# Safety and Security in Large Language Models

**Messi H.J. Lee**
Washington University in St. Louis
hojunlee@wustl.edu

**Patrick Farley**
Washington University in St. Louis
pfarley@wustl.edu

**Sangwook Suh**
Washington University in St. Louis
sangwooksuh@wustl.edu

## Abstract

The involvement of Large Language Models (LLMs) in everyday life is growing by the day. However, there continues to be a wide array of issues and concerns pertaining to LLMs, particularly those related to safety and security. In this survey, we introduce three major safety- and security-related issues in LLMs. Then, we introduce state-of-the-art methods that are used to evaluate these issues. Finally, we discuss ongoing initiatives and methods being used to mitigate them in the literature.

## 1 The Three Issues

### 1.1 Bias

LLM bias refers to the tendency of LLMs to reproduce societal stereotypes or associations between social groups and semantic attributes. Studies in this literature have documented stereotypes pertaining to gender (Kotek et al., 2023; Lucy and Bamman, 2021), religion (Abid et al., 2021), and politics (Motoki et al., 2023; Rozado, 2023), among many others. A recent work by Kotek et al. (2023) gave the four most recently published LLMs (as of August 2023) ambiguous sentences where a feminine pronoun (i.e., she) either referred to a male stereotypical occupation or a female stereotypical occupation. When the LLMs were asked which occupation the pronoun was referring to, all four models were more likely to respond with the female stereotypical occupation, suggesting that LLMs associate the female gender with female stereotypical occupations. Another recent work by Motoki et al. (2023) demonstrates that ChatGPT responses tend to align with the political views of ChatGPT impersonating a Democrat, suggesting its tendency to manifest left-leaning political views. These studies, altogether, point to the LLMs' propensity to reproduce societal stereotypes and raises concerns for bias propagation and the sustenance of existing societal inequalities (see Bender et al., 2021).

### 1.2 Hallucinations

Another central issue is hallucination. Hallucination refers to the tendency of text generative models to generate text that is either unfaithful to the input text or nonfactual (Maynez et al., 2020). OpenAI (2023) labels the two types of hallucinations as "open domain" and "closed domain" hallucination where open domain hallucination refers to instances where the model makes up information that is not provided in the input text whereas closed domain hallucination refers to instances where the model provides false information about the world. Either type of hallucinations have been robustly documented in state of the art LLMs such as ChatGPT (Bang et al., 2023; Borji, 2023) and GPT-4 (OpenAI, 2023), with GPT-4 performing slightly better. Hallucinations of LLMs lead users to question the veracity and reliability of the model outputs (Weidinger et al., 2021) and further raises concerns for the application of LLMs in the real world. For example, the use of LLMs to summarize patient information, in the presence of hallucinations, can pose a direct threat to patients' lives. Furthermore, cases have shown that the use of LLMs for academic writing can result in an array of issues from non-factual references to making claims that are unsupported by scientific evidence (e.g., Alkaissi and McFarlane, 2023; Azamfirei et al., 2023). These studies highlight the potential for LLMs to mislead users and to propagate misinformation.

### 1.3 Privacy and Harmful Responses

The final issue is privacy and harmful responses. LLMs are trained on massive collections of text (e.g., Zhang et al., 2022; Chowdhery et al., 2022; Touvron et al., 2023). The issue with this is that the large bodies of text contain, among many others, publicly available personal information, and LLMs tend to memorize "atypical data points" that may contain sensitive information of the authors

or subjects of the text (Carlini et al., 2019, 2021, 2023). Such behavior of LLMs have raised concerns for *training data extraction attacks* where an adversary recovers individual training examples, such as social security numbers, using simple LLM queries (Carlini et al., 2021). The authors identified 46 named individuals and extracted contact information of 32 individuals from GPT-2 generated texts, highlighting the potential for personal data leakage facilitated by LLMs. OpenAI (2023) explains that the memorization tendency of LLMs combined with basic tasks pertaining to personal and geographic information can lead to responses that directly harm the privacy of individuals.

## 2 Evaluation

### 2.1 Bias

A common approach of evaluating bias in LLMs is to test whether LLMs generate biased responses when given unbiased prompts by having domain experts or members of the social group of interest manually review them for bias. However, given that the approach is not scalable, more recent efforts combine automated steps with human effort.

