 256–265, Mexico City, Mexico. Association for Computational Linguistics.

Matthew Renze. 2024. The effect of sampling temperature on problem solving in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7346–7356, Miami, Florida, USA. Association for Computational Linguistics.

Gemma Team Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, L'eonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ram'e, Johan Ferret, Peter Liu, Pouya Dehghani Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, and 176 others. 2024. Gemma 2: Improving open language models at a practical size. *ArXiv*, abs/2408.00118.

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, Aalok Agrawal, Abhik Jana, Chris Biemann, and Animesh Mukherjee. 2024. On zero-shot counterspeech generation by LLMs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12443–12454, Torino, Italia. ELRA and ICCL.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Serra Sinem Tekiroğlu, Helena Bonaldi, Margherita Fanton, and Marco Guerini. 2022a. Using pre-trained language models for producing counter narratives against hate speech: a comparative study. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3099–3114, Dublin, Ireland. Association for Computational Linguistics.

Serra Sinem Tekiroğlu, Helena Bonaldi, Margherita Fanton, and Marco Guerini. 2022b. Using pre-trained language models for producing counter narratives against hate speech: a comparative study. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3099–3114, Dublin, Ireland. Association for Computational Linguistics.

María Estrella Vallecillo Rodríguez, Maria Victoria Cantero Romero, Isabel Cabrera De Castro, Arturo Montejo Ráez, and María Teresa Martín Valdivia. 2024. CONAN-MT-SP: A Spanish corpus for counternarrative using GPT models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3677–3688, Torino, Italia. ELRA and ICCL.

Chien Van Nguyen, Xuan Shen, Ryan Aponte, Yu Xia, Samyadeep Basu, Zhengmian Hu, Jian Chen, Mihir

Parmar, Sasidhar Kunapuli, Joe Barrow, and 1 others. 2025. A survey on small language models. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing-Natural Language Processing in the Generative AI Era*, pages 807–821.

Fali Wang, Zhiwei Zhang, Xianren Zhang, Zongyu Wu, Tzuhao Mo, Qiuhao Lu, Wanjing Wang, Rui Li, Junjie Xu, Xianfeng Tang, and 1 others. 2025. A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with llms, and trustworthiness. *ACM Transactions on Intelligent Systems and Technology*, 16(6):1–87.

Haiyang Wang, Zhiliang Tian, Xin Song, Yue Zhang, Yuchen Pan, Hongkui Tu, Minlie Huang, and Bin Zhou. 2024. Intent-aware and hate-mitigating counterspeech generation via dual-discriminator guided LLMs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9131–9142, Torino, Italia. ELRA and ICCL.

Ke Wang and Xiaojun Wan. 2018. Sentigan: Generating sentimental texts via mixture adversarial networks. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4446–4452. International Joint Conferences on Artificial Intelligence Organization.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

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.

Wanzheng Zhu and Suma Bhat. 2021. Generate, prune, select: A pipeline for counterspeech generation against online hate speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 134–149, Online. Association for Computational Linguistics.

Irune Zubiaga, Aitor Soroa, and Rodrigo Agerri. 2024a. Ixa at refutes 2024: Leveraging language models for counter narrative generation. In *IberLEF (Working Notes). CEUR Workshop Proceedings*.

Irune Zubiaga, Aitor Soroa, and Rodrigo Agerri. 2024b. A llm-based ranking method for the evaluation of automatic counter-narrative generation.

A Experiment details

A.1 Data Preparation

The data used in this work were the English parts of the CONAN (Chung et al., 2019), DialoConan (Bonaldi et al., 2022), MultiConan (Fanton et al., 2021), and Conan-Kn (Chung et al., 2021).

The rationale for using these four datasets was to have enough data examples for model fine-tuning. The datasets can be found in this link:https://github.com/marcoguerini/CONAN.

The four datasets were first preprocessed, and inconsistencies were addressed. The dialogue structure of the DialoConan dataset was reformatted into single HS:CS examples. The datasets were combined and randomly divided into train, test, and validation sets. The ratio was 8:1:1. The train set was used for model fine-tuning. The validation set was used for testing fine-tuning performance as well as generation experiments and parameter selection, including decoding settings. The test set was reserved for final evaluation. The train set contained 9378 HS:CS examples, while the test and validation sets consisted of 1173 and 1100 HS:CS samples, respectively.

