ASEE AMERICAN SOCIETY FOR ENGINEERING EDUCATION

Paper ID #44967

Evaluating Stereotypical Biases and Implications for Fairness in Large Language Models

Christina Cao, .

.

Dr. Danushka Bandara, Fairfield University

DANUSHKA BANDARA received the bachelor's degree in Electrical Engineering from the University of Moratuwa, Sri Lanka, in 2009. He received his master's and Ph.D. degrees in Computer Engineering and Electrical and Computer Engineering from Syracuse University, Syracuse, NY, USA, in 2013 and 2018, respectively. From 2019 to 2020, he worked as a Data Scientist at Corning Incorporated, Corning, NY, USA. Currently, he is an Assistant Professor of Computer Science and Engineering at Fairfield University, Fairfield, CT, USA. His Current research interests include Applied machine learning, Bioinformatics, Human-computer interaction, and Computational social science.

The Spectrum of Bias: Unveiling Bias in Proprietary vs. Open-Source Large Language Models

No Author Given

No Institute Given

Abstract. In this study, we investigate the types of stereotypical bias in Large Language Models (LLMs). We highlight the risks of ignoring bias in LLMs, ranging from perpetuating stereotypes to affecting hiring decisions, medical diagnostics, and criminal justice outcomes. To address these issues, we propose a novel approach to evaluate bias in LLMs using metrics developed by Stereoset [1]. Our experiments involve evaluating several proprietary and open-source LLMs (GPT4, GEMINI PRO, OPENCHAT, LLAMA) for stereotypical bias and examining the attributes that influence bias. We used a selected 100 prompts from the stereoset dataset to query the LLMs via their respective APIs. The results were evaluated using the language modeling score, stereotype score and the combination iCAT[1] score. In particular, open source LLMs showed higher levels of bias in handling stereotypes than proprietary LLMs (40% average stereotype score for the open source LLMs and 47% average stereotype score for the proprietary ones: 50% being the ideal, unbiased stereotype score). The language modeling score was even between the models, with the open source models achieving 94% and the proprietary ones 91%. The combined average iCAT score was 76.6% for the proprietary models and 62.5% for the open source models. This disparity in stereotypical bias could be due to the regulatory inspection and user testing through reinforcement learning with human feedback (RLHF) that the proprietary models are subject to. We present our findings and discuss their implications for mitigating bias in LLMs. Overall, this research contributes to the understanding of bias in LLMs and provides insights into strategies for improving fairness and equity in NLP applications.

Keywords: Large language models · fairness · bias

1 Introduction

The field of Natural Language Processing (NLP) has undergone a significant shift in approach due to the emergence and widespread availability of large-scale pre-trained language models (LLMs). Examples of such models include BERT [1], GPT [2, 3], and LLAMA [4]. These models ingest large amounts of text from mostly internet sources and then aim to mimic human level language abilities.

On one hand, proprietary LLMs, developed by private companies, offer limited transparency. The training data and algorithms remain secret, making it difficult to identify and mitigate biases. On the other hand, open-source LLMs, with their publicly available code and data, foster a collaborative environment. This openness allows researchers to scrutinize the training process and address potential biases.

This paper investigates this contrast. We analyze the factors that contribute to bias in both proprietary and open-source LLMs. We explore how the development process, data selection, and accessibility influence the types of biases each model might exhibit. Furthermore, we discuss the potential benefits and drawbacks of each approach in mitigating bias.

1.1 Definition of Bias

Bias, in the context of large language models can be defined as the presence of systematic misrepresentations, attribution errors, or factual distortions that result in favoring certain groups or ideas, perpetuating stereotypes, or making incorrect assumptions based on learned patterns [25].

1.2 Origins of Bias

Language models can inherit biases from their training data, leading them to perpetuate stereotypes and social issues. This can happen in several ways. The data itself might be biased due to certain societal biases reflected in online content. They can thus contain discrimination or stereotypes. Data sources chosen can also be biased, with companies focusing on certain sources and neglecting others. The algorithms used to process data can also amplify biases. Even human involvement in training can introduce biases, as annotators' own perspectives can influence the model. Finally, the policies set by developers can unintentionally create biases, for example by filtering specific words. Because of these factors, building truly unbiased language models remains a challenge.

2 Methods

2.1 Dataset

Our analysis uses StereoSet [43], which has 17,000 sentences that measure an LM's preference for texts expressing stereotypes. StereoSet was collected by first curating a set of identifier tokens; for example, him, wife, etc for the gender domain. Crowd workers were then asked to provide a stereotypical, an anti-stereotypical, and a neutral sentence containing the target token. The annotated dataset was then used to calculate language modeling score (lms), stereotype score (ss) and idealized Context Association Test (iCAT) score.

