Is Explanation All You Need? An Expert Survey on LLM-generated Explanations for Abusive Language Detection

Chiara Di Bonaventura[1,2,*,†], Lucia Siciliani[3], Pierpaolo Basile[3], Albert Meroño-Peñuela[1] and Barbara McGillivray[1]

[1]King's College London, London, United Kingdom
[2]Imperial College London, London, United Kingdom
[3]Department of Computer Science, University of Bari Aldo Moro, Italy

Abstract

Explainable abusive language detection has proven to help both users and content moderators, and recent research has focused on prompting LLMs to generate explanations for why a specific text is hateful. Yet, understanding the alignment of these generated explanations with human expectations and judgements is far from being solved. In this paper, we design a before-and-after study recruiting AI experts to evaluate the usefulness and trustworthiness of LLM-generated explanations for abusive language detection tasks, investigating multiple LLMs and learning strategies. Our experiments show that expectations in terms of usefulness and trustworthiness of LLM-generated explanations are not met, as their ratings decrease by 47.78% and 64.32%, respectively, after treatment. Further, our results suggest caution in using LLMs for explanation generation of abusive language detection due to (i) their cultural bias, and (ii) difficulty in reliably evaluating them with empirical metrics. In light of our results, we provide three recommendations to use LLMs responsibly for explainable abusive language detection.

Keywords

Large Language Models, Hate Speech Detection, Explanation Generation, Human Evaluation

1. Introduction

Explainability is a crucial open challenge in Natural Language Processing (NLP) research on abusive language [1] as increasing models' complexity [2], models' intrinsic bias [3], and international regulations [4] call for a shift in perspective from performance-based models to more transparent models. Moreover, recent studies have shown the benefits of explanations for users [5, 6] and content moderators [7] on social media platforms. The former can benefit from receiving an explanation for why a certain post has been flagged or removed whereas the latter are shown to annotate toxic posts faster and solve doubtful annotations thanks to explanations.

Several efforts have moved towards explainable abusive language detection in the past years, like the development of datasets containing rationales (i.e., the tokens in the text that suggest why the text is hateful) [8] or implied statements (i.e., description of the implied meaning of the text) [9, 10], and shared tasks on explainable hate speech detection [11, 12], *inter alia*. With Large Language Models (LLMs) like FLAN-T5 [13] showing remarkable performance across tasks and human-like text generation [14, 15, 16], recent studies have explored LLMs for explainable hate speech detection, wherein classification predictions are described through natural language explanations [17, 18]. For instance, [19] used chain-of-thought prompting [20] of LLMs to generate explanations for implicit hate speech detection.

However, most of these studies rely on empirical metrics like BLEU [21] to evaluate the generated explanations automatically. Consequently, the human perception and implications of these explanations remain understudied, as well as the extent to which empirical metrics approximate human judgements. [22] recruited crowdworkers to evaluate the level of hatefulness in tweets and the quality of explanations generated by GPT-3. Instead, we conduct an expert survey investigating four LLMs and five learning strategies across multi-class abusive language detection tasks to answer the following questions: **RQ1:** How well do LLM-generated explanations for abusive language detection match human expectations? **RQ2:** How well do empirical metrics align with human judgements? **RQ3:** What makes LLM-generated explanations good, according to experts?

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy
*Corresponding author.
†Work partially funded by the Trustworthy AI Research award received by The Alan Turing Institute and the the Italian Future AI Research Foundation (FAIR).
chiara.di_bonaventura@kcl.ac.uk (C. D. Bonaventura); lucia.siciliani@uniba.it (L. Siciliani); pierpaolo.basile@uniba.it (P. Basile); albert.merono@kcl.ac.uk (A. Meroño-Peñuela); barbara.mcgillivray@kcl.ac.uk (B. McGillivray)
0000-0002-1438-280X (L. Siciliani); 0000-0002-0545-1105 (P. Basile); 0000-0003-4646-5842 (A. Meroño-Peñuela)

2. Experimental Setup

To answer these research questions, we design a before-and-after study, surveying participants about their prior expectations about LLM-generated explanations and then showing them examples generated by several LLMs with diverse learning strategies[1], followed by further interviews. To ensure robustness of our results, we recruited experts in the field, i.e., AI researchers, as described below.

