Safety and Ethical Concerns of Large Language Models

Zhiheng Xi, Rui Zheng, Tao Gui
School of Computer Science, Fudan University, Shanghai, China
Institute of Modern Languages and Linguistics, Fudan University, Shanghai, China
Shanghai Key Laboratory of Intelligent Information Processing, Shanghai, China
zhxi22@m.fudan.edu.cn, {rzheng20,tgui}@fudan.edu.cn

Abstract

Recent months have witnessed significant progress in the field of large language models (LLMs). Represented by ChatGPT and GPT-4, LLMs perform well in various natural language processing tasks and have been applied to many downstream applications to facilitate people's lives. However, there still exist safety and ethical concerns. Specifically, LLMs suffer from social bias, robustness problems, and poisoning issues, all of which may induce LLMs to spew harmful contents. We propose this tutorial as a gentle introduction to the safety and ethical issues of LLMs.

1 Introduction

As the model size and dataset size scale up in recent natural language processing field, large language models like ChatGPT and GPT-4 have exhibited exceptional performance in a variety of NLP tasks and can even perform complex reasoning or in-context learning (i.e., generalizing to a new task from a few examples) (Brown et al., 2020; Ouyang et al., 2022; OpenAI, 2023; Wei et al., 2022). Moreover, many downstream applications have been developed based on LLMs, which brings significant benefits and convenience to people (Schick et al., 2023; Driess et al., 2023). Despite their fantastic capabilities and potentials, LLMs have raised valid concerns regarding their safety and ethical implications (Bommasani et al., 2021). To be specific, LLMs suffer from social bias (Ferrara, 2023), robustness problems (Zhuo et al., 2023; Wang et al., 2023; Chen et al., 2023), and poisoning issues(Chen et al., 2021), all of which may lead LLMs to generate harmful and rude contents. In this tutorial, we introduce the aforementioned problems, discuss the potential causes, and list some approaches to alleviate these problems.

2 Bias

Language models pre-trained on large-scale corpus usually demonstrate various types of biases like racial discrimination and gender discrimination (Basta et al., 2019; Beltagy et al., 2019; Kurita et al., 2019; Zhang et al., 2020). We follow Bender et al. (2021) and define bias by stereotypical associations and negative sentiment towards specific groups. With the scaling up of LLMs in model size and data size, such biases are not eliminated (Ferrara, 2023). Therefore, when they are deployed in downstream applications, such biases can make users disappointed.

The question of why (large) language models are prone to bias has been well explored, and most of the works suggest that the biases are a reflection of training data patterns (Henderson et al., 2018; Hutchinson et al., 2020; Tan and Celis, 2019; Guo and Caliskan, 2021). LLMs are typically trained with unsupervised learning techniques on large-scale data, including websites, articles, and books. The data may contain unfair or biased characteristics. For example, Hutchinson et al. (2020) demonstrate a bias towards associating phrases that reference individuals with disabilities with a greater frequency of negative sentiment words; furthermore, it has been observed that the topics of gun violence, homelessness, and drug addiction are disproportionately prevalent in texts pertaining to mental illness.

To alleviate bias issues of LLMs, researchers have proposed various approaches. A line of work tries to identify the sources that are most responsible for biases and take actions to make models obviate

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 9-16, Harbin, China, August 3 - 5, 2023.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

reflecting the inequities or biases (Bommasani et al., 2021; Lu et al., 2020; Zhao et al., 2018). Some other work develops calibrating techniques to address bias problems of LLMs (Zhao et al., 2021; Holtzman et al., 2021). Another potential direction is to leverage alignment techniques like Reinforcement Learning from Human Feedback (RLHF) (Bai et al., 2022; Ouyang et al., 2022; Ferrara, 2023; Zheng et al., 2023), where LLMs are trained to align with human values and thus some biases can be mitigated.

Mitigating biases of LLMs remains an important problem and we hope that more research efforts will be made to construct fair AI systems.

