[16] Jack Collens, Rachel Reimer, Gerald Schifman, and Pamela Wilkinson. AI Survey: Where Artificial Intelligence Stands in the Legal Industry, April 2024. URL https://assets.law360news.com/1826000/1826128/law360_pulse-ai_survey.pdf.

[17] David Caswell. AI and journalism: What's next?, 2023. URL https://reutersinstitute.politics.ox.ac.uk/news/ai-and-journalism-whats-next.

[18] Luke Hurst. Robot reporters? here's how news organisations are using ai in journalism. *Euronews*, 2023. URL https://www.euronews.com/next/2023/08/24/robot-reporters-heres-how-news-organisations-are-using-ai-in-journalism.

[19] M Sweney. Mirror and express owner publishes first articles written using ai. *The Guardian*, 2023. URL https://www.theguardian.com/business/2023/mar/07/mirror-and-express-owner-publishes-first-articles-written-using-ai.

[20] Rose E. Wang, Ana T. Ribeiro, Carly D. Robinson, Susanna Loeb, and Dora Demszky. Tutor copilot: A human-ai approach for scaling real-time expertise. *arXiv preprint arXiv:2410.03017*, 2024.

[21] Andy Extance. Chatgpt has entered the classroom: how llms could transform education. *Nature*, 623(7987):474–477, 2023.

[22] Dorottya Demszky and Jing Liu. M-powering teachers: Natural language processing powered feedback improves 1: 1 instruction and student outcomes. In *Proceedings of the Tenth ACM Conference on Learning@ Scale*, pages 59–69, 2023.

[23] Dorottya Demszky, Jing Liu, Heather C Hill, Dan Jurafsky, and Chris Piech. Can automated feedback improve teachers' uptake of student ideas? evidence from a randomized controlled trial in a large-scale online course. *Educational Evaluation and Policy Analysis*, page 01623737231169270, 2023.

[24] Davide Castelvecchi. Researchers built an 'AI Scientist' – what can it do? *Nature*, 2024. URL https://www.nature.com/articles/d41586-024-02842-3.

[25] Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. Researchagent: Iterative research idea generation over scientific literature with large language models. *arXiv preprint arXiv:2404.07738*, 2024.

[26] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

[27] Alex Kim, Maximilian Muhn, and Valeri Nikolaev. Financial statement analysis with large language models. *arXiv preprint arXiv:2407.17866*, 2024.

[28] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*, 2023.

[29] Xiao-Yang Liu, Guoxuan Wang, Hongyang Yang, and Daochen Zha. FinGPT: Democratizing internet-scale data for financial large language models. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*, 2023. URL https://openreview.net/forum?id=5BqWC1Fz8F.

[30] Alejandro Lopez-Lira and Yuehua Tang. Can chatgpt forecast stock price movements? return predictability and large language models. *arXiv preprint arXiv:2304.07619*, 2023.

[31] David Robson. Ai wisdom: What happens when you ask an algorithm for relationship advice. 2024. URL https://www.bbc.com/future/article/20240515-ai-wisdom-what-happens-when-you-ask-an-algorithm-for-relationship-advice.

[32] Britney Nguyen. I asked chatgpt for online and in-person dating advice, here's what relationship coaches think of its answers. 2023. URL https://www.businessinsider.com/what-real-life-relationship-coaches-think-chatgpts-dating-advice-2023-1.

[33] Ryan Liu, Howard Yen, Raja Marjieh, Thomas L Griffiths, and Ranjay Krishna. Improving interpersonal communication by simulating audiences with language models. *arXiv preprint arXiv:2311.00687*, 2023.

[34] Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. Neural theory-of-mind? on the limits of social intelligence in large LMs. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3762–3780, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.248. URL https://aclanthology.org/2022.emnlp-main.248.

[35] Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. Understanding social reasoning in language models with language models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=8bqjirgxQM.

[36] Michal Kosinski. Theory of mind might have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*, 2023.