For example, Huang et al. (2023) first trained a classifier that detects bias in LLM-generated code. The classifier, also based on an LLM, was used to determine whether a piece of LLM-generated code was biased with respect to nine pre-defined bias categories (e.g., age, gender, race ... etc.). To account for cases where the classifier is unable to identify bias, they incorporated a human review stage where the researchers manually reviewed the code for bias. Another example of an approach combining both human and automated evaluation is WinoQueer (Felkner et al., 2023). WinoQueer is a benchmark dataset used to evaluate bias against members of the LGBTQ+ community through a community-in-the-loop method. The team first distributed online surveys to LGBTQ+ community members and asked them to share stereotypes about their LGBTQ+ identity that caused them harm. Using the information from the survey, they first generated biased statements about LGBTQ+ community members. Then, they replaced LGBTQ+ terms with non-LGBTQ+ terms to produce counterfactual statements, adapting methods from Zhao et al. (2018). Finally, they used the benchmark to measure the pseudo-log-likelihood probability of LLMs predicting a biased statement over the counterfactual one. Similar studies have developed different benchmarks targeting specific groups for LLM bias evaluation, such as Chinese cultures and values (Huang and Xiong, 2023) and under-represented societies in New Zealand (Yogarajan et al., 2023).

Another approach is prompting LLMs to impersonate members of particular social groups and to analyze the generated responses for latent biases that LLMs may have for the group. Salewski et al. (2023) asked LLMs to impersonate personas differing with respect to age, expertise, race, and gender and found that the model would display performance disparities consistent with societal biases. For example, a woman persona did a better job of describing birds whereas a man persona did a better job at identifying cars. Furthermore, Motoki et al. (2023) asked LLMs to impersonate individuals that identify with different political orientation. They assessed bias by comparing the personas' responses to questions pertaining to political viewpoints to responses of the default model.

### 2.2 Hallucinations

Hallucinations in LLMs can be evaluated using either the supervised or the unsupervised method. Supervised methods involve comparing the models' responses with the ground truth whereas unsupervised methods detect hallucinations without any reference to the ground truth.

The most common supervised method is to use a benchmark set of questions and verified answers to assess accuracy (e.g., Min et al., 2023; Muhlgay et al., 2023). One example is TruthfulQA (Lin et al., 2022). TruthfulQA is designed so that LLMs generate *imitative falsehoods*, responses that seem highly plausible given its distribution in the training data but are incorrect (e.g., popular misconceptions). Using this benchmark, Lin et al. (2022) found that larger models tend to be less truthful, generating responses that align with popular misconceptions. Similarly, Chen et al. (2023b) had LLMs respond to a question-answering dialogue dataset and evaluated the responses using four different metrics. The four metrics all assess the quality of the LLM responses by comparing them to the ground truth. Another supervised method is adversarial testing performed by domain experts. As part of the evaluation, experts craft prompts and evaluate the model's responses based on the degree, pattern, and harmfulness of the hallucination (e.g., OpenAI, 2023).

An example of an unsupervised method is SelfCheckGPT, a sampling based approach to evaluating hallucinations (Manakul et al., 2023). In this

approach, LLMs are asked to respond to the same prompt multiple times and the similarity of the responses is used to detect hallucinations. Similar responses across samples indicate that the LLM-generated information is accurate, whereas dissimilar responses across samples suggest the possibility of a hallucination. Manakul et al. (2023) demonstrates that the approach outperforms the supervised methods in a range of tasks, despite it not requiring any external resources (i.e., the ground truth). Another unsupervised method is the detection of self-contradictions. Mündler et al. (2023) triggers LLMs to generate pairs of sentences within the same context and uses a Natural Language Inference (NLI) based algorithm to detect self-contradictions. Using this method, they find that ChatGPT is prone to self-contradictions.

### 2.3 Privacy and Harmful Responses

Directly assessing *training data extraction attacks* (Carlini et al., 2021) within an LLM is difficult. Instead, researchers have proposed evaluating for *membership inference attacks* where researchers probe LLMs to see if a set of sensitive data was used to train the model (Shokri et al., 2017). A large body of work relies on "targeted" prompts to see if LLM responses contain personal information (Huang et al., 2022; Lukas et al., 2023; Kim et al., 2023), but recent efforts have started to focus on more general prompts to account for the fact that LLMs avoid verbatim memorization and output similar yet harmful memorizations of personal data (Ippolito et al., 2023).