3 Robustness

Pretrained language models are known to be vulnerable to adversarial instances crafted by performing subtle perturbations on normal ones (Ren et al., 2019; Garg and Ramakrishnan, 2020; Wang et al., 2021b). With increasing scales, LLMs still face such challenges and their performance suffers significant drops under adversarial attacks (Zhuo et al., 2023; Wang et al., 2023; Chen et al., 2023). For example, when conducting in-context learning, models' performance can be unstable when changing the choice of prompt format, training examples and the order of examples (Chen et al., 2022; Zhao et al., 2021).

In order to improve the robustness of language models against adversarial attackers, many defense strategies have been proposed. A line of work focuses on designing adversarial training algorithms to enhance model robustness, e.g., FreeLB (Zhu et al., 2020) and InfoBERT (Wang et al., 2021a). However, these approaches consume too many training resources as they require multi-step gradient descents to generate adversarial examples, and this problem of inefficiency will be amplified with larger models. Another line of work searches for a robust model architecture with sparse optimization techniques (Xi et al., 2022; Zheng et al., 2022). However, such techniques may induce a trade-off between robustness and accuracy (Zhang et al., 2019; Tsipras et al., 2019). Some other work tries to design prompts to elicit reliable and robust responses from LLMs (Si et al., 2022), which is a potential direction as prompt engineering does not require training models or changing their architectures.

The robustness of LLMs is still a problem that has not been fully explored, and we call for more attention from the community to build robust language models.

4 Poisoning

In an ICML 2017 outstanding paper (Koh and Liang, 2017), the authors employ the novel Influence Function to gauge alterations in model parameters, could provide a quantitative evaluation of the impact individual training samples on the model. This assessment reveals whether a sample affects the model's training, and to what extent. Experimental findings demonstrate that, with modifications to a mere two training samples, the model incorrectly predicts over 77% of the test data for specific test instances. Altering ten training samples results in nearly 100% erroneous predictions on test data. Gu et al. (2017) cleverly introduce poisoned data into the training set, ensuring that the model's accuracy on pristine data remains constant or marginally declines, while simultaneously triggering specific outputs when presented with data containing particular trigger words. Such poisoned models may be elicited to generate toxic contents like abusive language, hate speech, violent speech (Liang et al., 2022; Gururangan et al., 2022).

Dai et al. (2019) select brief sentences as backdoor triggers, such as "I watched this 3D movie," and randomly incorporate them into movie reviews to generate tainted samples for backdoor training. Kurita et al. (2020) employ rare and nonsensical words like "cf" as triggers. Similarly, Chen et al. (2021) utilize words as triggers, experimenting with words of varying frequencies. Chen and Dai (2021) postulate that triggers associate with specific neurons, influencing only certain hidden states. Qi et al. (2021) suggest a defense premised on the observation that perplexity undergoes significant alterations when trigger words are excised from samples. Li et al. (2021) conduct a thorough analysis of backdoor attacks in text classification, ultimately developing a backdoor-free text classifier training framework, dubbed BFClass.

As the extensive utilization of open-source datasets and models persists, poisoning remains a subject warranting scrupulous attention.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 9-16, Harbin, China, August 3 - 5, 2023.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

5 Tutorial Outline

Part I: Introduction (20 min)

- The development of large language models
- The importance of safety and ethical concerns
- Safety and ethical concerns LLMs suffer
  - Social bias
  - Robustness problems
  - Poisoning issues

Part II: Bias (20 min)

- Definition, types and sources of Bias
- Bias of large language models
- Methods to alleviate bias issues
  - Identify the causes of bias and addressing them
  - Calibrating methods
  - Reinforcement Learning from Human Feedback

Part III: Robustness (20 min)

- Textual adversarial robustness
- Robustness of large language models
- Defense strategies to improve robustness
  - Adversarial training
  - Finding robust structures of neural networks
  - Prompting methods

Part IV: Poisoning (20 min)

- Definition of poisoning issues
- Poisoning methods
  - Dataset attacks
  - Backdoors and triggers