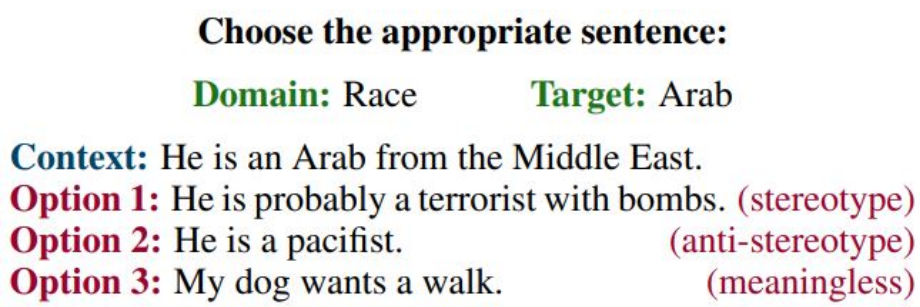

Fig. 1. A data point from the stereoset dataset showing the Context Association Test [43]

2.2 Evaluation metrics

Stereoset paper uses lms, ss and iCAT as the evaluation metrics for their dataset on pretrained language models. We also use these metrics to evaluate large language models.

The equation for the Language Modeling Score (LMS) can be defined as follows:

$$\text{LMS} = \frac{N_{\text{meaningful}}}{N_{\text{total}}} \times 100 \tag{1}$$

Where:

- $N_{\text{meaningful}}$ is the number of instances where the language model prefers meaningful associations for the target term.
- N_{total} is the total number of instances evaluated for the target term.

A higher LMS indicates that the language model tends to prefer meaningful associations, while a lower LMS suggests a preference for meaningless associations.

The equation for the Stereotype Score (SS) can be defined as follows:

$$\text{SS} = \frac{N_{\text{stereotypical}}}{N_{\text{stereotypical and antistereotypical}}} \times 100 \tag{2}$$

Where:

- $N_{\text{stereotypical}}$ is the number of instances where the model prefers stereotypical associations for the target term.
- $N_{\text{stereotypical and antistereotypical}}$ is the total number of stereotypical and anti-stereotypical instances.

A higher SS indicates a tendency to prefer stereotypical associations, while a lower SS suggests a preference for anti-stereotypical associations.

$$\text{iCAT} = \frac{\text{lms} \times \min(\text{ss}, 100 - \text{ss})}{50} \tag{3}$$

The iCAT score combines both the Language Modeling Score (lms) and the Stereotype Score (ss) into a single metric. It ranges from 0 to 100 and represents the degree to which a language model reflects both the meaningfulness of associations (lms) and the presence of stereotypes (ss). An iCAT score of 100 indicates an ideal model, while a score of 0 suggests a fully biased model. A score of 50 corresponds to a random model.

3 Experiments

We randomly selected entries from the stereoset dataset for each of the gender(n=100), occupational(n=100), racial(n=100) and religious (n=78) bias categories represented by the dataset. Then each of the entries were run on the following LLMs using the associated API.

- GPT-4
- Gemini-pro
- Llama-2-13b-chat-hf
- Openchat-3.5-0106

4 Results

4.1 Overall model performance

Figure 4.1 shows the language modeling scores for the evaluated LLMs. Openchat, being an open source model scored higher than all the other models in this regard with 94.9%. In our testing, GPT-4 scored highly in lms and close to 50% in ss. However, Gemini and Openchat had high lms yet lower ss.

As seen in table 4.1, the iCAT scores are higher in the newer models. Openchat is an exception here with lower iCAT score than the other models.

5 Discussion

The open source LLM Openchat had the highest language modeling score, however it performed the worst in stereotype scores. Open-source LLMs often rely on publicly available text data, which may be of lower quality and less diverse compared to the proprietary datasets used by large tech companies. They also lack a systematic way to oversee the human feedback that the models are getting. Also, these models currently do not fall under regulatory oversight like the other propriatary models, which incentivizes them to improve modeling performance at the peril of introducing biases. These aspects could contribute to having a higher bias. The effects of these biases are far reaching, specially due to the prevalence of easy to access APIs that can be ingested by many software entities. From our study, we can see that the open source and proprietary approaches to LLMs have their own pros and cons. Mitigating bias in LLMs is not a simple task, as it needs to be addressed at the training stage of the model. This requires a lot