2.1. Data

For our experiments, we use the HateXplain [8] and the Implicit Hate Corpus [9] as they encompass different levels of offensiveness (i.e., hate speech, offensive, neutral), expressiveness (i.e., explicit hate, implicit hate, neutral), multiple targeted groups, and explanations for the hateful label (Table 1). These datasets contain unstructured explanations of the words that constitute abuse (in HateXplain) and the user's intent (in Implicit Hate). In view of previous research arguing the need for structured explanations in hateful content moderation [1], we use the following template to create structured explanations, that we will use as ground-truth: "*Explanation: it contains the following hateful words (implied statement):*" for abusive content in HateXplain (Implicit Hate Corpus), and "*The text does not contain abusive content.*" for neutral content.

Dataset	Labels	Target	Explanation
HateXplain	hate speech, offensive, neutral	women, black, ...	Token-level
Implicit Hate	implicit hate, explicit hate, neutral	Jews, whites, ...	Implied statement

Table 1
Summary of datasets used.

2.2. Methodology

We extensively investigate four popular LLMs across five learning strategies on their ability to detect multi-class offensiveness and expressiveness of abusive language and to generate explanations for the classification.

Models. We use different open-source LLMs (Table 2): the base versions of **FLAN-Alpaca** [23, 24], **FLAN-T5** [13], **mT0** [25], and the 7B foundational model **Llama 2** [26], which is an updated version of LlaMA [27].

Model	Instruction Fine-tuned	Toxicity Fine-tuned
FLAN-Alpaca	☒	☒
FLAN-T5	☒	☒
mT0	☒	-
Llama-2	-	-

Table 2
Summary of models used.

Learning strategies. As different prompting strategies might yield different results, we test five distinct learning strategies using the established Stanford Alpaca template[2] (cf. Appendix A for prompt details):

(1) zero-shot learning (zsl): we pass *"Classify the input text as* `list_of_labels`*, and provide an explanation"* in the instruction field of the template. The `list_of_labels` changes according to the dataset used;

(2) few-shot learning (fsl): we pass three additional examples to the aforementioned template, which are randomly sampled with equal probability among the labels to account for class imbalance in the datasets. We experimented with different numbers of examples (i.e., passing one, three or five examples), and chose three as it was the best strategy;

(3) knowledge-guided zero-shot learning (kg): instead of passing additional examples in the prompts, we add external knowledge retrieved by means of an entity linker[3], which first detects entities mentioned in the input text, and then retrieves the relevant information from the external knowledge base. We use Wikidata [28] for encyclopedic knowledge, KnowledJe [29] for hate speech temporal linguistic knowledge and ConceptNet [30] for commonsense knowledge. We modify the prompt template with an additional field called 'context' to account for this external knowledge;

(4) instruction fine-tuning (ft): we use the same prompts used in (1) to instruction fine-tune Llama-2;

(5) knowledge-guided instruction fine-tuning (kg_ft): we use the knowledge-guided prompts developed in (3) to instruction fine-tune Llama-2.

Empirical eval metrics. We evaluate how closely the LLM-generated explanations match the ground-truth across eight empirical similarity metrics due to the challenge of simultaneously assessing a wide set of criteria [31, 32, 33]. Following established NLG research [34, 35], we choose BERTScore [36] and METEOR [37] for semantic similarity. For syntactic similarity, we select BLEU [21], GBLEU [38], ROUGE [39], ChrF [40] with

[1]The data containing the LLM-generated explanations are publicly available at https://github.com/ChiaraDiBonaventura/is-explanation-all-you-need

[2]https://github.com/tatsu-lab/stanford_alpaca?tab=readme-ov-file#data-release

[3]If available, we use the API provided by the knowledge source, spaCy otherwise. https://spacy.io/

its derivates ChrF+ and ChrF++ [41, 42]. Additionally, we present an expert evaluation following our survey.

2.3. Survey Design

To evaluate how well LLMs align with human expectations and judgements in explanation generation, we design a before-and-after study as follows.