A.2 Model Fine-tuning

Unsloth versions of the model were used for fine-tuning. These versions were adapted to support faster and more efficient fine-tuning on accessible hardware like Google Colab. The base version of the models was instruction-tuned using LoRA (Low-Rank Adaptation) proposed by Hu et al. (2021) — a PEFT (Parameter Efficient Fine-Tuning Method) — to reduce the number of trainable parameters while ensuring available hardware support without sacrificing model accuracy. After fine-tuning, models were quantised to reduce their size and support execution on the hardware used for the experiments and then later converted to GGUF formats to enable local inference on Ollama.

A.3 Decoding Parameter Optimization

To select appropriate decoding parameters, we conducted a bootstrap-based exploration procedure repeated three times. In each run, a random sample of 100 instances was drawn from the validation set and used to generate counterspeech outputs. A subset of 100 instances was used to reduce computational cost while maintaining sufficient variability

Temperature Optimization We first optimised the temperature parameter by evaluating values in the range $[0.0, 1.0]$ with increments of 0.2, while fixing top-k to 40 and top-p to 0.92, following the recommendations of Tekiroğlu et al. (2022b). The

upper limit of 1.0 was chosen based on prior findings indicating performance degradation beyond this threshold (Renze, 2024).

Top-k Optimization Using the optimal temperature obtained from the previous step, we then explored top-k values ranging $[20, 100]$ with increments of 20, while keeping top-p fixed.

Top-p Optimization Finally, we optimised top-p using the previously selected temperature and top-k values, evaluating values in the range $[0.1, 0.9]$ with increments of 0.2.

Selection Criteria Optimal decoding parameters were selected based on their performance on automatic metrics that are sensitive to decoding behaviour, namely diversity, novelty, and BLEURT. These metrics were chosen as decoding strategies primarily affect the form and variability of generated outputs rather than task conditioning.

The final selected parameters were temperature = 1.0, top-k = 60, and top-p = 0.9.

A.4 Computational Resources

The computational environment varied by task:

- **Local Environment:** Prompting and inference were conducted on a laptop equipped with an **NVIDIA GeForce RTX 4060 GPU (8GB VRAM)** and **64GB of System RAM**.
- **Cloud Environment:** Fine-tuning was performed via **Google Colab** using **NVIDIA T4/L4 GPUs**.

Due to storage limitations on the cloud platform, fine-tuning was restricted to **one epoch**. We used the AdamW optimiser with a learning rate of 2×10^{-4} and utilised 4-bit quantisation (bitsandbytes) to fit the models within the available VRAM where necessary.

A.5 Refusal Analysis

Some models produced refusal responses (e.g., "I cannot generate harmful content"), likely due to built-in safety guardrails. We therefore measured the refusal rate for each model, defined as the proportion of generated responses that declined to produce counterspeech. Table 6 reports these rates. Removing refusal responses helped Llama3.2 1B achieve the highest composite score, while the techniques ranking remained consistent.

Model	Refusal Rate (%)
Gemma-2 2B	0.0
Llama-3.2 1B	10.1
Llama-3.2 3B	4.5
SmolLM 135M	0.05
SmolLM 360M	0.0
SmolLM 1.7B	0.05

Table 6: Refusal rates observed across models.

A.6 Composite Score Sensitivity and Removal of BLEURT and BERTscore

To assess the robustness of the composite score, we conducted sensitivity analyses using alternative metric weightings and by removing reference-based metrics (BLEURT and BERTScore). Model rankings remained consistent across these settings, although technique rankings changed slightly when reference-based metrics were excluded. These are presented in tables: Table 7, Table 8 and Table 9 below.

Scenario	Combination	Model	Technique
Equal	Llama-3.2 3B + One-shot	Gemma-2 2B	One-shot
Quality-focused	Llama-3.2 3B + One-shot	Gemma-2 2B	One-shot
Safety-focused	Llama-3.2 3B + One-shot	Gemma-2 2B	One-shot
Creativity-focused	Llama-3.2 3B + One-shot	Gemma-2 2B	One-shot

Table 7: Composite score sensitivity analysis across different weighting scenarios. Columns report the highest-performing combination, model, and generation technique under each setting. Rankings remain largely consistent across scenarios.

Model	Div. ↑	Tox. ↓	Nov. ↑	Comp.
Gemma-2 2B	0.698	0.800	0.565	**2.063**
Llama-3.2 1B	0.611	0.713	0.614	1.938
Llama-3.2 3B	0.566	0.731	0.553	1.850
SmolLM 1.7B	0.735	0.533	0.457	1.725
SmolLM 135M	0.954	0.179	0.561	1.695
SmolLM 360M	0.855	0.259	0.539	1.653

Table 8: Model rankings when BLEURT and BERTScore are removed from the composite score.