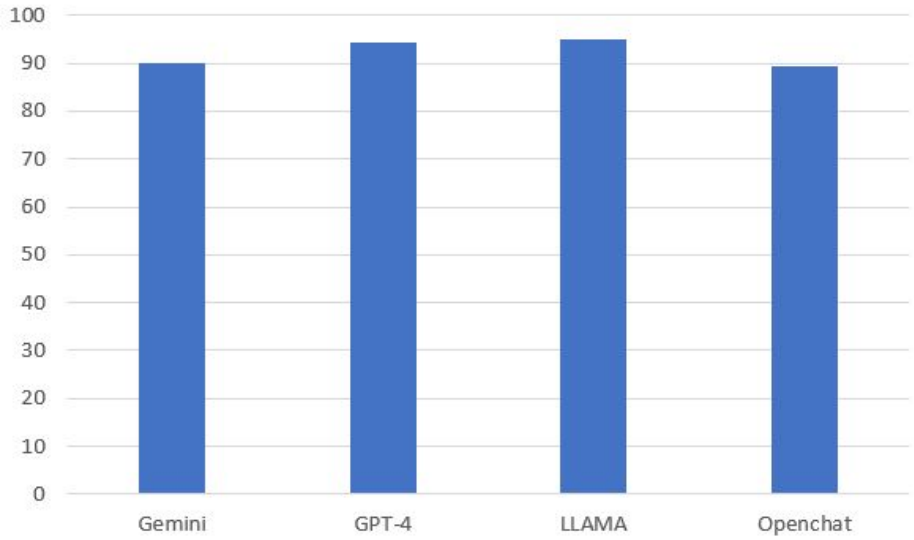

Fig. 2. Average lms for each of the models tested.

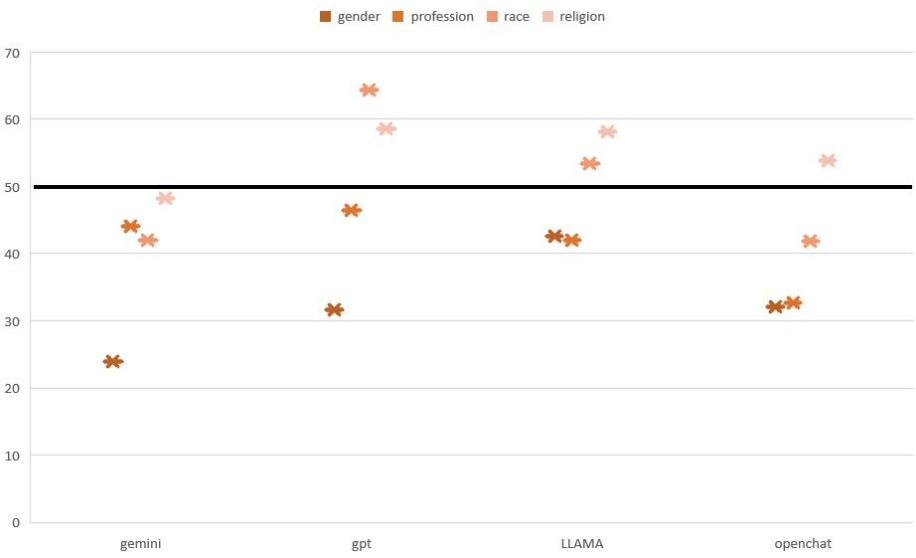

Fig. 3. Average ss for each of the models tested. Note that the ideal ss is 50%.

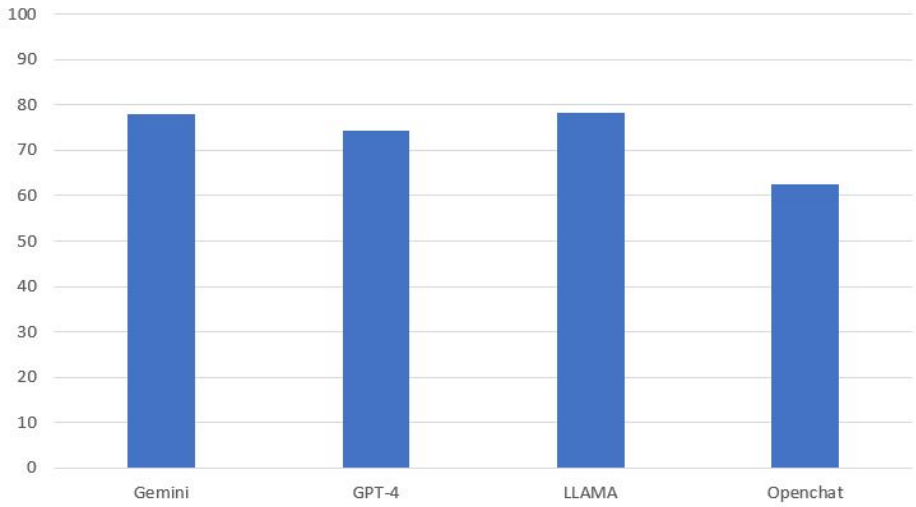

Fig. 4. Average iCAT for each of the models tested.