Before treatment. We ask for participant's background information, e.g., gender identity, native language and how they would rate the usefulness and trustworthiness of a language model for explanation generation. Specifically, we ask "How useful would you rate a system that provides you a textual explanation for its classification with respect to receiving only its classification?" and "How trustworthy would you rate a system that provides you a textual explanation for its classification with respect to receiving only its classification?" on a 1-5 Likert scale.

Treatment. As for the treatment, we show participants a sample of 70 texts from the datasets, paired with up to four different explanations. Specifically, given a text and ground-truth explanation, participants are asked if the text is correctly explained. If yes, they are asked to rate three different LLM-generated explanations with respect to the ground-truth on a 1-3 scale. These explanations are randomly sampled among the four LLMs and five learning strategies discussed in Section 2.2.

After treatment. Finally, we ask participants' opinion on the usefulness and trustworthiness of explanation generation, having seen the LLM-generated explanations. In addition, we ask general opinions related to what type of errors they observed most frequently, and what a good explanation would look like.

The full list of questions is in the Appendix B. The institutional ethical board of the first author's university approved our study design. We distributed the survey through channels that allow us to target individuals working in AI who are familiar with the field of language models and/or AI Ethics, including NLP reading groups and AI Ethics interest groups. To ensure the reliability of our before-and-after study, participants were given 1 hour to complete as many answers as they could. We collected answers from 15 participants, of which 33% (67%) identify as female (male), and 33% (67%) are (non) English native-speakers. The average level of participants' expertise in abusive language research is 2.47 out of 5 (self-described)[4], and their continents of origin include Europe (60%), Asia (26.67%), Africa (6.67%), and Latin America (6.67%).

[4]The list of levels to choose from was: 1=Novice, 2=Advanced beginner, 3=Competent, 4=Proficient, 5=Expert.

3. Results and Discussion

Our 15 participants reach a fair agreement, with Krippendorff's alpha [43] equal to 38.43%.

Fig. 1 shows changes in the relative frequencies of participant scores in the usefulness and trustworthiness of explanations before and after treatment. Participants' responses before treatment have expectations of textual explanations for classifications of being "highly useful" (above 50%; highest possible score) in terms of usefulness, and "moderately trustworthy" or "neutral" (above 40%; second and third best possible score) in terms of trustworthiness. However, scores for after treatment show participants changing their usefulness scores towards "moderately unuseful" (40-50%; second worst possible score) and their trustworthiness scores to "highly untrustworthy" (above 30%; worst possible score). Agreement differs in each category: usefulness is much more consensual, whereas trustworthiness is judged with higher variance. In general, LLM-generated explanations do not meet human expectations in terms of usefulness and trustworthiness. Specifically, exposing participants to these explanations leads to an average percentage decrease of 47.78% and 64.32% in the perception of the usefulness and trustworthiness of explanations, respectively.

Fig. 2 shows the scores of all empirical metrics and expert evaluation for all models on explanation generation. Overall, similarity metrics tend to be highly volatile with respect to each other. For instance, FLAN-Alpaca prompted with zero-shot learning (i.e., 'alpaca_zsl' in the figure) generates explanations that are more than 70% semantically similar to the ground-truth explanations according to BERTScore while being less than 20% semantically similar according to METEOR. Similarly for syntax: BLEU and GBLEU similarity scores are less than 3% whereas ROUGE and chrF/+/++ are in the range 9%-21%. Moreover, we observe that BERTScore has a tendency to over-score explanations compared to human evaluation scores. Contrarily, METEOR, BLEU, GBLEU, ROUGE and chrF/+/++ have a tendency to under-score explanations. Instruction fine-tuning helped all metrics to approximate expert evaluations better, especially when tuned on knowledge-guided prompts. We use the Spearman's rank correlation coefficient to compare the correlation between human scores and those provided by all the other metrics. In detail, we rank the models for each type of metric, and then we compute the Spearman correlation between the rank obtained by human scores and those obtained by other metrics. Table 3 reports all the correlation scores. We observe that BERTScore is the most correlated with humans in both tasks. Also,