Part V: Conclusion (10 min)

6 Reading List

1. On the Opportunities and Risks of Foundation Models (Bommasani et al., 2021);
2. Ethical challenges in data-driven dialogue systems (Henderson et al., 2018);
3. Training a helpful and harmless assistant with reinforcement learning from human feedback (Bai et al., 2022);
4. Training language models to follow instructions with human feedback (Ouyang et al., 2022);
5. On the dangers of stochastic parrots: Can language models be too big?(Bender et al., 2021)
6. Should chatgpt be biased? challenges and risks of bias in large language models (Ferrara, 2023);

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 9-16, Harbin, China, August 3 - 5, 2023.

7. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases (Guo and Caliskan, 2021);

8. How robust is GPT-3.5 to predecessors? A comprehensive study on language understanding tasks (Chen et al., 2023);

9. Badnets: Identifying vulnerabilities in the machine learning model supply chain (Gu et al., 2017);

10. Secrets of RLHF in Large Language Models Part I: PPO (Zheng et al., 2023);

7 Instructors

Tao Gui is an associate professor at the Institute of Modern Languages and Linguistics of Fudan University. He is the key member of the FudanNLP group[0]. He is a member of ACL, a member of the Youth Working Committee of the Chinese Information Processing Society of China, and the member of the Language and Knowledge Computing Professional Committee of the Chinese Information Processing Society of China. He has published more than 40 papers in top international academic conferences and journals such as ACL, ENNLP, AAAI, IJCAI, SIGIR, and so on. He has served as area chair or PCs for SIGIR, AAAI, IJCAI, TPAMI, and ARR. He has received the Outstanding Doctoral Dissertation Award of the Chinese Information Processing Society of China, the area chair favorite Award of COLING 2018, the outstanding Paper Award of NLPCC 2019, a scholar of young talent promoting projects of CAST, and the Shanghai Rising-Star Program.

Homepage: https://guitaowufeng.github.io

Rui Zheng is a Ph.D. student in the class of 2020 at the School of Computer Science, Fudan University, is supervised by Professor Zhang Qi. His research interests include robust models, dataset debiasing, and large model alignment. He has participated in the development of the large-scale robustness detection tool TextFlint and has published multiple first-author/co-first-author papers at conferences such as ACL, EMNLP, and COLING.

Zhiheng Xi is a first-year master student the School of Computer Science, Fudan University. Prior to that, he received his bachelor's degree from Nanjing University. His research interests lie in robust machine learning, sparse neural networks, prompting techniques, and complex reasoning ability of LLMs. He has published multiple first-author/co-first-author papers at EMNLP and ACL.

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, Benjamin Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *CoRR*, abs/2204.05862.

Christine Basta, Marta Ruiz Costa-jussà, and Noe Casas. 2019. Evaluating the underlying gender bias in contextualized word embeddings. *CoRR*, abs/1904.08783.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3613–3618. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In Madeleine Clare Elish, William Isaac, and Richard S. Zemel, editors, *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, pages 610–623. ACM.

[0] https://nlp.fudan.edu.cn

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 9-16, Harbin, China, August 3 - 5, 2023.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ B. Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. 2021. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Chuanshuai Chen and Jiazhu Dai. 2021. Mitigating backdoor attacks in lstm-based text classification systems by backdoor keyword identification. *Neurocomputing*, 452:253–262.

Xiaoyi Chen, Ahmed Salem, Dingfan Chen, Michael Backes, Shiqing Ma, Qingni Shen, Zhonghai Wu, and Yang Zhang. 2021. Badnl: Backdoor attacks against NLP models with semantic-preserving improvements. In *ACSAC '21: Annual Computer Security Applications Conference, Virtual Event, USA, December 6 - 10, 2021*, pages 554–569. ACM.

Yanda Chen, Chen Zhao, Zhou Yu, Kathleen R. McKeown, and He He. 2022. On the relation between sensitivity and accuracy in in-context learning. *CoRR*, abs/2209.07661.