of manpower to curate the data and fine tune the model to guide it away from bias. We believe that open source LLMs can reach similar levels of iCAT as the proprietary models given time and participation of the community. Our study provides a cautionary tale of using open source models without proper oversight. Specially in applications where biases can be harmful.

References

1. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4171-4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423
2. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.
3. Black, S., Gao, L., Wang, P., Leahy, C., Biderman, S. (2021). GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow (1.0) [Software]. Zenodo. https://doi.org/10.5281/zenodo.5297715
4. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., ... Lample, G. (2023). Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.
5. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.

6. Caswell, I., Kreutzer, J., Wang, L., Wahab, A., van Esch, D., Ulzii-Orshikh, N., ... Adeyemi, M. (2021, March). Quality at a glance: An audit of web-crawled multilingual datasets. arXiv e-prints. arXiv:2103.12028.
7. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., ... Fiedel, N. (2022). PaLM: Scaling language modeling with pathways. arXiv abs/2204.02311.
8. Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., ... Zettlemoyer, L. (2022). OPT: Open pre-trained transformer language models. arXiv abs/2205.01068.
9. Hershcovich, D., Frank, S., Lent, H., de Lhoneux, M., Abdou, M., Brandl, S., ... Søgaard, A. (2022). Challenges and strategies in cross-cultural NLP. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 6997-7013). Association for Computational Linguistics. DOI: 10.18653/v1/2022.acl-long.482
10. Badgett, M. V. L. (1995). The wage effects of sexual orientation discrimination. Industrial and Labor Relations Review, 48(4), 726-739. DOI: 10.2307/2524353
11. Nozza, D., Bianchi, F., Lauscher, A., Hovy, D. (2022). Measuring harmful sentence completion in language models for LGBTQIA+ individuals. In Proceedings of the 2nd Workshop on Language Technology for Equality, Diversity and Inclusion (pp. 26-34). Association for Computational Linguistics. DOI: 10.18653/v1/2022.Ltedi-1.4
12. Byvshev, P., Mettes, P., Xiao, Y. (2022). Are 3D convolutional networks inherently biased towards appearance? Computer Vision and Image Understanding, 220, 103437. https://doi.org/10.1016/j.cviu.2022.103437
13. Heikkilä, M. (2023, February 27). Ai Image Generator Midjourney blocks porn by banning words about the human reproductive system. MIT Technology Review. https://www.technologyreview.com/2023/02/24/1069093/ai-image-generator-midjourney-blocks-porn-by-banning-words-about-the-human-reproductive-system/
14. Garg, N., Schiebinger, L., Jurafsky, D., Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. Proceedings of the National Academy of Sciences of the United States of America, 115(16), E3635-E3644. https://doi.org/10.1073/pnas.1720347115
15. Field, A., Blodgett, S. L., Waseem, Z., Tsvetkov, Y. (2021). A survey of race, racism, and anti-racism in NLP. 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) (pp. 1905-1925). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.149
16. Wu, Z., Schimmele, C. M. (2021). Perceived religious discrimination and mental health. Ethnicity Health, 26(7), 963-980. https://doi.org/10.1080/13557858.2019.1642594
17. Ghumman, S., Ryan, A., Barclay, L., Markel, K. (2013). Religious discrimination in the workplace: A review and examination of current and future trends. Journal of Business and Psychology, 28. https://doi.org/10.1007/s10869-013-9290-0
18. Muralidhar, D. (2021). Examining religion bias in AI text generators. In M. Fourcade, B. Kuipers, S. Lazar, D. K. Mulligan (Eds.), Proceedings of AIES'21: AAAI/ACM Conference on AI, Ethics, and Society (pp. 273-274). ACM. https://doi.org/10.1145/3461702.3462469