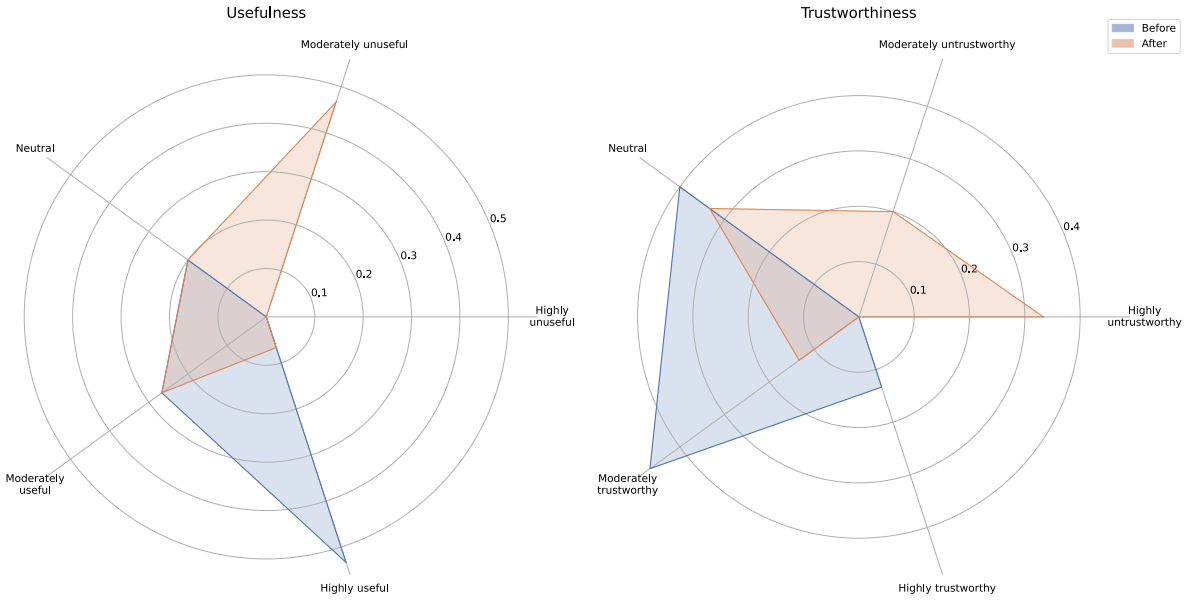

Figure 1: Relative frequencies of Likert scores before and after treatment on usefulness and trustworthiness of LLMs for explanation generation in abusive language detection.

chrF/+/++ metrics are highly correlated with humans while all the other metrics based on syntactic matches are slightly correlated with humans. Results show that semantic metrics are more similar to how humans evaluate the quality of the explanation generated by LLMs. Only one metric (ROUGE) shows a different behaviour between the two tasks.

Since 38.55% of the ground-truth explanations were not rated as good explanations by participants, we further investigated what are the most common errors and what makes an explanation good. Table 4 returns the most common error categories reported by participants. Most of them are related to logical fallacies (e.g., contradictory statements, hallucination), especially in the context of sarcasm and self-deprecating humour, rather than linguistic errors (e.g., grammar, misspellings). It is worth noticing that 13.33% of the participants reported that LLM-generated explanations contain cultural bias (e.g., stereotypes), with the implication of potentially perpetuating harms against the targeted victims of abusive language. As for desiderata, 73.33% of participants would like to receive textual explanations that are coherent with human reasoning and understanding, i.e., that are relevant and exhaustive to the text they refer to while being logically and linguistically correct. A remaining 20% thinks that a good explanation must be coherent with model reasoning instead. In other words, participants are much more concerned about how the explanation looks like rather than its reflection of the inner mechanism of the model reasoning. To quote a participant's perspective, "*I would want the explanation to be helpful to me and guide my own reasoning*".

Metric	Spearman Coeff.	
	Implicit Hate	HateXplain
bertscore	0,80	0,91
meteor	0,64	0,89
chrf1	0,60	0,83
chrf2	0,60	0,81
chrf	0,57	0,83
gbleu	0,53	0,25
rouge	0,50	0,86
bleu	0,27	0,11

Table 3
The Spearman coefficient between each metric and experts' scores.

Error Category	Relative Frequency
Logical Errors	26.67%
Vagueness	20.00%
Cultural Bias	13.33%
Hallucination	13.33%
Irrelevant Info	13.33%
Other	6.67%

Table 4
Percentage of error categories reported by participants.

