Alex, W., Cho, K., & Lewis, M. (2020). Asking and answering questions to evaluate the factual consistency of summaries. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 5008–5020). Association for Computational Linguistics.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V. 2. The Journal of Technology, Learning and Assessment, 4(3), 1-30.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (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 (pp. 610-623). ACM.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv. https://arxiv.org/abs/2108.07258

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. arXiv. https://arxiv.org/abs/2005.14165

Camburu, O. M., Rocktäschel, T., Lukasiewicz, T., & Blunsom, P. (2018). e-SNLI: Natural language inference with natural language explanations. In Advances in Neural Information Processing Systems (pp. 9539-9549). Curran Associates, Inc.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv. https://arxiv.org/abs/1810.04805

Dziri, N., Kamalloo, E., Mathewson, K. W., & Zaiane, O. (2021). On the origin of hallucinations in conversational models: Is it the datasets or the models? arXiv. https://arxiv.org/abs/2104.06245

European Commission. (2021). Proposal for a Regulation laying down harmonised rules on artificial intelligence. https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence

Farzi, N., & Dietz, L. (2023). Pencils down! Automatic rubric-based evaluation of retrieve/generate systems. In Proceedings of the 10th ACM SIGIR Conference on the Theory of Information Retrieval. ACM.

Federal Trade Commission. (2023). FTC report warns about using artificial intelligence to combat online problems. https://www.ftc.gov/news-events/news/press-releases/2023/06/ftc-report-warns-about-using-artificial-intelligence-combat-online-problems

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. Minds and Machines, 30(4), 681-694.

Gao, J., Galley, M., & Li, L. (2019). Neural approaches to conversational AI. Foundations and Trends® in Information Retrieval, 13(2-3), 127-298.

Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M. W. (2020). Retrieval augmented language model pre-training. In International Conference on Machine Learning (pp. 3929-3938). PMLR.

Ishida, T. (2024). Facilitating holistic evaluations with LLMs: Insights from scenario-based experiments. arXiv. https://arxiv.org/abs/2405.17728

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. Educational Research Review, 2(2), 130-144.

Kuechler, W. L., & Simkin, M. G. (2010). Why is performance on multiple-choice tests and constructed-response tests not more closely related? Theory and an empirical test. Decision Sciences Journal of Innovative Education, 8(1), 55-73.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. arXiv. https://arxiv.org/abs/2005.11401

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv. https://arxiv.org/abs/1907.11692

Liu, Y., Zhu, Y., Liu, P., & Peng, M. (2023). GPTEval: NLG evaluation using GPT-4 with better human alignment. arXiv. https://arxiv.org/abs/2303.16634

Marcus, G., & Davis, E. (2020). GPT-3, bloviator: OpenAI's language generator has no idea what it's talking about. MIT Technology Review. https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/

Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. arXiv. https://arxiv.org/abs/2005.00661

Mehri, S., & Eskenazi, M. (2020). USR: An unsupervised and reference free evaluation metric for dialog generation. arXiv. https://arxiv.org/abs/2005.00456

Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. Practical Assessment, Research, and Evaluation, 7(1), 10.

Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. Educational Research Review, 9, 129-144.

Petroni, F., Piktus, A., Fan, A., Lewis, P., Yazdani, M., De Cao, N., ... & Riedel, S. (2021). KILT: A benchmark for knowledge intensive language tasks. arXiv. https://arxiv.org/abs/2009.02252

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21, 1-67.

Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 469-481). ACM.

Reddy, Y. M., & Andrade, H. (2010). A review of rubric use in higher education. Assessment & Evaluation in Higher Education, 35(4), 435-448.

Rosenthal, S., Farra, N., & Nakov, P. (2017). SemEval-2017 task 4: Sentiment analysis in Twitter. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017) (pp. 502-518). Association for Computational Linguistics.

Senanayake, C., & Asanka, D. (2024). Rubric based automated short answer scoring using large language models (LLMs). In 2024 International Research Conference on Smart Computing and Systems Engineering (SCSE) (Vol. 7). IEEE.

Shermis, M. D., & Burstein, J. (Eds.). (2013). Handbook of automated essay evaluation: Current applications and new directions. Routledge.

Stahl, M., Grotov, A., Stremmel, H., & Schüssler, M. (2024). Exploring LLM prompting strategies for joint essay scoring and feedback generation. arXiv. https://arxiv.org/abs/2404.15845

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., ... & Christiano, P. (2020). Learning to summarize with human feedback. arXiv. https://arxiv.org/abs/2009.01325

Taghipour, K., & Ng, H. T. (2016). A neural approach to automated essay scoring. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (pp. 1882-1891). Association for Computational Linguistics.

Tamburri, D. A. (2020). Sustainable MLOps: Trends and challenges. In 2020 22nd International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC) (pp. 17-23). IEEE.

Tian, X., Pallay, C., Cloudt, E., Basu, D., Xu, Y., Peteranetz, M., ... & Wang, L. (2024). Examining LLM prompting strategies for automatic evaluation of learner-created computational artifacts. arXiv. https://arxiv.org/abs/2404.xxxxx

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (pp. 5998-6008). Curran Associates, Inc.

Kryscinski, W., McCann, B., Xiong, C., & Socher, R. (2020). Evaluating the factual consistency of abstractive text summarization. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 9332–9346). Association for Computational Linguistics.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., ... & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In International Conference on Machine Learning (pp. 2048-2057). PMLR.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). BERTScore: Evaluating text generation with BERT. arXiv. https://arxiv.org/abs/1904.09675

Zhao, T., Wang, X., Clevert, D. A., Horng, T., Tao, C., & Liu, T. Y. (2021). Calibrate before use: Improving few-shot performance of language models. arXiv. https://arxiv.org/abs/2102.09690

Zhong, M., Liu, Y., Yin, D., Mao, Y., Jiao, Y., Liu, P., ... & Han, J. (2022). Towards a unified multidimensional evaluator for text generation. arXiv. https://arxiv.org/abs/2210.07197