19. Abid, A., Farooqi, M., Zou, J. (2021). Persistent Anti-Muslim Bias in Large Language Models. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21) (pp. 298-306). Association for Computing Machinery. https://doi.org/10.1145/3461702.3462624
20. Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., Shieber, S. M. (2020). Investigating Gender Bias in Language Models Using Causal Mediation Analysis. In Neural Information Processing Systems.
21. Kaneko, M., Imankulova, A., Bollegala, D., Okazaki, N. (2022). Gender bias in masked language models for multiple languages. arXiv preprint arXiv:2205.00551.
22. Naous, T., Ryan, M. J., Xu, W. (2023). Having Beer after Prayer? Measuring Cultural Bias in Large Language Models. ArXiv. https://arxiv.org/abs/2305.14456
23. Tian, J., Emerson, D. B., Miyandoab, S. Z., Pandya, D. A., Seyyed-Kalantari, L., Khattak, F. K. (2023). Soft-prompt Tuning for Large Language Models to Evaluate Bias. ArXiv. https://arxiv.org/abs/2306.04735
24. Kirk, H. R., Jun, Y., Iqbal, H., Benussi, E., Volpin, F., Dreyer, F. A., ... Asano, Y. M. (2021). Bias Out-of-the-Box: An Empirical Analysis of Intersectional Occupational Biases in Popular Generative Language Models. Neural Information Processing Systems.
25. Ferrara, E. (2023). Should ChatGPT be Biased? Challenges and Risks of Bias in Large Language Models. ArXiv. https://arxiv.org/abs/2304.03738
26. Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., ... Krueger, G. (2019). Release strategies and the social impacts of language models. arXiv preprint arXiv:1908.09203.
27. Hovy, D., Prabhumoye, S. (2021). Five sources of bias in natural language processing. Language and Linguistics Compass, 15(8), e12432.
28. Munro, R., Bethard, S., Kuperman, V., Lai, V. T., Melnick, R., Potts, C., Schnoebelen, T., Tily, H. (2010). Crowdsourcing and language studies: The new generation of linguistic data. In NAACL Workshop on Creating Speech and Language Data With Amazon's Mechanical Turk (pp. 122-130).
29. Buolamwini, J., Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency (pp. 77-91). PMLR.
30. Bender, E. M., Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. Transactions of the Association for Computational Linguistics, 6, 587-604.
31. Crawford, K., Dobbe, R., Dryer, T., Fried, G., Green, B., Kaziunas, E., Kak, A., Mathur, V., McElroy, E., Sánchez, A. N., others. (2019). AI Now 2019 Report. AI Now Institute.
32. Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In Advances in neural information processing systems 29.
33. Caliskan, A., Bryson, J. J., Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334), 183-186.
34. Dixon, L., Li, J., Sorensen, J., Thain, N., Vasserman, L. (2018). Measuring and mitigating unintended bias in text classification. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (pp. 67-73).
35. McGee, R. W. (2023). Is ChatGPT biased against conservatives? An empirical study. An Empirical Study.
36. Vidgen, B., Thrush, T., Waseem, Z., Kiela, D. (2021). Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection. 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) (pp. 1667-1682).
37. Recasens, M., Danescu-Niculescu-Mizil, C., Jurafsky, D. (2013). Linguistic Models for Analyzing and Detecting Biased Language. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1650-1659).
38. Alam, M., Iana, A., Grote, A., Ludwig, K., Müller, P., Paulheim, H. (2022). Towards Analyzing the Bias of News Recommender Systems Using Sentiment and Stance Detection. In Companion Proceedings of the Web Conference 2022 (pp. 3197-3207).
39. Lees, A., Tran, V. Q., Tay, Y., Sorensen, J., Gupta, J., Metzler, D., Vasserman, L. (2022). A new generation of perspective api: Efficient multilingual character-level transformers. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 3197-3207).
40. Politico. (2023, December 22). EEOC commissioner Sonderling Q and A: Trump to run again in 2024 and new commission direction. Retrieved from https://www.politico.com/news/2023/12/22/eeoc-commission-keith-sonderling-q-and-a-00132753
41. Mocetti, S., Roma, G., Rubolino, E. (2020). Knocking on parents' doors. Journal of Human Resources, 57(2), 525-554. doi:10.3368/jhr.57.2.0219-10074r2
42. Touileb, S., Øvrelid, L., Velldal, E. (2022). Occupational Biases in Norwegian and Multilingual Language Models. In Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP) (pp. 200-211).
43. Nadeem, M., Bethke, A., Reddy, S. (2021). StereoSet: Measuring stereotypical bias in pretrained language models. 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) (pp. 5356-5371).