[37] Tomer Ullman. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*, 2023.

[38] Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. Clever hans or neural theory of mind? stress testing social reasoning in large language models. In Yvette Graham and Matthew Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2257–2273, St. Julian's, Malta, March 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.eacl-long.138.

[39] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

[40] Sean Trott, Cameron Jones, Tyler Chang, James Michaelov, and Benjamin Bergen. Do large language models know what humans know? *Cognitive Science*, 47(7):e13309, 2023.

[41] Jaan Aru, Aqeel Labash, Oriol Corcoll, and Raul Vicente. Mind the gap: Challenges of deep learning approaches to theory of mind. *Artificial Intelligence Review*, 56(9):9141–9156, 2023.

[42] Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 2024. URL https://web.mit.edu/bcs/nklab/media/pdfs/Mahowald.TICs2024.pdf.

[43] Matthew Le, Y-Lan Boureau, and Maximilian Nickel. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5872–5877, 2019.

[44] Xiaomeng Ma, Lingyu Gao, and Qihui Xu. Tomchallenges: A principle-guided dataset and diverse evaluation tasks for exploring theory of mind. In *Conference on Computational Natural Language Learning*, 2023. URL https://api.semanticscholar.org/CorpusID:258865295.

[45] Kanishk Gandhi, Jan-Philipp Franken, Tobias Gerstenberg, and Noah D. Goodman. Understanding social reasoning in language models with language models. *ArXiv*, abs/2306.15448, 2023. URL https://api.semanticscholar.org/CorpusID:259262573.

[46] Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. Hi-ToM: A benchmark for evaluating higher-order theory of mind reasoning in large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10691–10706, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.717. URL https://aclanthology.org/2023.findings-emnlp.717.

[47] Cameron Robert Jones, Sean Trott, and Ben Bergen. EPITOME: Experimental protocol inventory for theory of mind evaluation. In *First Workshop on Theory of Mind in Communicating Agents*, 2023. URL https://openreview.net/forum?id=e5Yky8Fnvj.

[48] Pei Zhou, Aman Madaan, Srividya Pranavi Potharaju, Aditya Gupta, Kevin R. McKee, Ari Holtzman, Jay Pujara, Xiang Ren, Swaroop Mishra, Aida Nematzadeh, Shyam Upadhyay, and Manaal Faruqui. How far are large language models from agents with theory-of-mind?, 2024. URL https://openreview.net/forum?id=xnUIMz5u2s.

[49] Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. OpenToM: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8593–8623, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.466.

[50] Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1819–1862, 2024.

[51] Victoria Basmov, Yoav Goldberg, and Reut Tsarfaty. Llms' reading comprehension is affected by parametric knowledge and struggles with hypothetical statements. *arXiv preprint arXiv:2404.06283*, 2024.

[52] Victoria Basmov, Yoav Goldberg, and Reut Tsarfaty. Simple linguistic inferences of large language models (llms): Blind spots and blinds. *arXiv preprint arXiv:2305.14785*, 2023.

[53] Wesley H Holliday and Matthew Mandelkern. Conditional and modal reasoning in large language models. *arXiv preprint arXiv:2401.17169*, 2024.

[54] OpenAI. Hello GPT-4o, 2024. URL https://openai.com/index/hello-gpt-4o/.

[55] OpenAI. GPT-4 Technical Report, March 2023. URL https://arxiv.org/abs/2303.08774.

[56] OpenAI. Introducing ChatGPT, November 2023. URL https://openai.com/blog/chatgpt.

[57] Anthropic. The Claude 3 Model Family: Opus, Sonnet, Haiku, March 2024. URL https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf.

[58] AI@Meta. Llama 3 Model Card, 2024. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

[59] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[60] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[61] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

[62] Mirac Suzgun, Stuart Shieber, and Dan Jurafsky. string2string: A modern python library for string-to-string algorithms. In Yixin Cao, Yang Feng, and Deyi Xiong, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 278–285, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-demos.26.