References

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, oct 2016. doi: 10.1145/2976749.2978318. URL `https://doi.org/10.1145%2F2976749.2978318`.

Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. Llm-deliberation: Evaluating llms with interactive multi-agent negotiation games. *arXiv preprint arXiv:2309.17234*, 2023.

Meeraj Ajam. Intelligent meeting recap in teams premium, now available. `https://techcommunity.microsoft.com/t5/microsoft-teams-blog/intelligent-meeting-recap-in-teams-premium-now-available/ba-p/3832541`, 2023. Accessed: 2023-09-23.

Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. What does it mean for a language model to preserve privacy? In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2280–2292, 2022.

Lindsey J Byom and Bilge Mutlu. Theory of mind: Mechanisms, methods, and new directions. *Frontiers in human neuroscience*, 7:413, 2013.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646*, 2022.

Yang Chen, Ethan Mendes, Sauvik Das, Wei Xu, and Alan Ritter. Can language models be instructed to protect personal information? *arXiv preprint arXiv:2310.02224*, 2023.

Malinda J Colwell, Kimberly Corson, Anuradha Sastry, and Holly Wright. Secret keepers: children's theory of mind and their conception of secrecy. *Early Child Development and Care*, 186(3):369–381, 2016.

Haonan Duan, Adam Dziedzic, Nicolas Papernot, and Franziska Boenisch. Flocks of stochastic parrots: Differentially private prompt learning for large language models. *arXiv preprint arXiv:2305.15594*, 2023.

Benj Edwards. Ai-powered bing chat spills its secrets via prompt injection attack. `https://tinyurl.com/injection-attack`, 2023. Accessed: 2023-09-23.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 653–670, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.48. URL `https://aclanthology.org/2020.emnlp-main.48`.

Carl Franzen. Oops! google search caught publicly indexing users' conversations with bard ai. `https://tinyurl.com/google-bard-leak`, 2023. Accessed: 2023-09-27.

Lauren Good. Chatgpt can now talk to you—and look into your life. `https://www.wired.com/story/chatgpt-can-now-talk-to-you-and-look-into-your-life/`, 2023. Accessed: 2023-09-28.

John Hart, J. Gratch, and Stacy Marsella. How virtual reality training can win friends and influence people. 2013. URL `https://api.semanticscholar.org/CorpusID:146359723`.

Paul M Heider, Jihad S Obeid, and Stéphane M Meystre. A comparative analysis of speed and accuracy for three off-the-shelf de-identification tools. *AMIA Summits on Translational Science Proceedings*, 2020:241, 2020.

Liwei Jiang, Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny Liang, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jon Borchardt, Saadia Gabriel, Yulia Tsvetkov, Oren Etzioni, Maarten Sap, Regina A. Rini, and Yejin Choi. Can machines learn morality? the delphi experiment. 2021. URL `https://api.semanticscholar.org/CorpusID:250492927`.