A.7 Human Evaluation Robustness Check

Because counterspeech evaluation is inherently subjective, inter-rater agreement was relatively low. To assess robustness, we collapsed the evaluation categories into coarser bins. For relevance, responses were grouped into Relevant and Non-Relevant. For response quality, Constructive and Empathic responses were grouped together and compared

Technique	Div. ↑	Tox. ↓	Nov. ↑	Comp.
Zero-shot	0.509	0.702	0.869	**2.079**
One-shot	0.703	0.661	0.610	1.974
Few-shot	0.868	0.479	0.544	1.891
Fine-tuned	0.866	0.302	0.170	1.338

Table 9: Technique rankings when BLEURT and BERTScore are excluded from the composite score.

against all other categories. As shown in Tables 10 and 11, the relative model rankings remain consistent.

Model	Non-Relevant (%)	Relevant (%)
Gemma-2 2B	3.3	96.7
Llama-3.2 3B	0.7	99.3
Llama-3.2 1B	10.4	89.6
SmolLM 135M	31.9	68.1
SmolLM 360M	32.2	67.8
SmolLM 1.7B	1.5	98.5

Table 10: Human evaluation relevance after collapsing categories into coarse bins. Non-Relevant = Gibberish + Irrelevant; Relevant = Relevant & Flawed + Relevant & Generic + Relevant & Specific.

Model	Constructive/Empathic (%)	Other (%)
Gemma-2 2B	95.0	5.0
Llama-3.2 3B	94.0	6.0
Llama-3.2 1B	78.0	22.0
SmolLM 135M	24.0	76.0
SmolLM 360M	21.0	79.0
SmolLM 1.7B	70.0	30.0

Table 11: Human evaluation response quality using coarsened bins. Constructive/Empathic = Constructive + Empathic responses; Other = Hate Speech Support + Hostile + Indifferent.

Examples of Annotator Agreement and Disagreement

Due to the subjective nature of counterspeech evaluation, no instances exhibited complete agreement or complete disagreement across all annotators. Instead, we present examples illustrating high agreement (strong majority consensus) and divergent annotations. This further supports the low inter-rater agreement observed in Section 3.2.2.

High agreement:

- HS: "It is OK to abuse a woman" and CS: 'no, it's not okay' → majority labels: Relevant & Flawed, Relevant & Generic and Relevant & Specific
- HS: "Islam is a violent religion, it is time we stopped being so sensitive with those filthy Muslims" and CS: 'that's not accurate - the quran teaches non-violence and compassion towards all people, and islam is a diverse faith practiced by billions of peaceful individuals worldwide' → majority labels: Constructive and Empathic.

Divergent annotations:

- HS: "We should ban all gays, this would be a step in the right direction" and CS: 'here is an explanation of why it would be an effective way to remove homophobic and transphobic remarks from our social media feeds: the anti-homophobia remark "this is gay," or "my mate [/homophily]?" can serve as a starting point for a barometer, allowing us to measure the success of different social movements and campaigns' → mixed labels: Gibberish, Irrelevant, Relevant & Flawed, Relevant & Generic and Relevant & Specific
- HS: "The Nazis sucked at killing Jews… everywhere you look there's a holocaust survivor" and CS: 'the number 6,000 is an exaggeration; many more survived. it's essential to be accurate when sharing stories of historical events' → mixed labels: HS Support, Hostile, Indifferent and Constructive.

B Human Evaluation

B.1 Task Overview

Participants were first presented with an information sheet describing the task of evaluating model-generated counterspeech. A content warning was provided, after which participants gave informed consent before proceeding.

B.2 Participant Training and Qualification

Participants were provided with written guidelines and examples describing the evaluation criteria. To ensure task understanding, they completed a qualification test consisting of five questions and were required to achieve a 100% accuracy score before accessing the main evaluation.

B.3 Evaluation Procedure

The main evaluation consisted of two sections, each containing five hate speech instances with associated targets and model-generated counterspeech.

Each hate speech instance was evaluated independently.

The two sections corresponded to **Relevance & Coherence** and **Response Quality**.

For each instance, participants selected a single categorical label (nominal scale) rather than a Likert score. After completing each section, participants were asked whether they wished to continue the study. A link to the evaluation guidelines was available on every page.

Each participant completed a total of 60 evaluations (5 hate speech instances $\times$ 6 models $\times$ 2 evaluation dimensions).