Issues related to privacy are closely related to harmful responses where LLMs aid the user in doing illicit activities or harm members of marginalized groups through hate speech (Bommasani et al., 2021). To evaluate harmful responses within LLMs, Gehman et al. (2020) built a benchmark dataset using naturally occurring corpus of English text and toxicity scores from widely used classifiers. Another approach to evaluate harmful responses is *jailbreak attacks*. These attacks prompt LLMs to respond in ways that circumvent or contradict their own safety protocols. Wei et al. (2023) evaluated the performance of jailbreak prompts by tallying the number of LLM responses that are labeled as "GOOD BOT" if the model refused to respond, "BAD BOT" if the model generated an on-topic response, and "UNCLEAR" if neither.

## 3 Mitigation

### 3.1 Debiasing

One way to mitigate bias in LLMs is to *debias* them. One way to debias LLMs is to remove bias in the training data altogether and to re-train the model using the new dataset (Sun et al., 2019). This method is effective as it tackles the bias at its source, but removing bias in the training data and re-training resource-intensive models are both costly and ineffective. Another way of debiasing is counterfactual data augmentation (CDA; Zmigrod et al., 2019; Webster et al., 2021; Zhao et al., 2018). Zhao et al. (2018) proposed a data-augmentation strategy where all female entities in the training corpus are swapped with male entities, and vice versa. The authors found that re-training the model with the augmented dataset not only removed bias, the model performed equivalently well on the co-reference resolution task. However, data augmentation doubles the size of the dataset, which makes the strategy resource-intensive (Sun et al., 2019). Another debiasing strategy that has been shown to be effective is Self-Debias (Schick et al., 2021). The authors leverage the self-diagnostic capabilities of LLMs to discourage them from generating biased responses. Using a self-diagnosis template, the model determines whether their continuation to a prompt is desired or undesired, and based on that diagnosis, the probability distribution is modified to discourage undesired behavior. In a review of five LLM debiasing strategies including both CDA and Self-Debias, Meade et al. (2022) identified Self-Debias as the most effective strategy.

### 3.2 Hallucinations

One strategy that has been shown to be effective in mitigating hallucinations is Reinforcement Learning with Human Feedback (RLHF; Christiano et al., 2023; Stiennon et al., 2022; Ouyang et al., 2022) or the training of LLMs using human feedbacks on helpfulness and harmfulness of LLM responses. Ouyang et al. (2022) had human labelers provide "model answers" for input prompts and fine-tuned GPT-3 using supervised learning. They found that responses of the fine-tuned model were better, that they were more truthful (i.e., less hallucinations), and that they were less toxic (i.e., less harmful responses). More recent work has used a similar approach to directly address the "harmful" aspects of the model (Bai et al., 2022). Bai et al. (2022) created a dataset where labelers had LLMs give in-

structions or generate harmful information. At each time step of the conversation, the labelers chose which of the responses were more helpful or harmful. Then, with the two “helpful" and “harmful" datasets, the authors used reinforcement learning to fine-tune the LLMs. Consistent with findings from Ouyang et al. (2022), their fine-tuned model performed better with respect to truthfulness and honesty.

A more recent strategy proposed to mitigate hallucinations is Petite Unsupervised Research and Revision (PURR; Chen et al., 2023a). PURR works by first introducing corruption into pieces of text (e.g., randomly deleting words, shuffling the order of words, replacing words with synonyms). Then they have an LLM generate a collection of relevant evidence which is then used to fine-tune “compact editors" to de-noise the corrupted text. The model is evaluated on two criteria: preservation, how much of the original output was edited by PURR, and attribution, how much the statement can be inferred from the evidence. Chen et al. (2023a) found that PURR would make fewer large or unnecessary edits, more good edits, and edits that improve attribution while preserving the context of the query compared to two other state of the art editors devised for similar purposes: Evidence-based Factual Error Correction (EFEC; Thorne and Vlachos, 2021) and Retrofit Attribution using Research and Revision (RARR; Gao et al., 2023), suggesting that PURR could be a promising method to mitigate hallucinations in LLMs.