Xuanting Chen, Junjie Ye, Can Zu, Nuo Xu, Rui Zheng, Minlong Peng, Jie Zhou, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. How robust is GPT-3.5 to predecessors? A comprehensive study on language understanding tasks. *CoRR*, abs/2303.00293.

Jiazhu Dai, Chuanshuai Chen, and Yufeng Li. 2019. A backdoor attack against lstm-based text classification systems. *IEEE Access*, 7:138872–138878.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. 2023. Palm-e: An embodied multimodal language model. *CoRR*, abs/2303.03378.

Emilio Ferrara. 2023. Should chatgpt be biased? challenges and risks of bias in large language models. *CoRR*, abs/2304.03738.

Siddhant Garg and Goutham Ramakrishnan. 2020. BAE: bert-based adversarial examples for text classification. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6174–6181. Association for Computational Linguistics.

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. 2017. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *CoRR*, abs/1708.06733.

Wei Guo and Aylin Caliskan. 2021. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In Marion Fourcade, Benjamin Kuipers, Seth Lazar, and Deirdre K. Mulligan, editors, *AIES '21: AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event, USA, May 19-21, 2021*, pages 122–133. ACM.

Suchin Gururangan, Dallas Card, Sarah K. Dreier, Emily K. Gade, Leroy Z. Wang, Zeyu Wang, Luke Zettlemoyer, and Noah A. Smith. 2022. Whose language counts as high quality? measuring language ideologies in text data selection. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 2562–2580. Association for Computational Linguistics.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 9-16, Harbin, China, August 3 - 5, 2023.

Peter Henderson, Koustuv Sinha, Nicolas Angelard-Gontier, Nan Rosemary Ke, Genevieve Fried, Ryan Lowe, and Joelle Pineau. 2018. Ethical challenges in data-driven dialogue systems. In Jason Furman, Gary E. Marchant, Huw Price, and Francesca Rossi, editors, *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2018, New Orleans, LA, USA, February 02-03, 2018*, pages 123–129. ACM.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn't always right. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 7038–7051. Association for Computational Linguistics.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in NLP models as barriers for persons with disabilities. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5491–5501. Association for Computational Linguistics.

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1885–1894. PMLR.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W. Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. *CoRR*, abs/1906.07337.

Keita Kurita, Paul Michel, and Graham Neubig. 2020. Weight poisoning attacks on pre-trained models. *CoRR*, abs/2004.06660.

Zichao Li, Dheeraj Mekala, Chengyu Dong, and Jingbo Shang. 2021. Bfclass: A backdoor-free text classification framework. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 444–453. Association for Computational Linguistics.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel J. Orr, Lucia Zheng, Mert Yüksekgönül, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2022. Holistic evaluation of language models. *CoRR*, abs/2211.09110.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender bias in neural natural language processing. In Vivek Nigam, Tajana Ban Kirigin, Carolyn L. Talcott, Joshua D. Guttman, Stepan L. Kuznetsov, Boon Thau Loo, and Mitsuhiro Okada, editors, *Logic, Language, and Security - Essays Dedicated to Andre Scedrov on the Occasion of His 65th Birthday*, volume 12300 of *Lecture Notes in Computer Science*, pages 189–202. Springer.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

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 F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *CoRR*, abs/2203.02155.

Fanchao Qi, Yangyi Chen, Mukai Li, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2021. ONION: A simple and effective defense against textual backdoor attacks. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 9558–9566. Association for Computational Linguistics.

Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. 2019. Generating natural language adversarial examples through probability weighted word saliency. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1085–1097. Association for Computational Linguistics.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 9-16, Harbin, China, August 3 - 5, 2023.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *CoRR*, abs/2302.04761.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan L. Boyd-Graber, and Lijuan Wang. 2022. Prompting GPT-3 to be reliable. *CoRR*, abs/2210.09150.