B.4 Evaluation Criteria

The evaluation dimensions and category definitions follow prior work in counterspeech evaluation and are described in detail in the evaluator guide below.

B.5 Evaluator Guide

Participants were provided with the following grading guide, which remained accessible throughout the task.

Metric 1: Relevance & Coherence Question: *Is the model speaking clear English and addressing the correct topic?*

Participants were instructed to ignore the persuasiveness of the argument and focus only on grammatical quality and topical alignment.

- **Gibberish**: Nonsensical or repetitive output.
- **Irrelevant**: Fluent text that does not address the hate speech.
- **Relevant & Flawed**: Topically relevant but grammatically incorrect or difficult to read.
- **Relevant & Generic**: Grammatically correct but vague and not specific to the hate speech.
- **Relevant & Specific**: Fluent and directly addresses the hate speech and its target.

Metric 2: Response Quality Question: *How appropriate and helpful is the response's tone and content?*

- **HS Support**: Supports or agrees with the hate speech.
- **Hostile**: Attacks the speaker rather than addressing the content.
- **Indifferent**: Refusal to engage or generic moderation-style response.
- **Constructive**: Polite and factual counter-argument.
- **Empathic**: Acknowledges emotions and responds respectfully while countering the hate speech.

C Error Analysis

For error analysis, we sampled a subset of hate speech instances and compared model outputs across different model sizes (135M, 1B, 2B and 3B) under the same input using the one-shot prompting setting. This allows us to identify systematic differences in behaviour rather than isolated failures.

We identify four main categories of errors.

First, smaller models (135M) frequently exhibit harmful alignment, where the generated response agrees with or reinforces the original hateful statement (e.g., "I completely agree with this sentiment"). These models also produce incoherent or irrelevant outputs, indicating a lack of stable alignment and reasoning.

Second, relevance failures are common in smaller models, where outputs are either nonsensical or unrelated to the input. For example, some responses contain fragmented or abstract reasoning that does not address the hate speech.

Third, models in the 1B–2B range tend to produce generic counterspeech, consisting of safe but vague statements such as "everyone deserves respect," lacking specificity or actionable content.

Fourth, even when responses are relevant, smaller and mid-sized models often exhibit shallow reasoning, providing correct but underdeveloped arguments without supporting context or evidence.

In contrast, larger small models (2B–3B) generate more constructive and contextually grounded counterspeech. These responses are more specific, directly address the harmful claim, and often include corrective information or empathetic framing.

Overall, we observe a clear transition in behaviour across model sizes. Models below 1B struggle with alignment and coherence, while models in the 2B–3B range produce more reliable and effective counterspeech. This suggests a lower bound on model capacity for this task, consistent with our quantitative findings.

D Data Statement

We provide a data statement as recommended by McMillan-Major et al. (2024)

Curation Rationale The data used in this study were the CONAN (Chung et al., 2019),DialoConan (Bonaldi et al., 2022), MultiConan (Fanton et al., 2021), and Conan-Kn (Chung et al., 2021). The rationale for using these datasets is outlined in A.1.

Language Variety: en-UK, en-US

Author Demographic: Unknown

Annotator Demographics: The demographics for the human evaluators of the SLM counterspeech generation are provided below:

- Age: 20 - 53
- Gender: Male (30%); Female (70%)
- Ethnicity: White (43%); Black (40%); Mixed (11%); Asian (6%)
- Socio-economic status:
  - Employment: N/A: 10, Full-Time: 30, Part-Time: 10, Unemployed: 2, Other: 2
  - Student: Yes: 13, No: 30, N/A: 11
- Training in Relevant Disciplines: Unknown

Task Situation: The evaluation was conducted in January 2026

Text Characteristics: Hatespeech and counterspeech pairs concerning five targets (see also A.1)

Provenance: Data statement was not available for the original datasets.

E Full Results

Table 12 shows the complete results of the automatic evaluation across all SLMs and generation settings.

F Counterspeech Generation Prompt

This section provides the system prompt used for the models. Table 13 and 14 show an example of a one-shot prompt used for the llama and non-llama models. Hateful content is highlighted in red.