One last strategy that can be used to mitigate hallucinations is ExplanatoryGPT (Sovrano et al., 2023). ExplanatoryGPT integrates ChatGPT with an information retrieval software tool to produce responses that are grounded on reliable sources of information like textbooks. In the first step of ExplanatoryGPT, the information retrieval software extracts all information that is related to the user’s query from pertinent textbooks and documentation. The information is reorganized to serve the user’s specific query and then given to ChatGPT to format the response for better readability and coherence. By using external software to retrieve accurate information, ExplanatoryGPT can circumvent LLM hallucinations. Despite the promise that the strategy offers, there is yet to be work documenting its efficacy and there are questions surrounding its feasibility since it would require extensive and accurate databases on a wide range of technical areas.

### 3.3 Privacy and Harmful Responses

Some common strategies that the state of the art LLMs are using to mitigate privacy-related attacks like training data extraction attacks and membership inference attacks include: fine-tuning the models so that the models avoid answering queries that might violate a person’s security, creating automated evaluations of such attempts, and monitoring and restricting such queries (OpenAI, 2023). A more recent attempt to address a possible source of memorization tendency of LLMs is *deduplication*. Lee et al. (2022) identifies duplicate text and repetitive long strings as a cause of LLM memorization and propose tools to remove duplicate strings from LLM training data. While the authors do not experimentally establish the proposed method as an effective strategy to reduce memorization in LLMs, they discuss its potential to mitigating privacy concerns as well as reducing training time of the models to achieve equivalent model performance.

As for mitigating harmful responses and jailbreak attempts, *SmoothLLM* (Robey et al., 2023) is gaining traction as an effective mitigation strategy. The authors find that adversarial prompts (i.e., jailbreak prompts) are vulnerable to character-level changes, meaning that just a small change to an adversarial prompt can render the jailbreak attempt harmless. Based on this finding, the authors propose a defense mechanism where SmoothLLM randomly adds noise to the prompt (i.e., replacing a few characters, removing words ... etc.). If the generated text is inconsistent across perturbations, SmoothLLM classifies the prompt as adversarial and blocks any future attempts. This strategy showed massive success in preventing jailbreaks in state of the art LLMs including and not limited to GPT-4, Llama2, and Vicuna.

## 4 Conclusion

Bias, hallucination, and privacy (and harmful responses) are just three of many safety- and security-related issues pertaining to LLMs. These concerns pose direct threats to users in the form of amplification of societal stereotypes, the spread of misinformation, and privacy violations. In this survey, we discussed these issues and explored state of the art evaluation and mitigation strategies that are gaining traction in the literature. While there are improvements to be made, there are indications that LLMs are becoming more safe and secure with the collective effort of researchers in the field of NLP.

## References


Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent Anti-Muslim Bias in Large Language Models.

Hussam Alkaissi and Samy I McFarlane. 2023. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus*.

Razvan Azamfirei, Sapna R. Kudchadkar, and James Fackler. 2023. Large language models and the perils of their hallucinations. *Critical Care*, 27(1):120.

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 Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. ArXiv:2302.04023 [cs].

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 *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, Virtual Event Canada. ACM.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. 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 F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir P. Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the opportunities and risks of foundation models. *ArXiv*.

Ali Borji. 2023. A Categorical Archive of ChatGPT Failures. ArXiv:2302.03494 [cs].

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. Quantifying Memorization Across Neural Language Models. ArXiv:2202.07646 [cs].

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks. ArXiv:1802.08232 [cs].

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting Training Data from Large Language Models. ArXiv:2012.07805 [cs].

Anthony Chen, Panupong Pasupat, Sameer Singh, Hongrae Lee, and Kelvin Guu. 2023a. PURR: Efficiently Editing Language Model Hallucinations by Denoising Language Model Corruptions. ArXiv:2305.14908 [cs].

Yuyan Chen, Qiang Fu, Yichen Yuan, Zhihao Wen, Ge Fan, Dayiheng Liu, Dongmei Zhang, Zhixu Li, and Yanghua Xiao. 2023b. Hallucination detection: Robustly discerning reliable answers in large language models. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 245–255, New York, NY, USA. Association for Computing Machinery.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny

Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways. ArXiv:2204.02311 [cs].