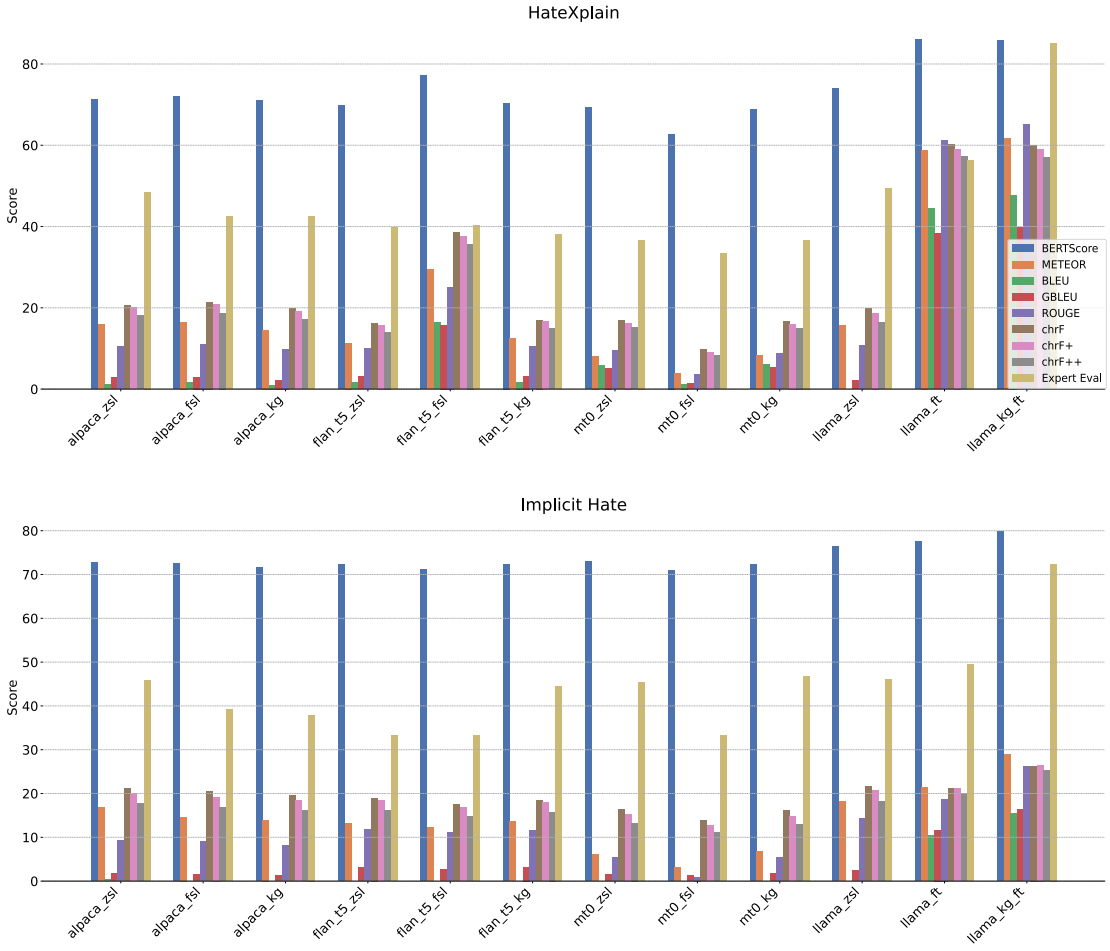

Figure 2: Evaluation of explanation generation by LLMs across empirical metrics and human eval.

4. Conclusion

In this paper, we conducted a before-and-after study to understand human expectations and judgements of LLM-generated explanations for multi-class abusive language detection tasks. Contrarily to previous research [22], we investigated multiple LLMs and learning techniques, and we surveyed AI experts who are familiar with abusive language research instead of crowdworkers. We found that human expectations in terms of usefulness and trustworthiness of LLM-generated explanations are not met: after seeing these explanations, the usefulness and trustworthiness ratings decrease by 47.78% and 64.32%, respectively. Secondly, our results show that empirical metrics commonly used to evaluate textual explanations are highly volatile with respect to each other, even when they measure the same type of similarity (i.e., semantic vs. syntactic), and therefore pointing at the need of more reliable metrics for the empirical evaluation of textual explanations. In general, BERTScore and METEOR metrics exhibit the strongest correlation with human judgements. Lastly, our study provides evidence of the desiderata for LLM-generated explanations, suggesting that explanations should be coherent with human reasoning rather than model reasoning. Participants value the most textual explanations that are relevant and exhaustive to the text they refer to, while being logically and linguistically correct. Justifications for this preference lie on the fact that abusive language detection heavily relies on additional context and knowledge about slang and slurs, for which receiving an explanation is helpful to participants' understanding of the text. Future work should investigate whether this preference holds for other domains as well. In light of our findings, we conclude with