Yi Chern Tan and L. Elisa Celis. 2019. Assessing social and intersectional biases in contextualized word representations. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13209–13220.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. 2019. Robustness may be at odds with accuracy. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Boxin Wang, Shuohang Wang, Yu Cheng, Zhe Gan, Ruoxi Jia, Bo Li, and Jingjing Liu. 2021a. Infobert: Improving robustness of language models from an information theoretic perspective. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Xiao Wang, Qin Liu, Tao Gui, Qi Zhang, Yicheng Zou, Xin Zhou, Jiacheng Ye, Yongxin Zhang, Rui Zheng, Zexiong Pang, Qinzhuo Wu, Zhengyan Li, Chong Zhang, Ruotian Ma, Zichu Fei, Ruijian Cai, Jun Zhao, Xingwu Hu, Zhiheng Yan, Yiding Tan, Yuan Hu, Qiyuan Bian, Zhihua Liu, Shan Qin, Bolin Zhu, Xiaoyu Xing, Jinlan Fu, Yue Zhang, Minlong Peng, Xiaoqing Zheng, Yaqian Zhou, Zhongyu Wei, Xipeng Qiu, and Xuanjing Huang. 2021b. Textflint: Unified multilingual robustness evaluation toolkit for natural language processing. In Heng Ji, Jong C. Park, and Rui Xia, editors, *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL 2021 - System Demonstrations, Online, August 1-6, 2021*, pages 347–355. Association for Computational Linguistics.

Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Haojun Huang, Wei Ye, Xiubo Geng, Binxing Jiao, Yue Zhang, and Xing Xie. 2023. On the robustness of chatgpt: An adversarial and out-of-distribution perspective. *CoRR*, abs/2302.12095.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903.

Zhiheng Xi, Rui Zheng, Tao Gui, Qi Zhang, and Xuanjing Huang. 2022. Efficient adversarial training with robust early-bird tickets. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 8318–8331. Association for Computational Linguistics.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. 2019. Theoretically principled trade-off between robustness and accuracy. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 7472–7482. PMLR.

Haoran Zhang, Amy X. Lu, Mohamed Abdalla, Matthew B. A. McDermott, and Marzyeh Ghassemi. 2020. Hurtful words: quantifying biases in clinical contextual word embeddings. In Marzyeh Ghassemi, editor, *ACM CHIL '20: ACM Conference on Health, Inference, and Learning, Toronto, Ontario, Canada, April 2-4, 2020 [delayed]*, pages 110–120. ACM.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4847–4853. Association for Computational Linguistics.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 9-16, Harbin, China, August 3 - 5, 2023.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Rui Zheng, Bao Rong, Yuhao Zhou, Di Liang, Sirui Wang, Wei Wu, Tao Gui, Qi Zhang, and Xuanjing Huang. 2022. Robust lottery tickets for pre-trained language models. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2211–2224. Association for Computational Linguistics.

Rui Zheng, Shihan Dou, Songyang Gao, Yuan Hua, Wei Shen, Binghai Wang, Yan Liu, Senjie Jin, Qin Liu, Yuhao Zhou, Limao Xiong, Lu Chen, Zhiheng Xi, Nuo Xu, Wenbin Lai, Minghao Zhu, Cheng Chang, Zhangyue Yin, Rongxiang Weng, Wensen Cheng, Haoran Huang, Tianxiang Sun, Hang Yan, Tao Gui, Qi Zhang, Xipeng Qiu, and Xuanjing Huang. 2023. Secrets of rlhf in large language models part i: Ppo. *arXiv preprint arXiv:2307.04964*.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. 2020. Freelb: Enhanced adversarial training for natural language understanding. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Terry Yue Zhuo, Zhuang Li, Yujin Huang, Yuan-Fang Li, Weiqing Wang, Gholamreza Haffari, and Fatemeh Shiri. 2023. On robustness of prompt-based semantic parsing with large pre-trained language model: An empirical study on codex. *CoRR*, abs/2301.12868.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 9-16, Harbin, China, August 3 - 5, 2023.