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2023. Deep reinforcement learning from human preferences. ArXiv:1706.03741 [cs, stat].

Virginia K. Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. 2023. Winoqueer: A community-in-the-loop benchmark for anti-lgbtq+ bias in large language models.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Y. Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023. Rarr: Researching and revising what language models say, using language models.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models.

Dong Huang, Qingwen Bu, Jie Zhang, Xiaofei Xie, Junjie Chen, and Heming Cui. 2023. Bias Assessment and Mitigation in LLM-based Code Generation. ArXiv:2309.14345 [cs] version: 1.

Jie Huang, Hanyin Shao, and Kevin Chen-Chuan Chang. 2022. Are large pre-trained language models leaking your personal information?

Yufei Huang and Deyi Xiong. 2023. CBBQ: A Chinese Bias Benchmark Dataset Curated with Human-AI Collaboration for Large Language Models. ArXiv:2306.16244 [cs].

Daphne Ippolito, Florian Tramèr, Milad Nasr, Chiyuan Zhang, Matthew Jagielski, Katherine Lee, Christopher A. Choquette-Choo, and Nicholas Carlini. 2023. Preventing verbatim memorization in language models gives a false sense of privacy.

Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. 2023. Propile: Probing privacy leakage in large language models.

Hadas Kotek, Rikker Dockum, and David Q. Sun. 2023. Gender bias and stereotypes in Large Language Models. In *Proceedings of The ACM Collective Intelligence Conference*, pages 12–24. ArXiv:2308.14921 [cs].

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating Training Data Makes Language Models Better. ArXiv:2107.06499 [cs].

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Li Lucy and David Bamman. 2021. Gender and Representation Bias in GPT-3 Generated Stories. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55, Virtual. Association for Computational Linguistics.

Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. 2023. Analyzing leakage of personally identifiable information in language models.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2022. An Empirical Survey of the Effectiveness of Debiasing Techniques for Pre-trained Language Models. ArXiv:2110.08527 [cs].

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation.

Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. 2023. More human than human: measuring ChatGPT political bias. *Public Choice*.

Dor Muhlgay, Ori Ram, Inbal Magar, Yoav Levine, Nir Ratner, Yonatan Belinkov, Omri Abend, Kevin Leyton-Brown, Amnon Shashua, and Yoav Shoham. 2023. Generating benchmarks for factuality evaluation of language models.

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2023. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation.

OpenAI. 2023. GPT-4 Technical Report. ArXiv:2303.08774 [cs].

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. ArXiv:2203.02155 [cs].

Alexander Robey, Eric Wong, Hamed Hassani, and George J. Pappas. 2023. Smoothllm: Defending large language models against jailbreaking attacks.

David Rozado. 2023. The Political Biases of ChatGPT. *Social Sciences*, 12(3):148. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.

Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. 2023. In-context impersonation reveals large language models' strengths and biases.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP. *Transactions of the Association for Computational Linguistics*, 9:1408–1424.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models.

Francesco Sovrano, Kevin Ashley, and Alberto Bacchelli. 2023. Toward Eliminating Hallucinations: GPT-based Explanatory AI for Intelligent Textbooks and Documentation. In *CEUR Workshop Proceedings*.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2022. Learning to summarize from human feedback. ArXiv:2009.01325 [cs].

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review.

James Thorne and Andreas Vlachos. 2021. Evidence-based factual error correction. 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 3298–3309, Online. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. ArXiv:2302.13971 [cs].

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. 2021. Measuring and Reducing Gendered Correlations in Pre-trained Models. ArXiv:2010.06032 [cs].

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023. Jailbroken: How does llm safety training fail?

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from Language Models. ArXiv:2112.04359 [cs].

Vithya Yogarajan, Gillian Dobbie, Timothy Pistotti, Joshua Bensemann, and Kobe Knowles. 2023. Challenges in Annotating Datasets to Quantify Bias in Under-represented Society. ArXiv:2309.08624 [cs] version: 1.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: Open Pre-trained Transformer Language Models. ArXiv:2205.01068 [cs].

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual Data Augmentation for Mitigating Gender Stereotypes in Languages with Rich Morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.