three recommendations to use LLMs responsibly for explainable abusive language detection: **(1)** be aware of the cultural bias these models might exhibit when generating free-text explanations, which can further harm targeted groups; **(2)** if possible, instruction fine-tune LLMs for explanation generation of abusive language detection. This not only could ensure the generation of structured explanations as advised by previous research [1] but it also returns the highest evaluation scores, both empirically and expert-wise, when using knowledge-guided prompts; **(3)** opt for a combination of empirical metrics to evaluate textual explanations when no human evaluation is possible, since no particular empirical metric seems to generalise across different learning techniques, models and datasets, making the ground-truth lie somewhere in between BERTScore (upper bound) and BLEU (lower bound).

Acknowledgments

This work was supported by the UK Research and Innovation [grant number EP/S023356/1] in the UKRI Centre for Doctoral Training in Safe and Trusted Artificial Intelligence (www.safeandtrustedai.org); by the Trustworthy AI Research award by The Alan Turing Institute, supported by the British Embassy Rome and the UK Science & Innovation Network; and by the PNRR project FAIR - Future AI Research (PE00000013), Spoke 6 - Symbiotic AI (CUP H97G22000210007) under the NRRP MUR program funded by the NextGenerationEU.

References

[1] P. Mishra, H. Yannakoudakis, E. Shutova, Tackling online abuse: A survey of automated abuse detection methods, arXiv preprint arXiv:1908.06024 (2019).

[2] P. Barceló, M. Monet, J. Pérez, B. Subercaseaux, Model interpretability through the lens of computational complexity, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 15487–15498. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/b1adda14824f50ef24ff1c05bb66faf3-Paper.pdf.

[3] M. Sap, D. Card, S. Gabriel, Y. Choi, N. A. Smith, The risk of racial bias in hate speech detection, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 1668–1678. URL: https://aclanthology.org/P19-1163. doi:10.18653/v1/P19-1163.

[4] The European Parliament and The Council of the European Union, Eu regulation 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation), Official Journal of the European Union (2016).

[5] O. L. Haimson, D. Delmonaco, P. Nie, A. Wegner, Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas, Proc. ACM Hum.-Comput. Interact. 5 (2021). URL: https://doi.org/10.1145/3479610. doi:10.1145/3479610.

[6] J. Brunk, J. Mattern, D. M. Riehle, Effect of transparency and trust on acceptance of automatic online comment moderation systems, in: 2019 IEEE 21st Conference on Business Informatics (CBI), volume 01, 2019, pp. 429–435. doi:10.1109/CBI.2019.00056.

[7] A. Calabrese, L. Neves, N. Shah, M. W. Bos, B. Ross, M. Lapata, F. Barbieri, Explainability and hate speech: Structured explanations make social media moderators faster, arXiv preprint arXiv:2406.04106 (2024).

[8] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, A. Mukherjee, Hatexplain: A benchmark dataset for explainable hate speech detection, in: Proceedings of the AAAI conference on artificial intelligence, volume 35, 2021, pp. 14867–14875.

[9] M. ElSherief, C. Ziems, D. Muchlinski, V. Anupindi, J. Seybolt, M. De Choudhury, D. Yang, Latent hatred: A benchmark for understanding implicit hate speech, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 345–363.

[10] M. Sap, S. Gabriel, L. Qin, D. Jurafsky, N. A. Smith, Y. Choi, Social bias frames: Reasoning about social and power implications of language, in: ACL, 2020.

[11] D. Nozza, A. T. Cignarella, G. Damo, T. Caselli, V. Patti, Hodi at evalita 2023: Overview of the first shared task on homotransphobia detection in italian, in: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop, EVALITA 2023, CEUR Workshop Proceedings (CEUR-WS. org), 2023.

