Tanya Chowdhury, Razieh Rahimi, and James Allan. Rank-lime: Local model-agnostic feature attribution for learning to rank, 2022.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

Prithiviraj Damodaran. Parrot: Paraphrase generation for nlu., 2021.

Amit Dhurandhar. Auto-correlation dependent bounds for relational data. In *Proc. of the 11th Workshop on Mining and Learning with Graphs. Chicago*, 2013.

Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M. Rush. Gltr: Statistical detection and visualization of generated text, 2019.

Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey, 2018.

Wang-Cheng Kang and Julian McAuley. Self-attentive sequential recommendation, 2018.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. Dense passage retrieval for open-domain question answering, 2020.

Su Young Kim, Hyeonjin Park, Kyuyong Shin, and Kyung-Min Kim. Ask me what you need: Product retrieval using knowledge from gpt-3, 2022.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models, 2023.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense, 2023.

Nir Kshetri and Jeffrey Voas. Deep learning–based social media misinformation detection. *IEEE Software*, 39(1):53–59, 2022. doi: 10.1109/MS.2022.3053106.

Gerhard C Langelaar, Iwan Setyawan, and Reginald L Lagendijk. Watermarking digital image and video data. a state-of-the-art overview. *IEEE Signal processing magazine*, 17(5):20–46, 2000.

Thomas Lavergne, Tanguy Urvoy, and François Yvon. Detecting fake content with relative entropy scoring. In *Pan*, 2008.

Lucien Le Cam. *Asymptotic methods in statistical decision theory*. Springer Science & Business Media, 2012.

Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. Gpt detectors are biased against non-native english writers, 2023.

William Lifferth. Fake news, 2018. URL `https://kaggle.com/competitions/fake-news`.

Manuel V. Loureiro, Steven Derby, and Tri Kurniawan Wijaya. Topics as entity clusters: Entity-based topics from language models and graph neural networks, 2023.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pp. 142–150, 2011.

Hasan Mesut Meral, Bülent Sankur, A. Sumru Özsoy, Tunga Güngör, and Emre Sevinç. Natural language watermarking via morphosyntactic alterations. *Comput. Speech Lang.*, 23(1):107–125, jan 2009. ISSN 0885-2308. doi: 10.1016/j.csl.2008.04.001. URL https://doi.org/10.1016/j.csl.2008.04.001.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. Detectgpt: Zero-shot machine-generated text detection using probability curvature, 2023.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1797–1807, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1206. URL https://aclanthology.org/D18-1206.

OpenAI. Ai text classifier. https://platform.openai.com/ai-text-classifier, 2023.

OpenAI. Gpt-4 technical report, 2023.

Hariom A. Pandya and Brijesh S. Bhatt. Question answering survey: Directions, challenges, datasets, evaluation matrices, 2021.

Yury Polyanskiy and Yihong Wu. Information theory: From coding to learning, 2022.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL https://aclanthology.org/D16-1264.

Alexander M Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 379–389, 2015.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. Can ai-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*, 2023.

Tyler Schildhauer. Fake news detection in the era of ai. In *Proceedings of the 25th ACM Conference on Computer-Supported Cooperative Work and Social Computing*, pp. 1–10. ACM, 2022. doi: 10.1145/1234567.1234567.

Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C Courville, and Joelle Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 2015.

Jin Dong Shin, Hyun Jae Kim, Kyung Min Lee, Seonghyeon Kim, and Seungwon Shin. Multilingual language generation and automatic writing evaluation with transformer models. *arXiv preprint arXiv:2104.06399*, 2021.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, Miles McCain, Alex Newhouse, Jason Blazakis, Kris McGuffie, and Jasmine Wang. Release strategies and the social impacts of language models, 2019.

Salil Pravin Vadhan. *A study of statistical zero-knowledge proofs*. PhD thesis, Massachusetts Institute of Technology, 1999.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30:5998–6008, 2017.

Harsh K Verma, Abhishek Narain Singh, and Raman Kumar. Robustness of the digital image watermarking techniques against brightness and rotation attack, 2009.

Shweta Wadhera, Deepa Kamra, Ankit Rajpal, Aruna Jain, and Vishal Jain. A comprehensive review on digital image watermarking, 2022.

Zhen Wang. Modern question answering datasets and benchmarks: A survey, 2022.

Larry Wasserman. Lecture notes for stat 705: Advanced data analysis. https://www.stat.cmu.edu/~larry/=stat705/Lecture27.pdf, 2013. Accessed on April 9, 2023.

Heqi Zheng, Xiao Zhang, Zewen Chi, Heyan Huang, Tan Yan, Tian Lan, Wei Wei, and Xian-Ling Mao. Cross-lingual phrase retrieval, 2022.

Xiaofei Zou and Xu Ling. Ai-based detection of misinformation in social media. *IEEE Access*, 9: 112408–112418, 2021. doi: 10.1109/ACCESS.2021.3104419.