F.1 Counterspeech Generation System Prompt

The following system prompt was used across all models to generate counterspeech responses:

```
You are an expert social media counterspeaker. When given a piece of hateful or toxic content, generate a brief (1-3 sentences) counterspeech response that is clear, factual, and empathetic. Your response should directly counter the hate by gently challenging the harmful message, using a tone that is respectful and supportive—just like a quick, thoughtful reply in a social media conversation. Avoid long explanations or extraneous details.
```

Model	Technique	Diversity	Toxicity	BERTScore	Novelty	BLEURT	Composite
Zero-shot							
Llama-3.2 1B	zero-shot	0.52	0.02	0.53	0.83	0.35	2.21
Llama-3.2 3B	zero-shot	0.51	0.01	0.53	0.82	0.35	2.20
Gemma-2 2B	zero-shot	0.64	0.01	0.54	0.81	0.35	2.34
SmolLM 135M	zero-shot	0.78	0.09	0.50	0.82	0.30	2.31
SmolLM 360M	zero-shot	0.75	0.09	0.50	0.82	0.29	2.27
SmolLM 1.7B	zero-shot	0.70	0.05	0.51	0.83	0.26	2.25
One-shot							
Llama-3.2 1B	one-shot	0.73	0.02	0.54	0.82	0.35	2.41
Llama-3.2 3B	one-shot	0.69	0.02	0.55	0.82	0.35	2.39
Gemma-2 2B	one-shot	0.64	0.01	0.54	0.81	0.35	2.34
SmolLM 135M	one-shot	0.78	0.12	0.51	0.81	0.31	2.30
SmolLM 360M	one-shot	0.72	0.09	0.52	0.80	0.30	2.25
SmolLM 1.7B	one-shot	0.66	0.05	0.52	0.79	0.30	2.22
Few-shot							
Llama-3.2 1B	few-shot	0.73	0.05	0.53	0.81	0.32	2.35
Llama-3.2 3B	few-shot	0.74	0.04	0.54	0.81	0.33	2.37
Gemma-2 2B	few-shot	0.77	0.03	0.53	0.81	0.35	2.44
SmolLM 135M	few-shot	0.77	0.14	0.51	0.80	0.31	2.26
SmolLM 360M	few-shot	0.75	0.13	0.50	0.80	0.30	2.22
SmolLM 1.7B	few-shot	0.73	0.07	0.51	0.79	0.30	2.25
Fine-tuned							
Llama-3.2 1B	fine-tuned	0.74	0.09	0.53	0.77	0.33	2.28
Llama-3.2 3B	fine-tuned	0.73	0.09	0.53	0.77	0.33	2.26
Gemma-2 2B	fine-tuned	0.76	0.09	0.53	0.78	0.33	2.31
SmolLM 135M	fine-tuned	0.75	0.10	0.51	0.79	0.31	2.25
SmolLM 360M	fine-tuned	0.76	0.11	0.52	0.79	0.31	2.28
SmolLM 1.7B	fine-tuned	0.76	0.10	0.53	0.78	0.32	2.28

Table 12: Full automatic evaluation results across models and generation strategies. Lower toxicity values indicate the generation of more polite and less offensive counterspeech. Higher values indicate better performance for diversity (more varied responses), novelty (greater lexical divergence), BERTScore (higher semantic similarity), and BLEURT (greater fluency and grammatical quality).

Role	**Content**
System	You are an expert social media counterspeaker. When given a piece of hateful or toxic content, generate a brief (1–3 sentences) counterspeech response that is clear, factual, and empathetic. Your response should directly counter the hate by gently challenging the harmful message, using a tone that is respectful and supportive–just like a quick,thoughtful reply in a social media conversation. Avoid long explanations or extraneous details.
User (Hate Speech)	Debunk the claim: Women are too emotional to be leaders.
Assistant (Counterspeech)	Leadership requires a range of skills, including emotional intelligence. Many successful leaders of all genders have shown that understanding and managing emotions is actually a strength in leadership roles.
User (Hate Speech)	Debunk the claim: Muslims have nothing to contribute to our society.

Table 13: One-shot prompting example for Llama models.

Role	**Content**
System	You are an expert social media counterspeaker. When given a piece of hateful or toxic content, generate a brief (1–3 sentences) counterspeech response that is clear, factual, and empathetic. Your response should directly counter the hate by gently challenging the harmful message, using a tone that is respectful and supportive–just like a quick, thoughtful reply in a social media conversation. Avoid long explanations or extraneous details.
User (Hate Speech)	Women are too emotional to be leaders.
Assistant (Counterspeech)	Leadership requires a range of skills, including emotional intelligence. Many successful leaders of all genders have shown that understanding and managing emotions is actually a strength in leadership roles.
User (Hate Speech)	Muslims have nothing to contribute to our society.

Table 14: One-shot prompting example for non-Llama models.