[12] H. Kirk, W. Yin, B. Vidgen, P. Röttger, Semeval-2023 task 10: Explainable detection of online sexism, in: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), 2023, pp. 2193–2210.

[13] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, et al., Scaling instruction-finetuned language models, arXiv preprint arXiv:2210.11416 (2022).

[14] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[15] T. Zhang, F. Ladhak, E. Durmus, P. Liang, K. McKeown, T. B. Hashimoto, Benchmarking large language models for news summarization, arXiv preprint arXiv:2301.13848 (2023).

[16] C. Ziems, W. Held, O. Shaikh, J. Chen, Z. Zhang, D. Yang, Can large language models transform computational social science?, arXiv preprint arXiv:2305.03514 (2023).

[17] S. Roy, A. Harshvardhan, A. Mukherjee, P. Saha, Probing LLMs for hate speech detection: strengths and vulnerabilities, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 6116–6128. URL: https://aclanthology.org/2023.findings-emnlp.407. doi:10.18653/v1/2023.findings-emnlp.407.

[18] Y. Yang, J. Kim, Y. Kim, N. Ho, J. Thorne, S.-Y. Yun, HARE: Explainable hate speech detection with step-by-step reasoning, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 5490–5505. URL: https://aclanthology.org/2023.findings-emnlp.365. doi:10.18653/v1/2023.findings-emnlp.365.

[19] F. Huang, H. Kwak, J. An, Chain of explanation: New prompting method to generate quality natural language explanation for implicit hate speech, in: Companion Proceedings of the ACM Web Conference 2023, WWW '23 Companion, Association for Computing Machinery, New York, NY, USA, 2023, p. 90–93. URL: https://doi.org/10.1145/3543873.3587320. doi:10.1145/3543873.3587320.

[20] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al., Chain-of-thought prompting elicits reasoning in large language models, Advances in neural information processing systems 35 (2022) 24824–24837.

[21] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[22] H. Wang, M. S. Hee, M. R. Awal, K. T. W. Choo, R. K.-W. Lee, Evaluating gpt-3 generated explanations for hateful content moderation, in: Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, 2023, pp. 6255–6263.

[23] R. Bhardwaj, S. Poria, Red-teaming large language models using chain of utterances for safety-alignment, arXiv preprint arXiv:2308.09662 (2023).

[24] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, T. B. Hashimoto, Stanford alpaca: An instruction-following llama model, https://github.com/tatsu-lab/stanford_alpaca, 2023.

[25] N. Muennighoff, T. Wang, L. Sutawika, A. Roberts, S. Biderman, T. Le Scao, M. S. Bari, S. Shen, Z. X. Yong, H. Schoelkopf, X. Tang, D. Radev, A. F. Aji, K. Almubarak, S. Albanie, Z. Alyafeai, A. Webson, E. Raff, C. Raffel, Crosslingual generalization through multitask finetuning, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 15991–16111. URL: https://aclanthology.org/2023.acl-long.891. doi:10.18653/v1/2023.acl-long.891.

[26] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models, arXiv preprint arXiv:2307.09288 (2023).

[27] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).

[28] D. Vrandečić, M. Krötzsch, Wikidata: a free collaborative knowledgebase, Communications of the ACM 57 (2014) 78–85.

[29] K. Halevy, A group-specific approach to nlp for hate speech detection, arXiv preprint arXiv:2304.11223 (2023).

[30] R. Speer, J. Chin, C. Havasi, Conceptnet 5.5: An open multilingual graph of general knowledge, in: Proceedings of the AAAI conference on artificial intelligence, volume 31, 2017.

[31] A. B. Sai, T. Dixit, D. Y. Sheth, S. Mohan, M. M. Khapra, Perturbation CheckLists for evaluating NLG evaluation metrics, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 7219–7234. URL: https://aclanthology.org/2021.emnlp-main.575. doi:10.18653/v1/2021.emnlp-main.575.

[32] E. Reiter, A structured review of the validity of BLEU, Computational Linguistics 44 (2018) 393–401. URL: https://aclanthology.org/J18-3002. doi:10.1162/coli_a_00322.

[33] J. Novikova, O. Dušek, A. Cercas Curry, V. Rieser, Why we need new evaluation metrics for NLG,

in: M. Palmer, R. Hwa, S. Riedel (Eds.), Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 2241–2252. URL: https://aclanthology.org/D17-1238. doi:10.18653/v1/D17-1238.

[34] A. B. Sai, A. K. Mohankumar, M. M. Khapra, A survey of evaluation metrics used for nlg systems, ACM Computing Surveys (CSUR) 55 (2022) 1–39.

[35] A. Celikyilmaz, E. Clark, J. Gao, Evaluation of text generation: A survey, arXiv preprint arXiv:2006.14799 (2020).

[36] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, in: International Conference on Learning Representations, 2019.

[37] A. Lavie, M. J. Denkowski, The meteor metric for automatic evaluation of machine translation, Machine translation 23 (2009) 105–115.

[38] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, Łukasz Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, J. Dean, Google's neural machine translation system: Bridging the gap between human and machine translation, 2016. arXiv:1609.08144.

[39] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://www.aclweb.org/anthology/W04-1013.

[40] M. Popović, chrF: character n-gram F-score for automatic MT evaluation, in: Proceedings of the Tenth Workshop on Statistical Machine Translation, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 392–395. URL: https://aclanthology.org/W15-3049. doi:10.18653/v1/W15-3049.

[41] M. Popović, chrF++: words helping character n-grams, in: Proceedings of the Second Conference on Machine Translation, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 612–618. URL: https://aclanthology.org/W17-4770. doi:10.18653/v1/W17-4770.

[42] M. Post, A call for clarity in reporting BLEU scores, in: Proceedings of the Third Conference on Machine Translation: Research Papers, Association for Computational Linguistics, Belgium, Brussels, 2018, pp. 186–191. URL: https://www.aclweb.org/anthology/W18-6319.

[43] K. Krippendorff, Computing krippendorff's alpha-reliability, 2011.

A. Prompt Details

Table 5 shows the two types of prompts we used in our experiments, following the template of the Stanford Alpaca project. The two categories differ for the 'context' that is passed in the knowledge-guided version, which contains the information extracted from the knowledge sources linked to the text. As described in the Section 2.2 of the paper, we used the vanilla prompts for zero-shot learning, few-shot learning, and instruction fine-tuning whereas we used the knowledge-guided prompts for knowledge-guided zero-shot learning and knowledge-guided instruction fine-tuning.

B. Survey Questions

Participants were presented with the questions shown in Table 6.

Category	Prompt Template
Vanilla	Below is an instruction that describes a task, paired with input text. Write a response that appropriately completes the instruction. Instruction: Classify the input text as `list_of_labels`, and provide an explanation. Input text: `text_to_classify`. Response:
Knowledge-guided	Below is an instruction that describes a task, paired with context and input text. Write a response that appropriately completes the instruction based on the context. Instruction: Classify the input text as `list_of_labels`, and provide an explanation. Context: `knowledge_source_linked`. Input text: `text_to_classify`. Response:

Table 5
Details of vanilla prompts and knowledge-guided prompts passed to the LLMs in our experiments.

Part	Questions
Before Treatment	"Which gender do you identify as?" "Are you an English native-speaker?" "What is your country of origin?" "What is your level of expertise on language models or abusive language?" "How useful would you rate a system that provides you a textual explanation for its classification with respect to receiving only its classification?" "How trustworthy would you rate a system that provides you a textual explanation for its classification with respect to receiving only its classification?"
Treatment	"Do you think explanation 1 provides a good explanation given the text?" "If your answer was yes, does explanation 2 mean the same thing as explanation 1?" "If your answer was yes, does explanation 3 mean the same thing as explanation 1?" "If your answer was yes, does explanation 4 mean the same thing as explanation 1?"
After Treatment	"Having seen these explanations, how useful would you rate a system that provides you a textual explanation for its classification?" "Having seen these explanations, how trustworthy would you rate a system that provides you a textual explanation for its classification?" "What was the main error you noticed in these explanations?" "What do you think makes a textual explanation good?" "Do you have any comment you would like to share?"

Table 6
List of questions asked to participants in our expert survey.